Tracking the neural correlates of learning to read with dense-sampling fMRI

Alexander Enge[1,2] & Michael A. Skeide[1]

[1] Max Planck Institute for Human Cognitive and Brain Sciences
[2] Humboldt-Universität zu Berlin

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Keywords: reading, literacy, learning, children, fMRI, neuroimaging

Introduction

Learning to read is a developmental milestone of lifelong importance. Reading and writing enables us to exchange messages across time and space, to remember things we would otherwise forget all too easily, to educate ourselves and others, or to get lost in fictional worlds whenever our own one seems a bit much. Without it, we would likely be missing out on many of humankind's greatest a achievements or, at the very least, you, our dear reader, would not be able to read this manuscript. It therefore seems fair to say that the process of learning to read, typically taking place through late kindergarten and early primary school education as well as through the help of parents and other caretakers, is a crucial step of children's education and cognitive development.

Children typically pick up knowledge about their first letters informally, for example by being shown how their own name is written or by memorizing the logo of their favorite TV series or cereal brand. However, at this first, *logographic* stage of reading development (Ehri, 2005; Frith, 1986), children are typically not yet aware of the systematic correspondence between individual letters (graphemes) and speech sounds (phonemes), and therefore unable to pronounce or understand novel words. In this stage, written words are probably perceived similar to other types of visual objects, and novel words that were never encounter before do not elicit any phonological or semantic representation as they would for a trained reader.

Through structured training in kindergarten, primary school, or at home, children advance to the second, *alphabetic* stage of reading development,[1] where they are taught that individual letters are associated with individual speech sounds. This allows children to process the individual letters of a written word sequentially, convert them to their corresponding speech sounds, and combine those speech sounds to vocalize the entire word. Since most healthy children will have become proficient in spoken language comprehension long before learning to read (typically with < 3 years of age), they will also be able to access the meaning of the written word based on its assembled spoken word form. For this way of reading to be successful, children need to possess the understanding and skill to break down words into their constituent phonemes, typically referred to as "phonological awareness." It is therefore unsurprising that phonological skills first-letter naming, rhyming, and verbal short-term memory are among the best predictors of individual differences in future reading ability (e.g., Melby-Lervåg et al., 2012). However, this alphabetic form of reading remains somewhat slow and error-prone due to the sequential conversion of letters into speech sounds.

Only in the third and final stage of reading development, the *orthographic* stage, children become able to read words as a whole and to access their meaning directly from the visual word form, without having to take the "detour" of phonological decoding. Presumably, it is the emergence of this shortcut that substantially reduces word reading times after approximately the first year of reading instruction (e.g., Hasenäcker et al., 2017) and which makes adult reading so efficient and seemingly effortless.

The analysis code for this study is openly available at XXX. We have no conflict of interest to disclose.

Correspondence concerning this article should be addressed to Alexander Enge, Stephanstraße 1a, 04103 Leipzig, Germany. E-mail: alex_enge@web.de

[1]Note, however, that as with any model of developmental "stages" or "phases," there typically is no hard and clear-cut boundary between one "stage" and the next, and there typically is large interindividual variation in the onsets and durations of the "stages."

Neural correlates of reading

Reading as a cognitive function has to be implemented on a neuroanatomical and functional level in the brain. To probe which brain areas contribute to different aspects reading in proficient readers, cognitive-neuroscientific methods such as electroencephalography (EEG) and magnetic resonance imaging (MRI) can be used to measure brain activity while participants perform reading tasks. Due to its comparatively high spatial resolution (on the order of a few millimeters), functional magnetic resonance imaging (fMRI), which measures changes in blood oxygen level as a proxy of local neuronal activity, can be used to isolate which brain areas which are more active during reading than during other control tasks.

One well-replicated finding using this methodology is that processing written words engages a relatively circumscribed region on the left ventral occipito-temporal (vOT) cortex. This region, typically referred to as the visual word form area (VWFA; e.g., Cohen et al., 2000; Dehaene et al., 2002; Dehaene & Cohen, 2011; McCandliss et al., 2003), is thought to be a purely visual and pre-lexical area that identifies written words based on its lower-level visual shapes. There is converging evidence that anomalies in, damage to, and stimulation of the VWFA can cause reading difficulties (Brem et al., 2020; Hillis et al., 2005; Hirshorn et al., 2016; but see also Price & Devlin, 2003). Furthermore, the VWFA is often described as a prime example of the neuronal recycling hypothesis (Dehaene & Cohen, 2007), according to which "modern" cognitive functions such as reading,[2] for which not enough evolutionary time has passed to develop dedicated brain circuitry, repurpose brain areas with similar but evolutionarily older functions. In the case of reading, the VWFA may reuse regions that had previously been specialized for recognizing other complex visual object categories such as faces, limbs, or tools (Dehaene et al., 2015; Kubota et al., 2024; Nordt et al., 2021; but see also Coltheart, 2014).

The brain areas involved in reading beyond the early stages of visual word form recognition are less well understood, presumably owing to large differences in task design between studies. A meta-analysis of fMRI study (Murphy et al., 2019) found that beyond the VWFA, single word reading reliable engages a left-lateralized set of regions including the left inferior frontal and left superior and middle temporal gyri, all of which are known to be involved in phonolical and semantic processing of spoken language. However, it remains an open question which brain areas serve as the interface between visual processing (word form recognition in the VWFA) and language processing (phonological and semantic processing in the left perisylvian language network) during reading. One candidate for the visual–phonological interface (linking written letters to speech sounds) is the left posterior supior temporal gyrus, as it has been show to integrate auditory and visual information when both are presented concurrently, e.g., during audio-visual letter perception or lip reading tasks (e.g., Blau et al., 2010; Calvert et al., 1997; van Atteveldt et al., 2004; Wilson et al., 2018).[3] One candidate for the visual–semantic interface is the left middle fusiform cortex (lmFFC), which lies anterior the high-level visual object recognition areas (including the VWFA) on the ventral surface of the occipital and temporal lobes. Using depth electrodes for recording and stimulation in epileptic patients,[4] it has been shown that this region is active for lexical retrieval from both auditory (spoken words) and visual (written words and object images) input (e.g., Forseth et al., 2018; Woolnough et al., 2020, 2022).

Neural correlates of reading development

Compared to hundreds if not thousands of studies in proficient adult readers, there has been a lot less research on how the neural correlates of reading develop in beginning readers. There are multiple reasons for this: (a) It is much harder to recruit children (and their families) as compared to undergraduate students who depend on obtaining course credit or monetary compensation; (b) Children in the relevant age range (late kindergarten to early primary school; typically ~5–8 years of age) have a shorter attention spand and show more in-scanner head movement, leading to fewer usable scans, lower data quality in usable scans, and fewer data points per scan; and (c) Accurately tracking the development of brain structure and function in individual children requires a longitudinal study design, which is very costly and demanding for study participants, typically leading to small sample sizes. Nevertheless, a few studies exist that have managed to obtain multiple longitudinal scans of children's brain activity as they are learning to read.

Dehaene-Lambertz et al. (2018) scanned scanned ten 6-year-old children longitudinally throughout the first year of schooling (6–7 scans per child) and presented them with visual objects from different object categories, including written words. Behaviorally, they found that reading performance (knowledge of grapheme–phoneme relations and number of words read per minutes in a standardized reading test) increased sharply during the first months of schooling. In the fMRI, they found that selectivity for words in the VWFA was not present at the beginning of the study but quickly emerged within less than 6 months in most of the children. Interestingly, the activation strength in the VWFA and other word-selective areas appeared to follow a curvilinear inverted "u"-shaped pattern, with a quick rise in BOLD activation strength in the first half of the study followed by a slight decline in the second half. Such a pattern would be predicted by the "expansion and renormalization" model of brain plasticity (Wenger et al., 2017), according to which a novel skill initially requires more resources (in terms of number of voxels or BOLD activation amplitude) but later on becomes more efficient and automatized. Regarding

[2] The first known scripts appeared approximately 3000–5000 B.C.E (e.g., Houston, 2004).

[3] Note that if the left pSTS does indeed play a role in reading, this could be viewed as another case of neuronal recycling, is its "new" skill of linking written letters to speech sounds may stem from its "original" skill to link lip movements to speech sounds.

[4] Unfortunately, the depth of this region and the associated signal dropout makes it hard to pick up BOLD activity changes in this region with fMRI.

the neuronal recycling hypothesis, Dehaene-Lambertz et al. (2018) found that the voxels that would later form the VWFA in individual children were weakly tuned to a different object category at the beginning of the study, namely tools. However, a few shortcomings of this study need to be noted: (a) The sample size (both in terms of number of children and number of scans per child) was very small, therefore leading to low statistical power to detect all but very large effects (see also Button et al., 2013; Ioannidis, 2005; Szucs & Ioannidis, 2017); (b) Their analysis of variance (ANOVA) model did not take into account the longitudinal nature of the data, with repeated measures of the same participants likely being positively correlated with one another; (c) It is unclear if the changes observed in this study were caused by reading instruction in particular or by other aspects of schooling or general cognitive development, since there was no non-reading control group (for obvious practical and ethical reasons); and (d) The study design captured only the very first stage of reading, namely visual word form recognition, while it remains an open question how visual word forms get linked to other aspects of (spoken) language comprehension, namely speech sounds (phonology) and word meaning (semantics).

Cultural biases in reading research

Most of psychological and cognitive-neuroscientific research is carried out in countries of the Global North, especially in Western Europe and Northern America. Within these countries, study participants are not sampled at random, but typically come from economically and educationally priviledged social backgrounds (e.g., psychology students). This restriction of study samples can limit the generalizability of research findings, as even basic and presumably "universal" effects such as some perceptual illusions do not necessarily replicate in participants from different cultural and socio-economic backgrounds (Blasi et al., 2022; Henrich et al., 2010).

In reading research, this problem is potentiated by the fact that different cultures developed—sometimes radically—different writing systems. Due to the concentration of research funds and technology in the Global North, research on reading, its development, and its neural correlates has almost exclusively been carried out in languages with *alphabetic* writing systems such as English, German, or French. In these writing systems, there is a corresponends[5] between individual units of written language (graphemes) and very small units of spoken language (phonemes). On the other end of the spectrum are *logographic* writing systems (such Chinese), in which individual graphemes correspond to large units of spoken languages, namely entire words or concepts (morphemes). In between these two extreems sit *syllabic* writing systems, in which individual graphemes correspond to intermediate units of spoken languages, namely sublexical consonant–vowel combinations (syllables). A special case are *alphasyllabic* languages (also "abugidas"; such as the Hindi writing system Devanagari), in which individual graphemes correspond to consonants with added diacritical marks above, below, or next to the letter for vowels. There is an abundance of research findings on different aspects of reading and its neural correlates in alphabetic languages,[6] but only very few studies that used logographic writing systems (typically Chinese) and next to none that used syllabic or alphasyllabic writing systems. These biases clearly limit the scope of empirical findings and current theories on reading and its development to a small set of regions and languages (Frost, 2012; Share, 2008, 2014, 2021).

The present study

Our goal here was to capture the developmental changes in brain activity related to written word processing as children are learning to read. To this end, we conducted a longitudinal fMRI study with children at 5–9 years of age who received explicit reading instruction for approximately 1.5 years. Children were scanned at relatively short intervals of approximately 2–3 months for a total of up to 6 scanning sessions per child. At each scanning session, children were presented with spoken and written words, pseudowords, and low-level sensory control stimuli. This allowed us to capture intra-individual changes in BOLD activity (as a proxy for local neuronal activity) in response to stimuli that differed in their sensory, phonological, and semantic content. To overcome the bias towards study participants from the Global North and alphabetic writing systems, we worked with children from socio-economically disadvantaged backgrounds in India who received reading instruction in their native language (Hindi) and alphasyllabic writing system (Devanagari). Our hypotheses were as follows:

- We expected that BOLD activity in response to written (pseudo-)words would increase as chilren are learning to read, either in a linear or quadratic (inverted "u"-shaped) pattern. This was based on previous findings (e.g., Dehaene-Lambertz et al., 2018) demonstrating the quick emergence of word selectivity in higher-level visual cortex.
- We expected that in audio-visual integration areas (especially the pSTS and vOT cortex), multi-voxel BOLD activity patterns would become more similar between written and spoken (pseudo-)words as children are learning to read. This was based on the intuition that only over the course of the study, children would become able to access the phonological and semantic content of written words.
- We expected that multi-voxel BOLD activity patterns for written (pseudo-)words would become more "stable," i.e., show less session-to-session variation as children are learning to read. This was

[5]Though the tightness/reliability of this correpsondence differing between relatively shallow (transparent) orthographies such as Spanish, Italian, or German, and relatively deep (opaque) orthographies such as English and French

[6]Unfortunately, most studies implicitly claim generalizability/universitality of their findings by not explicitly mentioning the writing system in the title, abstract, or conclusion of the paper.

based on the intuition that acitvity in reading-related areas should become more finely tuned and less noisy towards written words as children become able to decode and comprehend them.

Methods

Participants

We initially recruited a total of 32 children from the same village in the region of Uttar Pradesh, India, to participate in the reading intervention. In a neighboring village, we recruited an additional 25 children to participate in a mathematics intervention, serving as an active control group. However, data from these children was not analyzed for the purpose of the present manuscript. The two villages were selected in cooperation with a local non-governmental organization based on there being little to no access to primary school education due to geographic and socio-demographic constraints. All recruited children had to fulfil the following inclusion criteria: (1) age between 5 and 9 years at the beginning of the study, (2) not being able to decode letters and/or read words, (3) not attending school, (4) not fulfilling any contraindication for MRI scanning (e.g., no relevant implants or medication), (5) no hearing and/or vision impairment, (6) no language impairments, and (7) no attention deficits. The initial sample size of approximately 30 children per group was determined using an a priori power analysis based on a previous study (Dehaene-Lambertz et al., 2018; see Supplementary Methods for details). From the 32 children initially recruited for the reading intervention group, 17 children were excluded because they did not participate in at least two scanning sessions or because they did not meet our cutoff criterion for head motion (framewise displacement greater than 2.4 mm in less than 10% of fMRI volumes) in at least two scanning sessions. Therefore, the final sample size was $N = 15$ children. Of those, 9 identified as female and 6 identified as male. The mean age at the beginning of the study was 7.13 years (SD = 1.25 years, min = 5 years, max = 9 years).

Study design

The children in this longitudinal study completed regular MRI scanning sessions while participating in a structured reading program (or mathematics program for the control group, not shown in this manuscript). The first scanning session took place at the beginning of the program and the next scanning sessions (mean = 5.07 sessions per child, SD = 1.53 sessions, min = 2 sessions, max = 6 sessions) were spaced at intervals of approximately 2–3 months (mean = 79.9 days between sessions, SD = 35.8 days, min = 20 days, max = 182 days; see Figure 1). Each scanning session consisted of one localizer scan, one functional MRI run (see experimental design and scanning parameters below), one structural MRI scan (see scanning parameters below), and a series of standardized behavioral tests outside of the scanner (see behavioral testing below).

The structured intervention for the reading group focused on phonics (i.e., reading and writing the 46 primary Devanagari characters and their correspondence to sublexical consonant–vowel combinations), word decoding (i.e., reading and writing monosyllabic and more complex words), and sentence reading. The intervention was carried out by a local teacher and involved 2–4 hours of schooling on 5 days per week. Attendance of the classes was checked but not strictly enforced (mean = 3.74 days attended per week).

Experimental design

At each MRI scanning session, children performed an in-scanner language task with a block design where they were presented with short blocks of stimuli from six different conditions (visual words, auditory words, visual pseudowords, auditory pseudowords, visual low-level controls, and auditory low-level controls). All words were nouns of masculine grammatical gender, consisting of one or two syllables and three to six phonemes, and belonging to the same semantic category (animals). Pseudowords were generated from these words by replacing the initial consonant and vowel of the word. The substituted consonants were within the same articulatory place as the original consonants and the substituted graphemes matched the shape of the original graphemes as closely as possible. The low-level visual controls were false fonts that were created by rearranging the line segments of each grapheme of a word while preserving the position of the graphemes. The low-level auditory controls were created by spectrally rotating and noise-vocoding the spoken words. This was achieved by low-pass filtering the speech signal, inverting its spectrum around a center frequency of 2 kHz, dividing the speech signal into two logarithmically spaced frequency bands, extracting the amplitude envelope in each frequency band, using this envelope to module noise in the same frequency band, and recombining the frequency bands. All visual stimuli were presented in the middle of the screen in black font on a white background. The screen was placed behind the scanner bore and projected to the participant via a mirror mounted inside of the scanner. All auditory stimuli were recorded by a male native Hindi speaker.

In each of 108 blocks (18 per condition), 6 stimuli from the same condition were randomly presented with a duration of 1 s each. The order of blocks was random and not optimized for design efficiency. Between subsequent blocks, there was a random pause of 2.55, 3.82, or 5.09 s (equaling 1, 1.5, or 2 times the repetition time [TR]), during which a black fixation cross was presented in the middle of the screen. The total duration of the experiment was 17:40 min. To keep children attentive, they were asked to perform a simple target detection task. For this purpose, a target stimulus was inserted at a random location in 36 out of the 108 blocks. For visual blocks, this was the photograph of the face of a children's movie character, and for auditory blocks a short snippet of child-friendly human laughter. Children were asked to press a button on a MR-compatible button box whenever they saw or heard the target stimulus. They received auditory feedback in the form of a positive sound (after pressing the button when a target stimulus had appeared) or a negative sound (after not pressing the button when a target stimulus had appeared or after pressing the button when no target stimulus had appeared). The experiment was programmed and presented using PsychoPy (Version 2021) in Python. The code and stimuli for the experiment can be accessed via https://github.com/SkeideLab/SLANG-experiment/tree/main/scanner.

Standardized behavioral testing

Bla bla blub.

Scanning parameters

All scanning was conducted on a GE SIGNA Architect 3T MRI machine with a 48 channel head coil. After a short head localizer scan, there was one functional run, during which children performed the language experiment described above, and one structural run, during which children watched a TV episode or video of their choice.

The functional run was implemented using a gradient echo (GR) echo planar imaging (EPI) sequence with the following parameters: TE = 35 ms, TR = 2.547 s, flip angle = 88°, number of volumes = 420, field of view = 19.2 cm, in-plane matrix size = 80 × 80 voxels, slice thickness = 2.4 mm, gap between slices = 0.2 mm, voxel size = 2.4 × 2.4 × (2.4 + 0.2) mm, slice orientation = axial, phase encoding direction = anterior/posterior, slice order = interleaved/ascending, multiband acceleration (GE HyperBand) factor = 4. Slices covered the whole brain including the cerebellum. We did not collect any field maps and did not correct for potential inhomogeneity or spatial distortion.

The structural run was implemented using a T1-weighted MPRAGE sequence with the following parameters: TE = 3.188 ms, TR = 2.30568 s, TI = 900 ms, flip angle = 8°, field of view = 22.4 cm, in-plane matrix size = 256 × 256 voxels, slice thickness = 0.9 mm, no gap between slices, voxel size = 0.875 × 0.875 × 0.9 mm, slice orientation = axial, phase encoding direction = right/left.

Preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 24.0.1 (Esteban et al., 2018, RRID:SCR_016216; Esteban et al., 2019), which is based on *Nipype* 1.8.6 (Gorgolewski et al., 2011; Gorgolewski et al., 2018, RRID:SCR_002502).

Anatomical data preprocessing. A total of 2–6 T1-weighted (T1w) images were found within the input BIDS dataset. Each T1w image was corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.5.1 (Avants et al., 2008, RRID:SCR_004757). The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL, RRID:SCR_002823, Zhang et al., 2001). An anatomical T1w-reference map was computed after registration of 2–6 T1w images (after INU-correction) using `mri_robust_template` (FreeSurfer 7.3.2, Reuter et al., 2010). Brain surfaces were reconstructed using `recon-all` (FreeSurfer 7.3.2, RRID:SCR_001847, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.5.1), using brain-extracted versions of both T1w reference and the T1w template. The following template was were selected for spatial normalization and accessed with *TemplateFlow* (24.2.0, Ciric et al., 2022): *ICBM 152 Nonlinear Asymmetrical template version 2009c* (Fonov et al., 2009, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

Functional data preprocessing. For each of the 2–6 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume was generated, using a custom methodology of *fMRIPrep*, for use in head motion correction. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL, Jenkinson et al., 2002). The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve & Fischl, 2009). Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al. (2014)) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al. (2002)). FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contain a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's *aseg* segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance timeseries are calculated by means of principal components analysis of the signal found within a thin band (*crown*) of voxels around the edge of the brain, as proposed by (Patriat et al., 2017). The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): *fsnative*, *fsaverage5*. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `nitransforms`, configured with cubic B-spline interpolation. Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.10.4 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation(https://fmriprep.readthedocs.io/en/latest/workflows.html).

Additional preprocessing. After running *fMRIPrep*, the preprocessed BOLD fMRI time series data was spatially smoothed with a Gaussian kernel (FWHM = 5.0 mm) and masked using the whole-brain mask computed per-participant by *fMRIPrep*.

Session-level analysis

Separately for each participant and session, we modeled the preprocessed BOLD fMRI time series data using a mass-univariate general linear model (GLM) implemented in Nilearn (Version 0.10.4; Abraham et al., 2014) for Python (Version 3.9.19; Van Rossum & Drake, 2009). At each voxel, the BOLD time series (420 time points) was predicted using a design matrix with the following columns:

- A constant term (1 column).
- One task regressor for each of the six experimental conditions, created by convolving the block design (condition on/off) with the canonical "SPM" hemodynamic response function (HRF) implemented in Nilearn (6 columns). These were our regressors of interest.
- The second and third derivatives of each of the task regressors (12 columns), to capture age- and participant-specific deviations from the canonical HRF.
- The head motions parameters (translations and rotations in three directions) estimated by *fMRIPrep* during head motion correction (6 columns).
- Cosine regressors to high-pass filter the data at 128 s ($\approx$ 0.008 Hz), removing slow-frequency scanner drifts (15 columns).
- The top six anatomical CompCor (Behzadi et al., 2007) components estimated by *fMRIPrep* (6 columns).
- Spike regressors for each non-steady state outlier volume at the beginning of the scan as flagged by *fMRIPrep* (XXX–XXX columns, mean = XXX columns).
- Spike regressors for each high-motion outlier volume, defined as framewise displacement > 0.5 mm and flagged by *fMRIPrep* (XXX–XXX columns, mean = XXX columns).

For details on how these regressors were computed, see the "Outputs/Confounds" section in *fMRIPrep*'s documentation (https://fmriprep.org/en/stable/outputs.html#confounds).

From the fitted model, we computed an effect size map ("beta map") for each of the following contrasts, always reflecting the change in BOLD activation (in arbitrary units) between two experimental conditions:

- Auditory low-level vs. baseline
- Auditory pseudowords vs. baseline
- Auditory pseudowords vs. low-level
- Auditory words vs. baseline
- Auditory words vs. low-level
- Auditory words vs. pseudowords
- Visual low-level vs. baseline
- Visual pseudowords vs. baseline
- Visual pseudowords vs. low-level
- Visual words vs. baseline
- Visual words vs. low-level
- Visual words vs. pseudowords

We refer to all contrasts involving the baseline as "baseline contrasts," which capture the difference in BOLD activity between each different experimental condition and the fixation baseline period between experimental blocks. Therefore, these contrasts capture not only BOLD activity related to linguistic (phonological and semantic) processing but also BOLD activity related to low-level sensory processing (e.g., visual and auditory processing of the stimuli). We refer to all other contrasts as "experimental contrasts," which capture the difference in BOLD activity between two different experimental conditions. The experimental contrasts comparing pseudowords to low-level controls capture BOLD activity related to phonological processing, since pseudowords but not low-level controls contain phonological information, while both do not contain any semantic information. The experimental contrasts comparing words to pseudowords capture BOLD activity related to semantic processing, since words but not pseudowords contain semantic information, while both contain phonological information. The experimental contrasts comparing words to low-level controls capture BOLD activity related to both phonological and semantic processing, since words but not low-level controls contain both phonological and semantic information.

Group-level analysis

BOLD activity amplitude. To estimate reading related changes in BOLD activity amplitude, we fitted the beta maps from all participants and sessions using a linear mixed-effects model, separately for each contrast (see above). The dependent variable was the BOLD activation amplitude for a given participant and session, and the predictors were (1) a fixed intercept, implemented as a column of "1"s and reflecting the BOLD activity at the beginning of the study, (2) a fixed effect for linear time, implemented as the number of days elapsed since the beginning of the study and reflecting the linear change in BOLD activity due to the reading instruction, (3) a fixed effect for quadratic time, implemented as the square of linear time and reflecting the nonlinear ("u"-shaped or inverted "u"-shaped) change in BOLD activity due to the reading instruction, (4) a random intercept, reflecting individual participant's deviation from the global BOLD activity at the beginning of the study, and random slopes for (5) linear and (6) quadratic time, reflecting individual participant's deviations in the linear and nonlinear changes in BOLD activity due to the reading instruction. As is typical in frequentist linear mixed models, one parameter was estimated for each fixed effect (intercept, linear time, and quadratic time) and for each random effect (the standard deviation of the random intercepts and slopes for linear and quadratic time), as well as three correlation parameters between the three pairs of random effects. The mixed models (one for each contrast) were fitted separately at each voxel inside the brain mask in a mass-univariate fashion. Model fitting was performed in Julia (Version 1.10.4; Bezanson et al., 2017) using the MixedModels package (Version 4.25.3; Bates et al., 2024).

To correct for multiple comparisons across the XXX voxels inside the brain mask, we used the parametric cluster correction algorithm suggested by Cox et al. (2017a) and Cox et al. (2017b). Specifically, we first estimated the spatial smoothness of noise in our dataset by extracting and storing the residual time series from the session-level models (see above), separately for each participant and session. We then used the `3dFWHMx` function (with the `-acf` option) in AFNI (Version 24.2.01; Cox, 1996) to estimate a mixed Gaussian/mono-exponential spatial autocorrelation function with three parameters (Cox et al., 2017b; Cox et al., 2017a; see also https://afni.nimh.nih.gov/pub/dist/edu/data/CD.expanded/afni_handouts/afni07_ETAC.pdf). We averaged each of these three parameters across participants and sessions to obtain a single spatial autocorrelation function for the entire dataset. This function was then fed into the `3dClustSim` function in AFNI to generate novel noise-only maps and estimate a null distribution of cluster sizes. Using a cluster-forming voxel-level threshold of $p < .001$ and 10,000 iterations, this resulted in a final cluster-level extent threshold of XXX voxels to control the whole-brain family-wise error (FWE) rate at $p < .05$. We therefore deemed spatial clusters of BOLD activation statistically significant if they were larger than XXX voxels. To form clusters, neighboring voxels had to pass the cluster-forming voxel level threshold of $p < .001$ and touch each other with their faces (not just with their edges or nodes; the default "NN1" method in AFNI).

BOLD activity patterns. We also estimated reading related changes in the between-condition similarity and within-condition stability of BOLD activity patterns using multivariate analysis methods.

First, we investigated if the reading intervention made written word activity patterns more similar to those for spoken word activity patterns. For this, we used the beta maps from these two conditions and extracted the betas for different regions of interest (ROI; defined below). We then computed the linear correlation between these vectors and entered these (one value per subject and session) into a linear mixed-effects model, separately for each ROI. We specified the linear mixed-effects model in the same way as described above, with fixed and random effects for the intercept, the linear effect of time, and the quadratic effect of time. We repeated the same analysis for the correlation of activity patterns between written and spoken pseudowords. However, it is important to note at this point that with our block design, we are not comparing the activity patterns in response to individual items (e.g., the written word "monkey" and the spoken word "monkey"), but rather the general patterns of activity when processing written and spoken words. Though arguably more meaningful, the former comparison is not possible as our trials within each block were presented faster (1 s) than our TR (2.647 s), and we were therefore unable to obtain a beta map for each individual item.

Second, we investigated if the reading intervention made written word activity more "stable," i.e., more consistent across repeated presentations of written words within the same session. For this, we re-estimated the session-level beta maps as described above but with separate task regressors for each block (108 columns) instead of each condition (6 columns). We then took the beta map for each written word block (18 maps) and extracted the betas for different ROIs (defined below). We then computed the linear correlation between each pair of blocks (153 pairs) and averaged these to obtain a single stability (correlation) value for each participant and session. We entered these values into a linear-mixed effects model, separately for each ROI and specified in the same way as described above. We repeated the same analysis for the stability of activity patterns for written pseudowords. We also performed this analysis for spoken words and spoken pseudowords but reported the results in the appendix as they do not directly pertain to our hypotheses. F or all multivariate analyses, we used the same set of anatomically and functionally defined regions of interest. The anatomically defined regions of interest were the posterior superior temporal sulcus (pSTS), defined as the union of regions STSda, STSdp, STSvp, and STSva from the Glasser et al. (2016) atlas, and the ventral occipto-temporal cortex (vOT), defined as the union of regions V8, FFC, and VVC from the Glasser et al. (2016) atlas. The functionally defined regions were different for each condition: For the similarity between written and spoken words, we used the visual clusters from the "Written words vs. baseline" contrast and the auditory clusters from the "Spoken words vs. baseline" clusters. ikewise, for the similarity between written and spoken pseudowords, we used the visual clusters from the "Written pseudowords vs. baseline" contrast and the auditory clusters from the "Spoken pseudowords vs. baseline" clusters. For the within-condition pattern stability analysis, we used the visual or auditory clusters from that specific condition (e.g., for the stability analysis for written words, we used the visual clusters from the "Written words vs. baseline" contrast).

Results

BOLD activity amplitude

Auditory baseline contrasts. At the beginning of the study (intercept), spoken low-level controls, spoken pseudowords, and spoken words elicited BOLD activity in the left and right auditory cortex (superior temporal gyrus) and in the left inferior frontal gyrus (see orange clusters in Figures 1A, 2A, and 3A), as well as a few a BOLD deactivations in parietal and occipital areas (see blue clusters in Figures 1A, 2A, and 3A). For spoken low-level controls, there was no linear change in BOLD activity over the course of the study (see Figure 1C) but positive quadratic (i.e., "u"-shaped) change in BOLD activity three small clusters in the right middle/inferior temporal lobe and in the right parietal lobe (see orange clusters in Figure 1E). For spoken pseudowords, there was both negative linear and positive quadratic (i.e., "u"-shaped) change in one small cluster in the right parietal lobe (see blue cluster in Figure 2C and orange cluster in Figure 2E). For spoken words, there was positive linear change in BOLD activity over the course of the study in one small cluster in the left inferior parietal lobe (see orange cluster in Figure 3C), as well as positive quadratic change (i.e., "u"-shaped) change in one small cluster in the right superior parietal lobe (see orange cluster in Figure 3E).

Visual baseline contrasts. At the beginning of the study (intercept), written low-level controls, written pseudowords, and written words elicited BOLD activity in the left right occipital and inferior temporal cortices (see orange clusters in Figures 1B, 2B, and 3B) as well as BOLD deactivations in the left and right medial occiptial lobes (see blue clusters in Figures 1B, 2B, and 3B). For written low-level controls, there was no linear change in BOLD activity over the course of the study (see Figure 1D) but positive quadratic (i.e., "u"-shaped) change in BOLD activity two small clusters in the right occipital lobe (see orange clusters in Figure 1F). For written pseudowords, there was negative linear change in one small cluster in the left inferior frontal lobe (see left frontal blue cluster in Figure 2D) and both negative linear change and positive quadratic (i.e., "u"-shaped) change in one small cluster in the right occipital lobe (see posterior blue cluster in Figure 2D and orange cluster in Figure 2F). For written words, there was both negative linear and positive quadratic (i.e., "u"-shaped) change in one small cluster in the right occipital lobe (see small blue cluster in Figure 3D and posterior orange cluster in Figure 3F). Additionally, there was positive quadratic (i.e., "u"-shaped) change in one small cluster in the right parietal lobe (see anterior orange cluster in Figure 3F).

Auditory experimental contrasts. At the beginning of the study (intercept), the difference between spoken pseudowords and spoken low-level controls elicited widespread BOLD activity mainly in the left and right auditory cortex (superior temporal gyrus; see large orange clusters in Figure 4A). There was negative linear change in BOLD activity over the course of the study in one small cluster in the right anterior inferior parietal lobe (see small blue cluster in Figure 4C). There was positive quadratic change in BOLD activity over the course of the study in one small area in the left superior parietal lobe (see small blue orange in Figure 4E).

At the beginning of the study (intercept), the difference between spoken words and spoken low-level controls elicited widespread BOLD activity mainly in the left and right auditory cortex (superior temporal gyrus; see large orange clusters in Figures 5A). There was no evidence for linear or quadratic change over the course of the study in any brain areas (see Figure 5C and E).

The difference between spoken words and spoken pseudowords did not elicity any reliable BOLD activity at the beginning of the study and showed no evidence for change over the course of the study (see Figure 6C and E).

Visual experimental contrasts. At the beginning of the study (intercept), the difference between written pseudowords and written low-level controls elicited BOLD activity in one cluster in the right dorso-lateral prefrontal cortex (see orange cluster in Figure 4B) and BOLD deactivation in one cluster in the right secondary visual cortex (see blue cluster in Figure 4B). There was negative linear change over the course of the study in the right dorso-lateral prefrontal cortex (see blue cluster in Figure 4D).

The difference between written words and written low-level controls did not elicit any reliable BOLD activity at the beginning of the study (see Figure 5B) but there was negative quadratic (inverted "u"-shaped) change over the course of the study in one cluster in the left posterior

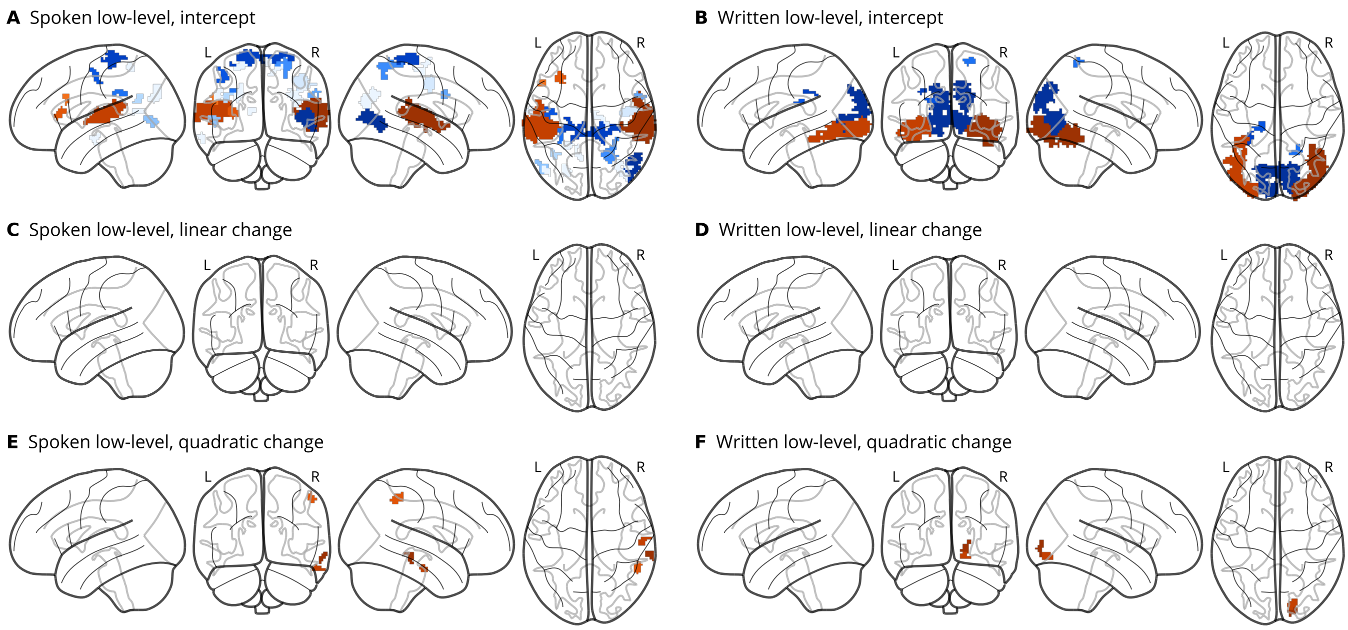

Figure 1. ***BOLD activity in response to low-level sensory controls.*** Panels **A** and **B** show statistically significant clusters (cluster-forming voxel threshold $p < .001$, uncorrected; cluster size threshold $p < .05$, FWE-corrected) for the contrast of low-level sensory control blocks versus fixation baseline at the beginning of the study (intercept; time = 0 months). Panels **C** and **D** show statistically significant clusters for the linear change in BOLD activity for the same contrast over the course of the study and panels **E** and **F** show statistically significant clusters for the quadratic change in BOLD activity for the same contrast over the course of the study. Panels **A**, **C**, and **E** show the auditory modality (noise-vocoded speech versus fixation basline) and panels **B**, **D**, and **F** show the visual modality (false fonts versus fixation baseline). In all panels **A**–**F**, clusters with positive BOLD amplitude (low-level > baseline) are shown in orange and clusters with negative BOLD amplitude (low-level < baseline) are shown in blue. Clusters with higher voxel-level peak statistics are shown in brighter colors. Panels **G** and **H** show individual participants' change in BOLD amplitude over time (colored dots and lines) as well as the best fitting linear mixed model (dashed black line) for the largest significant clusters in panels **E** and **F**, respectively, as indicated by the black asterisk (*) next to the cluster.

medial wall (near the ventral posterior cingulate cortex; see the blue cluster in Figure 5F).

The difference between written words and written pseudowords elicited BOLD deactivation in one cluster in the right dorso-lateral prefrontal cortex (see blue cluster in Figure 6B). There was no evidence for linear or quadratic change over the course of the study (see Figure 6D and F).

BOLD activity patterns

Audio-visual pattern similarity. For all baseline contrasts (low-level vs. baseline, pseudowords vs. baseline, and words vs. baseline) and ROIs, the auditory and visual BOLD response patterns were correlated significantly at the beginning of the study (intercept; all $ps < .001$; see Figures 7–9), except for words in the left spoken word ROI ($p = 0.072$; see Figure 9D). This is expected given that the exact same fixation baseline periods were used as the comparison condition for both the auditory and visual baseline contrasts. There was no evidence for linear change (all $ps > .059$) or quadratic change (all $ps > .128$) in pattern similarity over the course of the study for any contrast pair or ROI.

For all experimental contrasts (pseudowords vs. low-level, words vs. low-level, words vs. pseudowords) and ROIs, the auditory and visual BOLD response patterns were not significantly correlated at the beginning of the study (intercept; all $ps > .098$; see Figures 10–12). Likewise, there was no evidence for linear change (all $ps > .132$) or quadratic change (all $ps > .227$) in pattern similarity over the course of the study for any contrast pair or ROI.

Within-condition pattern stability. For almost all conditions and ROIs, blocks from the same condition were positively correlated at the beginning of the study (intercept; all $ps < .024$; see Figures 13–18), except for spoken low-level controls in the right written low-level controls ROI ($p = 0.074$; see Figure 13E), for spoken pseudowords in the right written pseudoword ROI ($p = 0.074$; see Figure 14E), for spoken words in the right written word ROI ($p = 0.185$; see Figure 15E), and for written words in the left spoken word ROI ($p = 0.061$; see Figure 18D). There was no evidence for linear change over the course of the study for any conditions and ROIs, except for written low-level controls in the left pSTS ROI ($b = -0.0018$, $p = .050$). Likewise, there was no evidence for quadratic change over the course of the study for any conditions and ROIs, except for written low-level controls in the right low-level controls ROI ($b = 0.0007$, $p = .043$). Note that we did not have any *a priori* hypothesis for longitudinal changes in the low-level control conditions but instead expected changes in pattern stability in the pseudoword and word conditions.

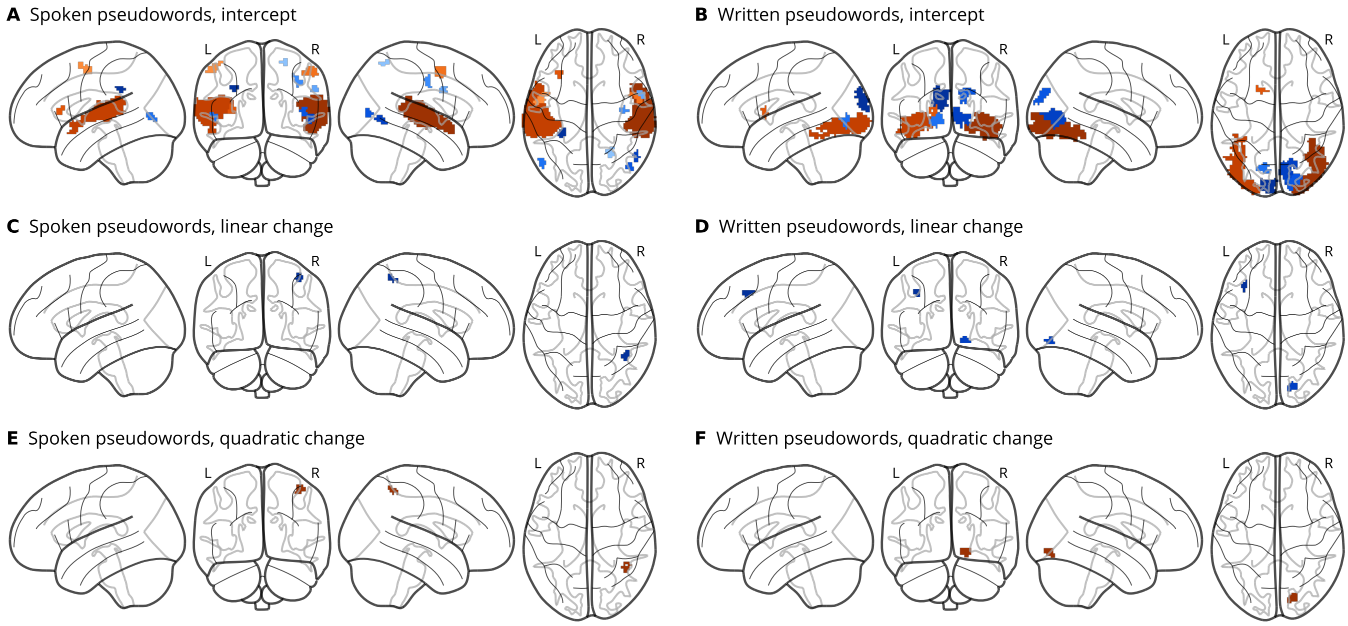

Figure 2. BOLD activity in response to pseudowords. Panels **A** and **B** show statistically significant clusters (cluster-forming voxel threshold $p < .001$, uncorrected; cluster size threshold $p < .05$, FWE-corrected) for the contrast of pseudoword blocks versus fixation baseline at the beginning of the study (intercept; time = 0 months). Panels **C** and **D** show statistically significant clusters for the linear change in BOLD activity for the same contrast over the course of the study and panels **E** and **F** show statistically significant clusters for the quadratic change in BOLD activity for the same contrast over the course of the study. Panels **A**, **C**, and **E** show the auditory modality (spoken pseudowords versus fixation basline) and panels **B**, **D**, and **F** show the visual modality (written pseudowords versus fixation baseline). In all panels **A**–**F**, clusters with positive BOLD amplitude (pseudowords > baseline) are shown in orange and clusters with negative BOLD amplitude (pseudowords < baseline) are shown in blue. Clusters with higher voxel-level peak statistics are shown in brighter colors. Panels **G** and **H** show individual participants' change in BOLD amplitude over time (colored dots and lines) as well as the best fitting linear mixed model (dashed black line) for the largest significant clusters in panels **E** and **F**, respectively, as indicated by the black asterisk (*) next to the cluster.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 1–10. https://doi.org/10.3389/fninf.2014.00014

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Bates, D., Alday, P., Kleinschmidt, D., Calderón, J. B. S., Zhan, L., Noack, A., Bouchet-Valat, M., Arslan, A., Kelman, T., Baldassari, A., Ehinger, B., Karrasch, D., Saba, E., Quinn, J., Hatherly, M., Piibeleht, M., Mogensen, P. K., Babayan, S., Holy, T., … Nazarathy, Y. (2024). *JuliaStats/MixedModels.jl: V4.25.3*. Zenodo. https://doi.org/10.5281/zenodo.13174525

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Blau, V., Reithler, J., van Atteveldt, N., Seitz, J., Gerretsen, P., Goebel, R., & Blomert, L. (2010). Deviant processing of letters and speech sounds as proximate cause of reading failure: A functional magnetic resonance imaging study of dyslexic children. *Brain*, *133*(3), 868–879. https://doi.org/dm3tgr

Brem, S., Maurer, U., Kronbichler, M., Schurz, M., Richlan, F., Blau, V., Reithler, J., van der Mark, S., Schulz, E., Bucher, K., Moll, K., Landerl, K., Martin, E., Goebel, R., Schulte-Körne, G., Blomert, L., Wimmer, H., & Brandeis, D. (2020). Visual word form processing deficits driven by severity of reading impairments in children with developmental dyslexia. *Scientific Reports*, *10*(1), 18728. https://doi.org/10.1038/s41598-020-75111-8

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

Ciric, R., Thompson, W. H., Lorenz, R., Goncalves, M., MacNicol, E., Markiewicz, C. J., Halchenko, Y. O., Ghosh, S. S., Gorgolewski, K. J., Poldrack, R. A., & Esteban, O. (2022). TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature*

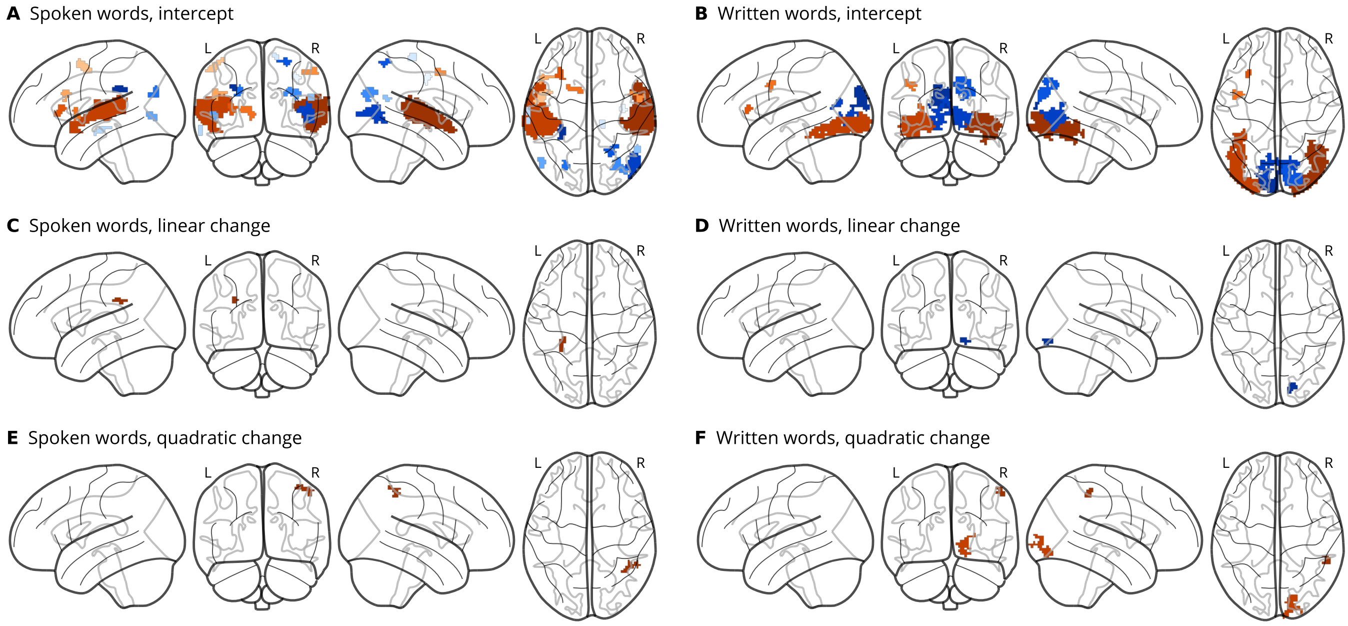

Figure 3. ***BOLD activity in response to words.*** Panels **A** and **B** show statistically significant clusters (cluster-forming voxel threshold $p < .001$, uncorrected; cluster size threshold $p < .05$, FWE-corrected) for the contrast of word blocks versus fixation baseline at the beginning of the study (intercept; time = 0 months). Panels **C** and **D** show statistically significant clusters for the linear change in BOLD activity for the same contrast over the course of the study and panels **E** and **F** show statistically significant clusters for the quadratic change in BOLD activity for the same contrast over the course of the study. Panels **A**, **C**, and **E** show the auditory modality (spoken words versus fixation basline) and panels **B**, **D**, and **F** show the visual modality (written words versus fixation baseline). In all panels **A**–**F**, clusters with positive BOLD amplitude (words > baseline) are shown in orange and clusters with negative BOLD amplitude (words < baseline) are shown in blue. Clusters with higher voxel-level peak statistics are shown in brighter colors. Panels **G** and **H** show individual participants' change in BOLD amplitude over time (colored dots and lines) as well as the best fitting linear mixed model (dashed black line) for the largest significant clusters in panels **E** and **F**, respectively, as indicated by the black asterisk (*) next to the cluster.

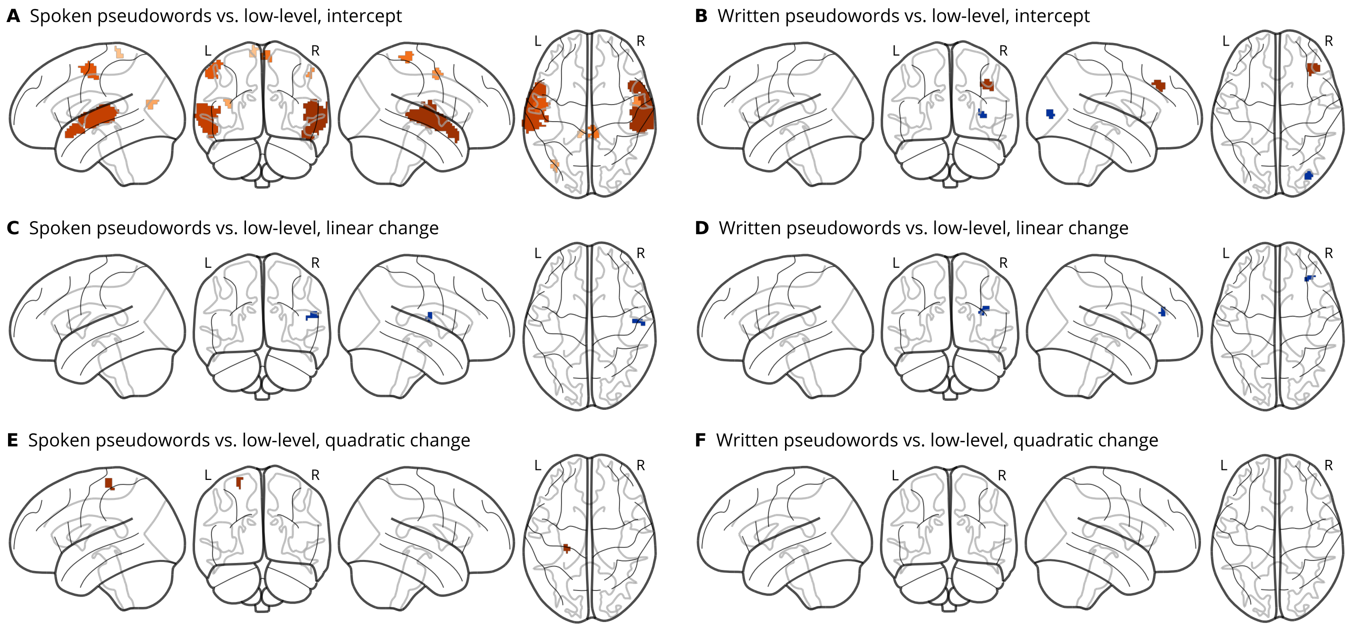

Figure 4. ***BOLD activity in response to pseudowords versus low-level controls.*** Panels **A** and **B** show statistically significant clusters (cluster-forming voxel threshold $p < .001$, uncorrected; cluster size threshold $p < .05$, FWE-corrected) for the contrast of pseudoword blocks versus low-level control blocks at the beginning of the study (intercept; time = 0 months). Panels **C** and **D** show statistically significant clusters for the linear change in BOLD activity for the same contrast over the course of the study and panels **E** and **F** show statistically significant clusters for the quadratic change in BOLD activity for the same contrast over the course of the study. Panels **A**, **C**, and **E** show the auditory modality (spoken pseudowords versus noise-vocoded speech) and panels **B**, **D**, and **F** show the visual modality (written pseudowords versus false fonts). In all panels **A**–**F**, clusters with positive BOLD amplitude (pseudowords > low-level) are shown in orange and clusters with negative BOLD amplitude (pseudowords < low-level) are shown in blue. Clusters with higher voxel-level peak statistics are shown in brighter colors. Panels **G** and **H** show individual participants' change in BOLD amplitude over time (colored dots and lines) as well as the best fitting linear mixed model (dashed black line) for the largest significant clusters in panels **E** and **F**, respectively, as indicated by the black asterisk (*) next to the cluster.

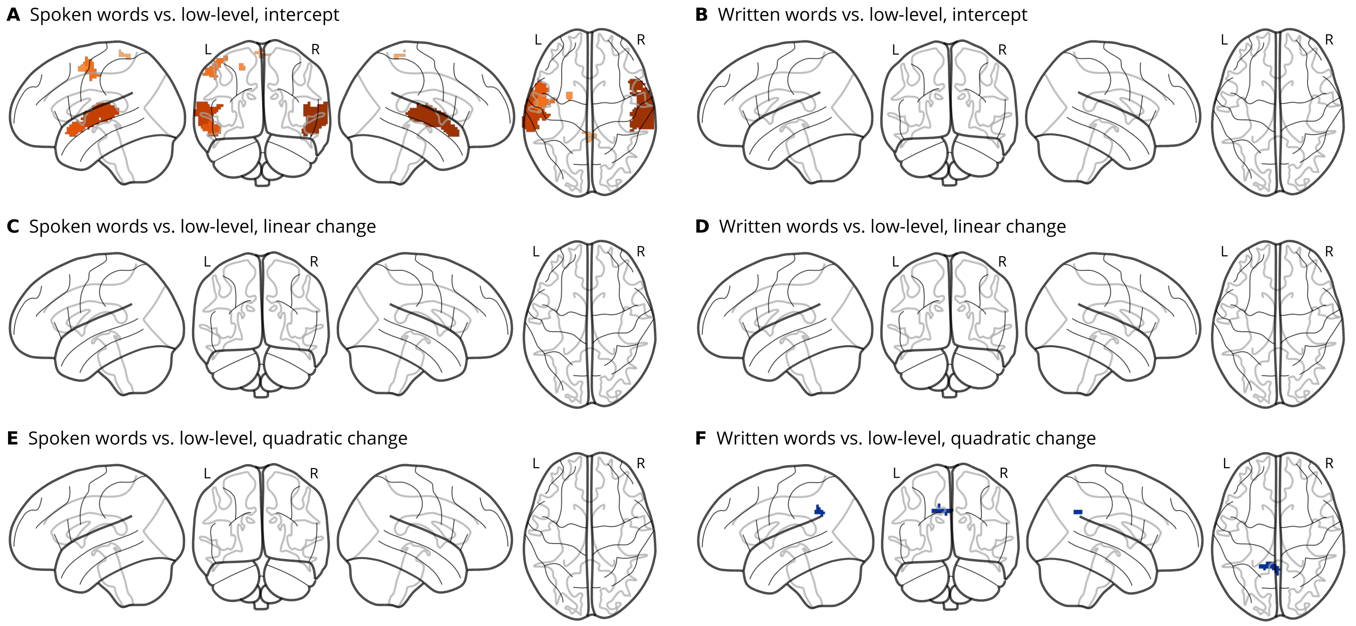

Figure 5. ***BOLD activity in response to words versus low-level controls.*** Panels **A** and **B** show statistically significant clusters (cluster-forming voxel threshold $p < .001$, uncorrected; cluster size threshold $p < .05$, FWE-corrected) for the contrast of word blocks versus low-level control blocks at the beginning of the study (intercept; time = 0 months). Panels **C** and **D** show statistically significant clusters for the linear change in BOLD activity for the same contrast over the course of the study and panels **E** and **F** show statistically significant clusters for the quadratic change in BOLD activity for the same contrast over the course of the study. Panels **A**, **C**, and **E** show the auditory modality (spoken words versus noise-vocoded speech) and panels **B**, **D**, and **F** show the visual modality (written words versus false fonts). In all panels **A**–**F**, clusters with positive BOLD amplitude (words > low-level) are shown in orange and clusters with negative BOLD amplitude (words < low-level) are shown in blue. Clusters with higher voxel-level peak statistics are shown in brighter colors. Panel **G** shows individual participants' change in BOLD amplitude over time (colored dots and lines) as well as the best fitting linear mixed model (dashed black line) for the largest significant cluster in panel **F**, as indicated by the black asterisk (*) next to the cluster.

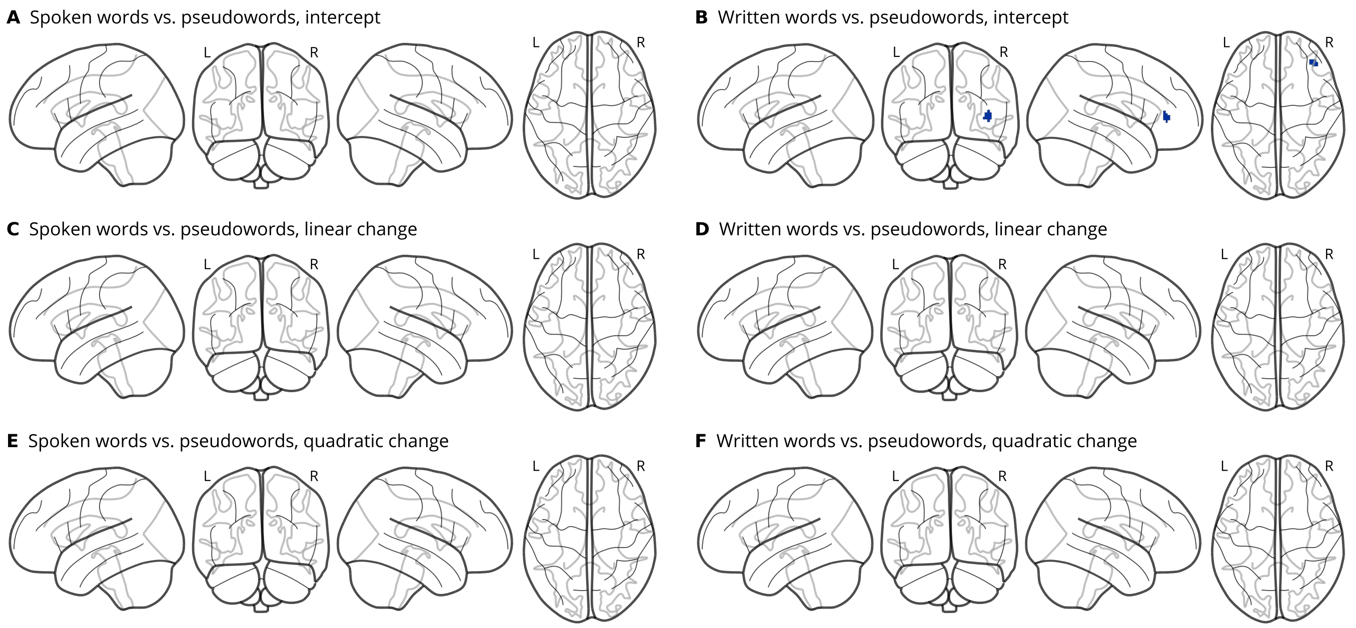

Figure 6. ***BOLD activity in response to words versus pseudowords.*** Panels **A** and **B** show statistically significant clusters (cluster-forming voxel threshold $p < .001$, uncorrected; cluster size threshold $p < .05$, FWE-corrected) for the contrast of word blocks versus pseudoword blocks at the beginning of the study (intercept; time = 0 months). Panels **C** and **D** show statistically significant clusters for the linear change in BOLD activity for the same contrast over the course of the study and panels **E** and **F** show statistically significant clusters for the quadratic change in BOLD activity for the same contrast over the course of the study. Panels **A**, **C**, and **E** show the auditory modality (spoken words versus spoken pseudowords) and panels **B**, **D**, and **F** show the visual modality (written words versus written pseudowords). In all panels **A**–**F**, clusters with positive BOLD amplitude (words > pseudowords) are shown in orange and clusters with negative BOLD amplitude (words < pseudowords) are shown in blue. Clusters with higher voxel-level peak statistics are shown in brighter colors.

Methods, *19*, 1568–1571. https://doi.org/10.1038/s41592-022-01681-2

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, *123*(2), 291–307. https://doi.org/10.1093/brain/123.2.291

Coltheart, M. (2014). The neuronal recycling hypothesis for reading and the question of reading universals. *Mind & Language*, *29*(3), 255–269. https://doi.org/10.1111/mila.12049

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance

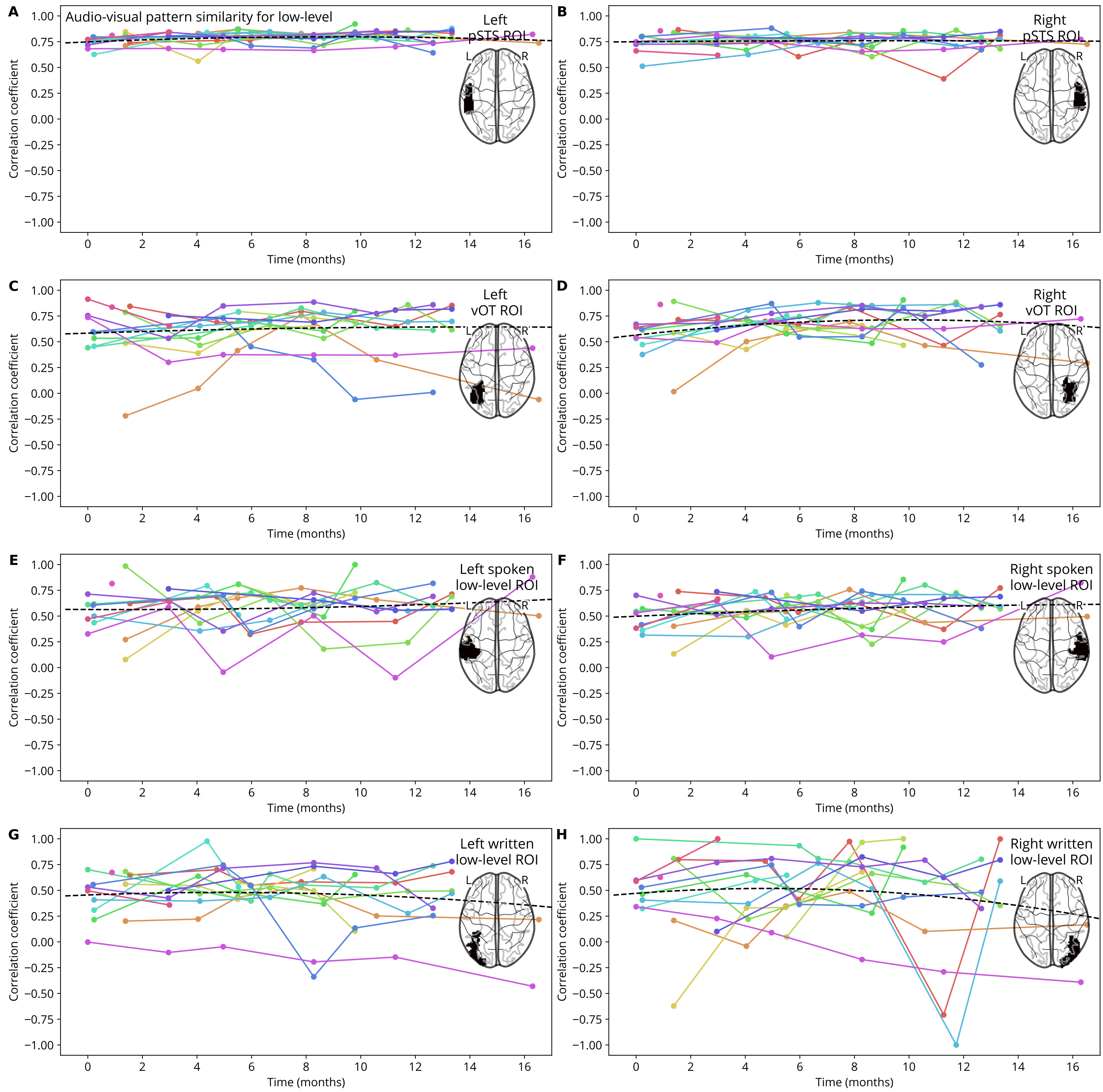

Figure 7. *Audio-visual pattern similarity for low-level controls.*

neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., & Taylor, P. A. (2017a). fMRI clustering and false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(17), E3370–E3371. https://doi.org/10.1073/pnas.1614961114

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., & Taylor, P. A. (2017b). fMRI clustering in AFNI: False-positive rates redux. *Brain Connectivity*, *7*(3), 152–171. https://doi.org/10.1089/brain.2016.0475

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398. https://doi.org/10.1016/j.neuron.2007.10.004

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

Dehaene, S., Cohen, L., Morais, J., & Kolinsky, R. (2015). Illiterate to literate: Behavioural and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, *16*(4), 234–244. https://doi.org/10.1038/nrn3924

Dehaene, S., Le Clec'H, G., Poline, J.-B., Le Bihan, D., & Cohen, L. (2002). The visual word form area: A prelexical representation of visual words in the fusiform gyrus. *Neuroreport*, *13*(3), 321–325. https://doi.org/10.1097/00001756-200203040-00015

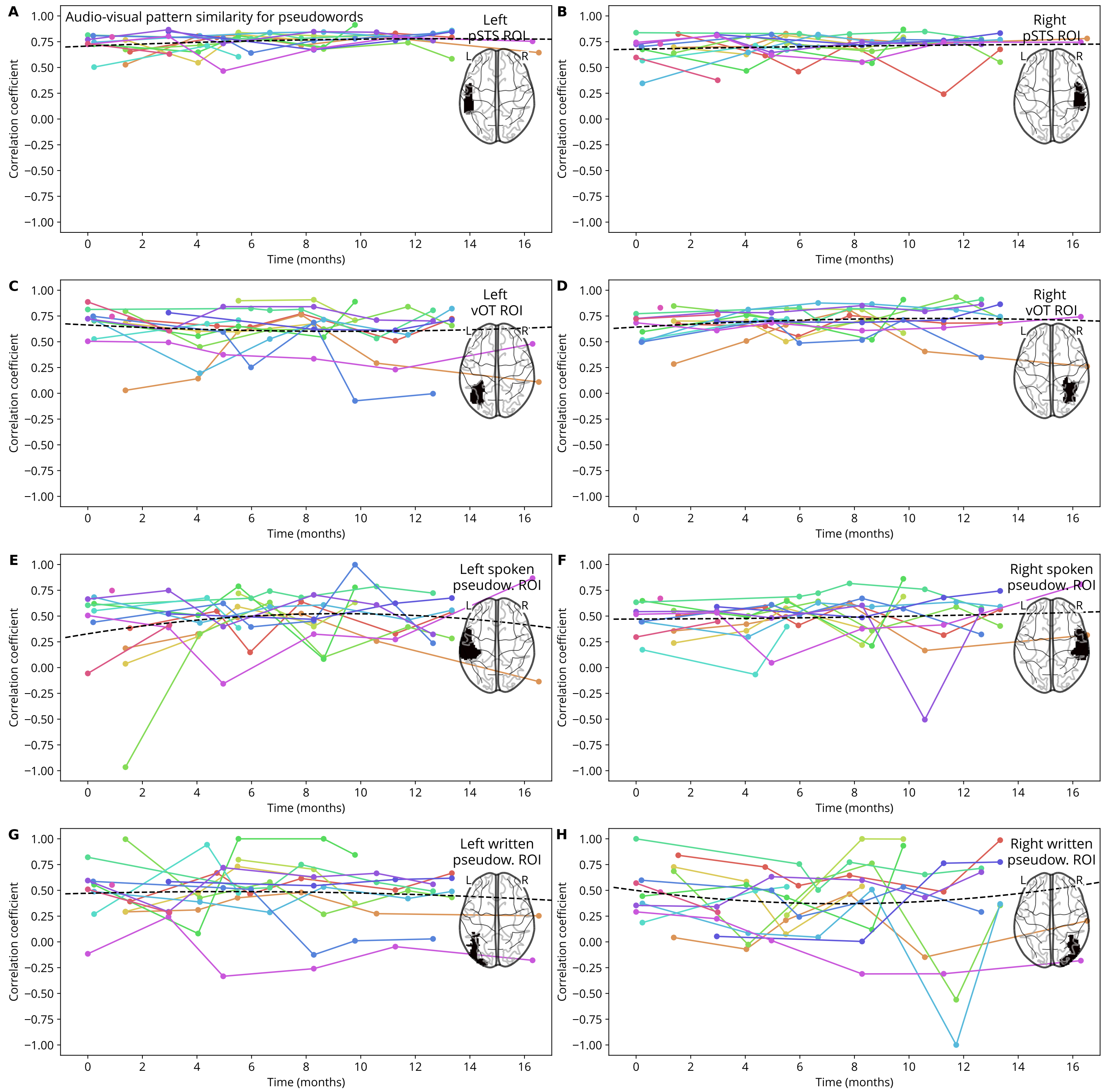

Figure 8. Audio-visual pattern similarity for pseudowords.

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLOS Biology*, *16*(3), 1–34. https://doi.org/10.1371/journal.pbio.2004103

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, *9*(2), 167–188. https://doi.org/fkh923

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, M., James D. andGoncalves, DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., & Gorgolewski, K. J. (2018). fMRIPrep. *Software*. https://doi.org/10.5281/zenodo.852659

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47, Supplement 1*, S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Forseth, K. J., Kadipasaoglu, C. M., Conner, C. R., Hickok, G., Knight, R. T., & Tandon, N. (2018). A lexical semantic hub for heteromodal naming in middle fusiform gyrus. *Brain*, *141*(7), 2112–2126. https://doi.org/10.1093/brain/awy120

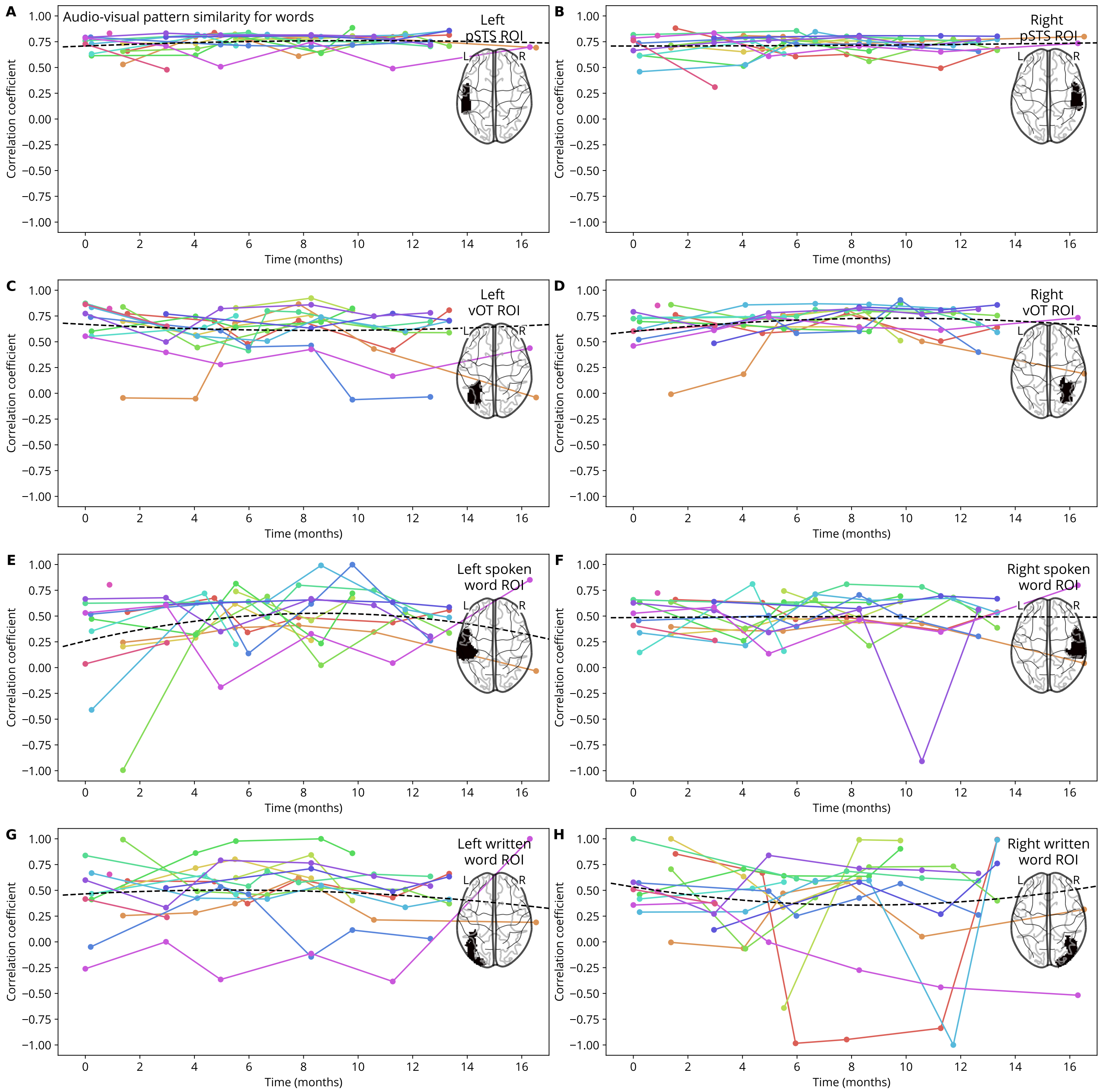

Figure 9. Audio-visual pattern similarity for words.

Frith, U. (1986). A developmental framework for developmental dyslexia. *Annals of Dyslexia*, *36*(1), 67–81. https://doi.org/10.1007/BF02648022

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences*, *35*(5), 263–279. https://doi.org/10.1017/S0140525X11001841

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Gorgolewski, K. J., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., … Ghosh, S. (2018). Nipype. *Software*. https://doi.org/10.5281/zenodo.596855

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Hasenäcker, J., Schröter, P., & Schroeder, S. (2017). Investigating developmental trajectories of morphemes as reading units in German. *Journal of Experimental Psychology: Learning, Memory, and*

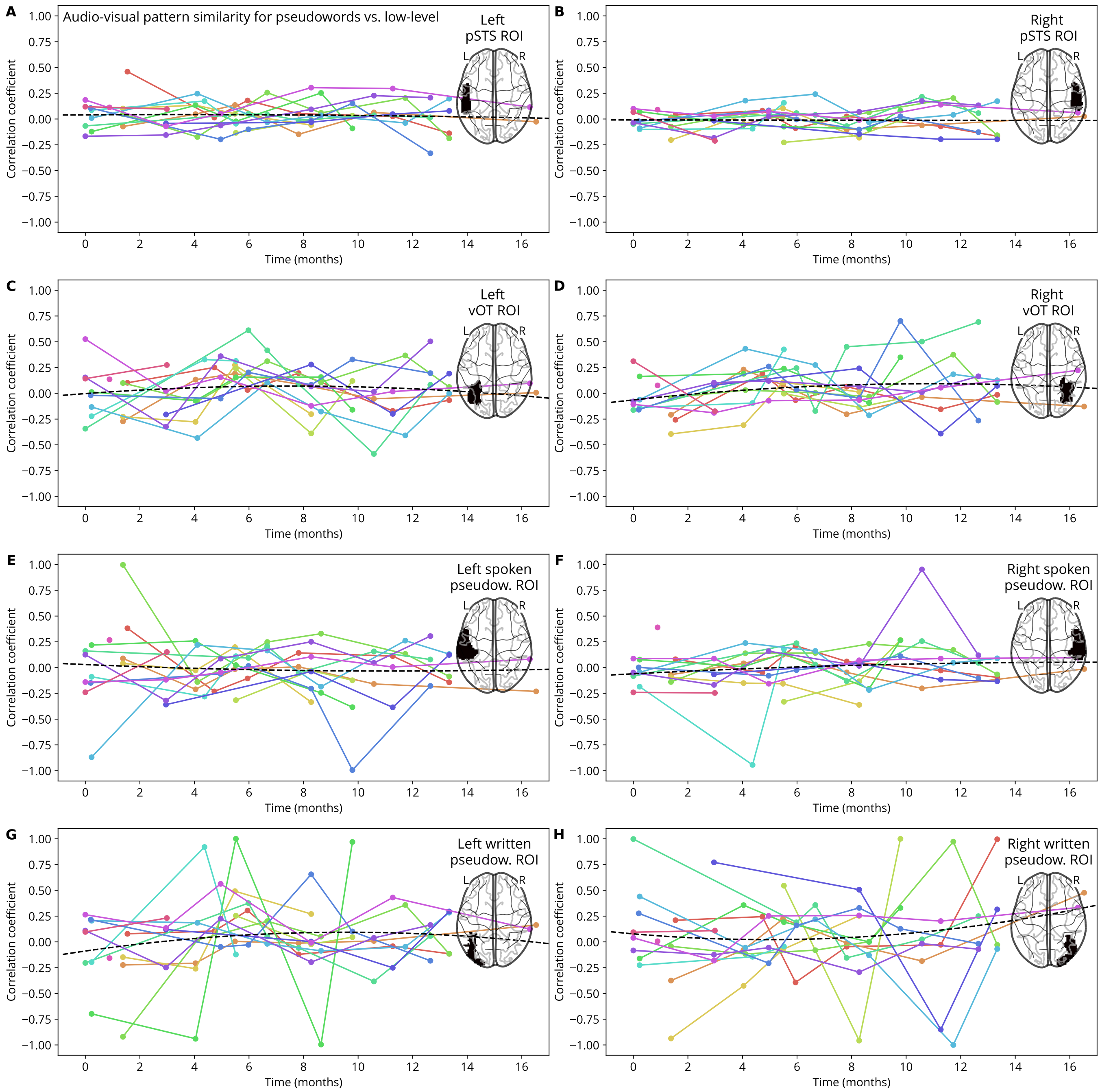

*Figure 10. **Audio-visual pattern similarity for pseudowords versus low-level controls.***

Cognition, *43*(7), 1093–1108. https://doi.org/10.1037/xlm0000353

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hillis, A. E., Newhart, M., Heidler, J., Barker, P., Herskovits, E., & Degaonkar, M. (2005). The roles of the "visual word form area" in reading. *NeuroImage*, *24*(2), 548–559. https://doi.org/10.1016/j.neuroimage.2004.08.026

Hirshorn, E. A., Li, Y., Ward, M. J., Richardson, R. M., Fiez, J. A., & Ghuman, A. S. (2016). Decoding and disrupting left midfusiform gyrus activity during word reading. *Proceedings of the National Academy of Sciences*, *113*(29), 8162–8167. https://doi.org/10.1073/pnas.1604126113

Houston, S. D. (Ed.). (2004). *The first writing: Script invention as history and process.* Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E. C., & Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLOS Computational Biology*, *13*(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350

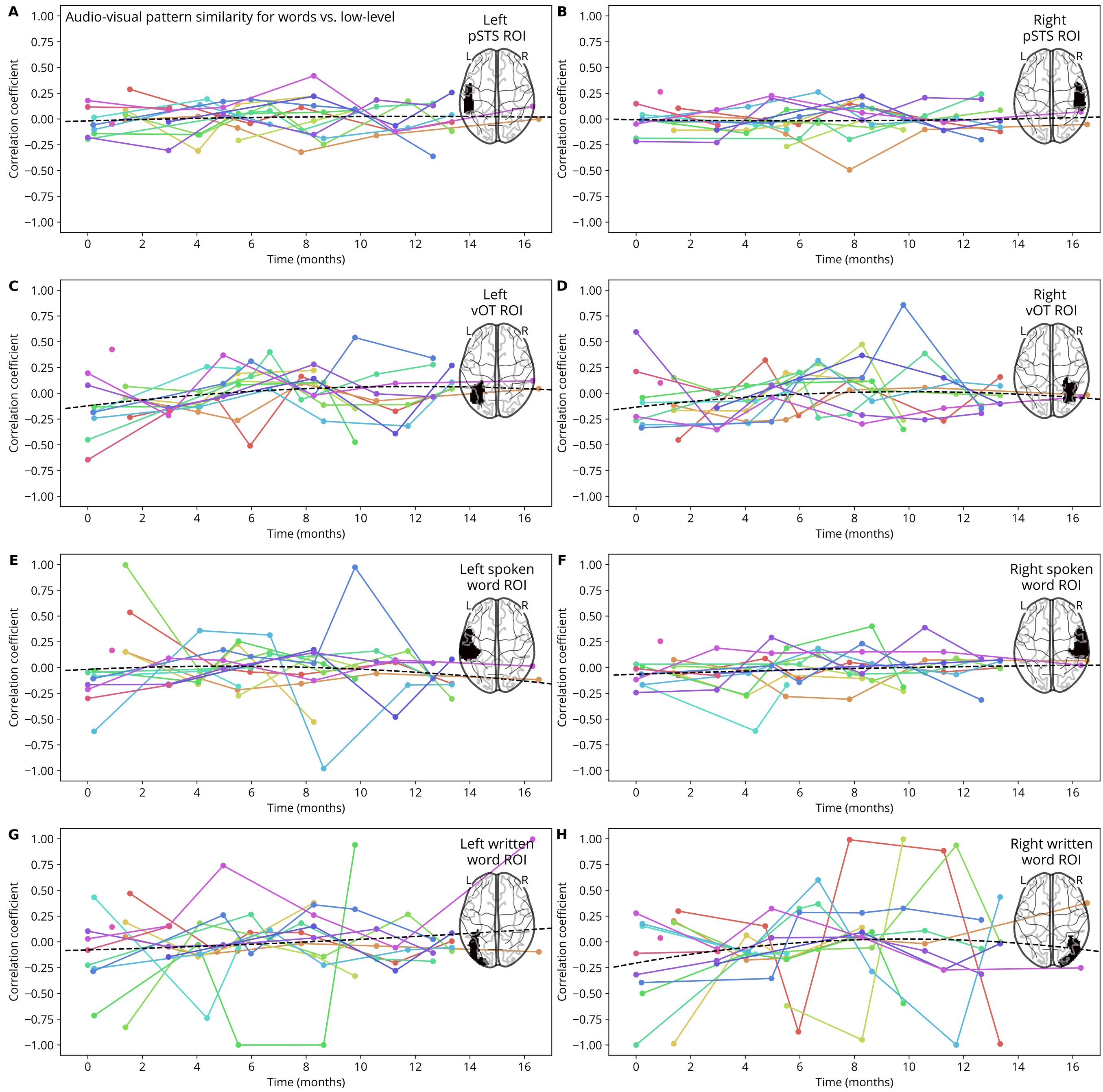

Figure 11. Audio-visual pattern similarity for words versus low-level controls.

Kubota, E., Grill-Spector, K., & Nordt, M. (2024). Rethinking cortical recycling in ventral temporal cortex. *Trends in Cognitive Sciences*, *28*(1), 8–17. https://doi.org/10.1016/j.tics.2023.09.006

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*(7), 293–299. https://doi.org/10.1016/S1364-6613(03)00134-7

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, *138*(2), 322–352. https://doi.org/10.1037/a0026744

Murphy, K. A., Jogia, J., & Talcott, J. B. (2019). On the neural basis of word reading: A meta-analysis of fMRI evidence using activation likelihood estimation. *Journal of Neurolinguistics*, *49*, 71–83. https://doi.org/10.1016/j.jneuroling.2018.08.005

Nordt, M., Gomez, J., Natu, V. S., Rezai, A. A., Finzi, D., Kular, H., & Grill-Spector, K. (2021). Cortical recycling in high-level visual cortex during childhood development. *Nature Human Behaviour*, 1–12. https://doi.org/10.1038/s41562-021-01141-5

Patriat, R., Reynolds, R. C., & Birn, R. M. (2017). An improved model of motion-related signal changes in fMRI. *NeuroImage*, *144, Part A*, 74–82. https://doi.org/10.1016/j.neuroimage.2016.08.051

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*(Supplement C), 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Price, C. J., & Devlin, J. T. (2003). The myth of the visual word form area. *NeuroImage*, *19*(3), 473–481. https://doi.org/10.1016/S1053-8119(03)00084-3

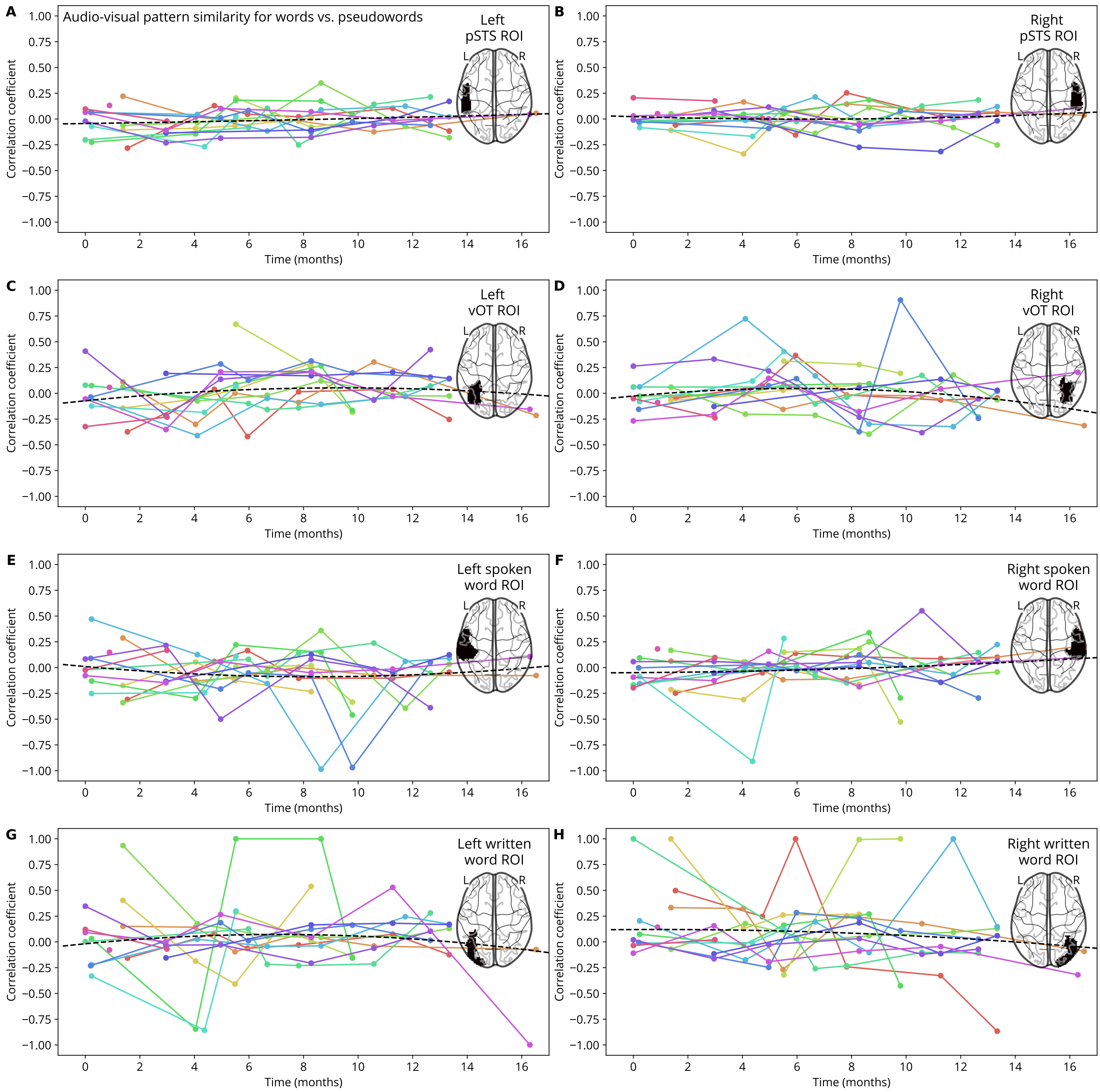

Figure 12. ***Audio-visual pattern similarity for words versus pseudowords.***

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, *53*(4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., & Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*(1), 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an ”outlier” orthography. *Psychological Bulletin*, *134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584

Share, D. L. (2014). Alphabetism in reading science. *Frontiers in Psychology*, *5*, 752. https://doi.org/10.3389/fpsyg.2014.00752

Share, D. L. (2021). Is the science of reading just the science of reading English? *Reading Research Quarterly*, *56*(S1), S391–S402. https://doi.org/10.1002/rrq.401

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

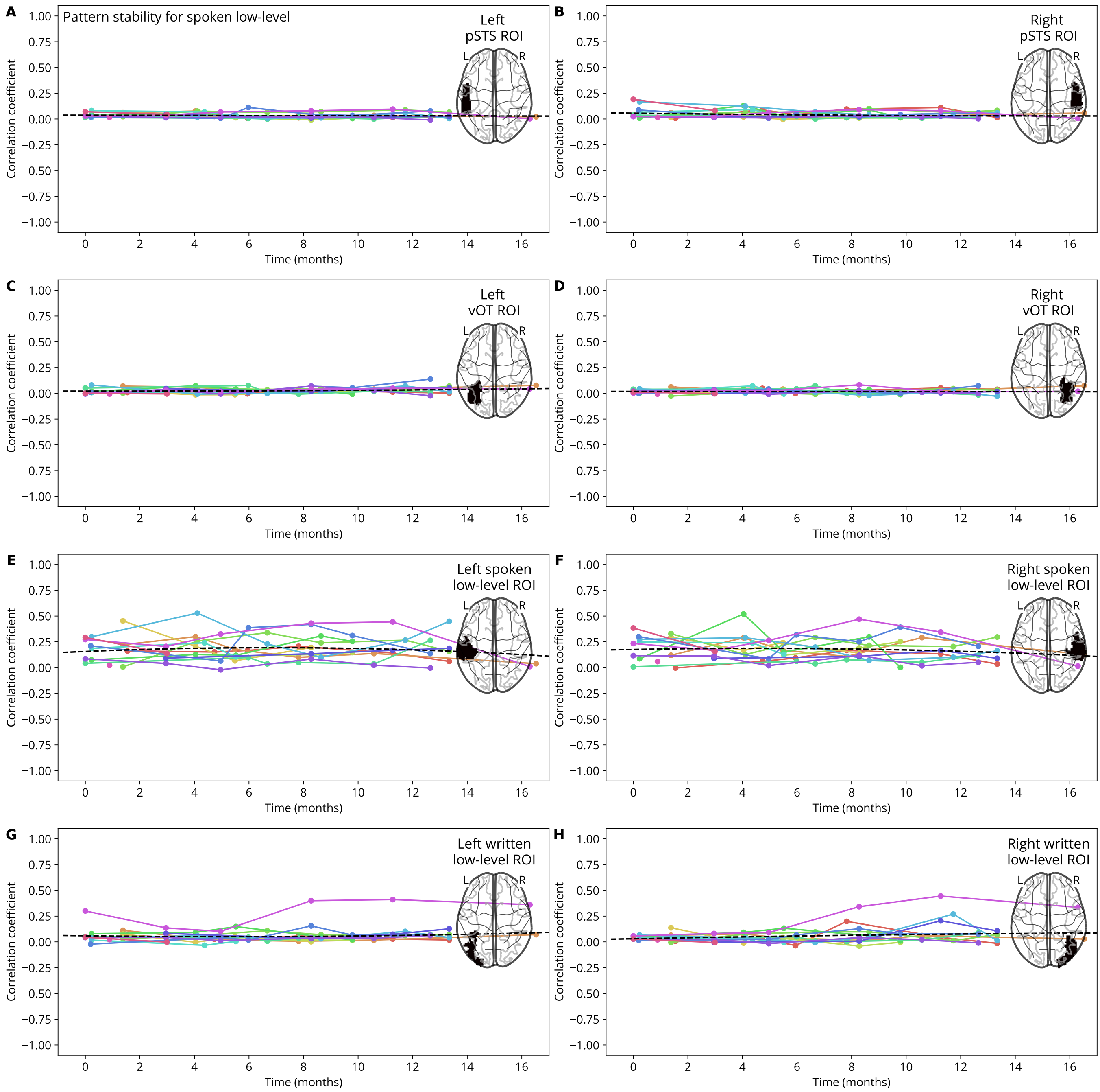

Figure 13. Pattern stability for spoken low-level contros.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, *43*(2), 271–282. https://doi.org/10.1016/j.neuron.2004.06.025

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Wenger, E., Brozzoli, C., Lindenberger, U., & Lövdén, M. (2017). Expansion and renormalization of human brain structure during skill acquisition. *Trends in Cognitive Sciences*, *21*(12), 930–939. https://doi.org/10.1016/j.tics.2017.09.008

Wilson, S. M., Bautista, A., & McCarron, A. (2018). Convergence of spoken and written language processing in the superior temporal sulcus. *NeuroImage*, *171*, 62–74. https://doi.org/gc8hq4

Woolnough, O., Donos, C., Curtis, A., Rollo, P. S., Roccaforte, Z. J., Dehaene, S., Fischer-Baum, S., & Tandon, N. (2022). A spatiotemporal map of reading aloud. *Journal of Neuroscience*, *42*(27), 5438–5450. https://doi.org/10.1523/JNEUROSCI.2324-21.2022

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2020). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, 1–10. https://doi.org/10.1038/s41562-020-00982-w

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57. https://doi.org/10.1109/42.906424

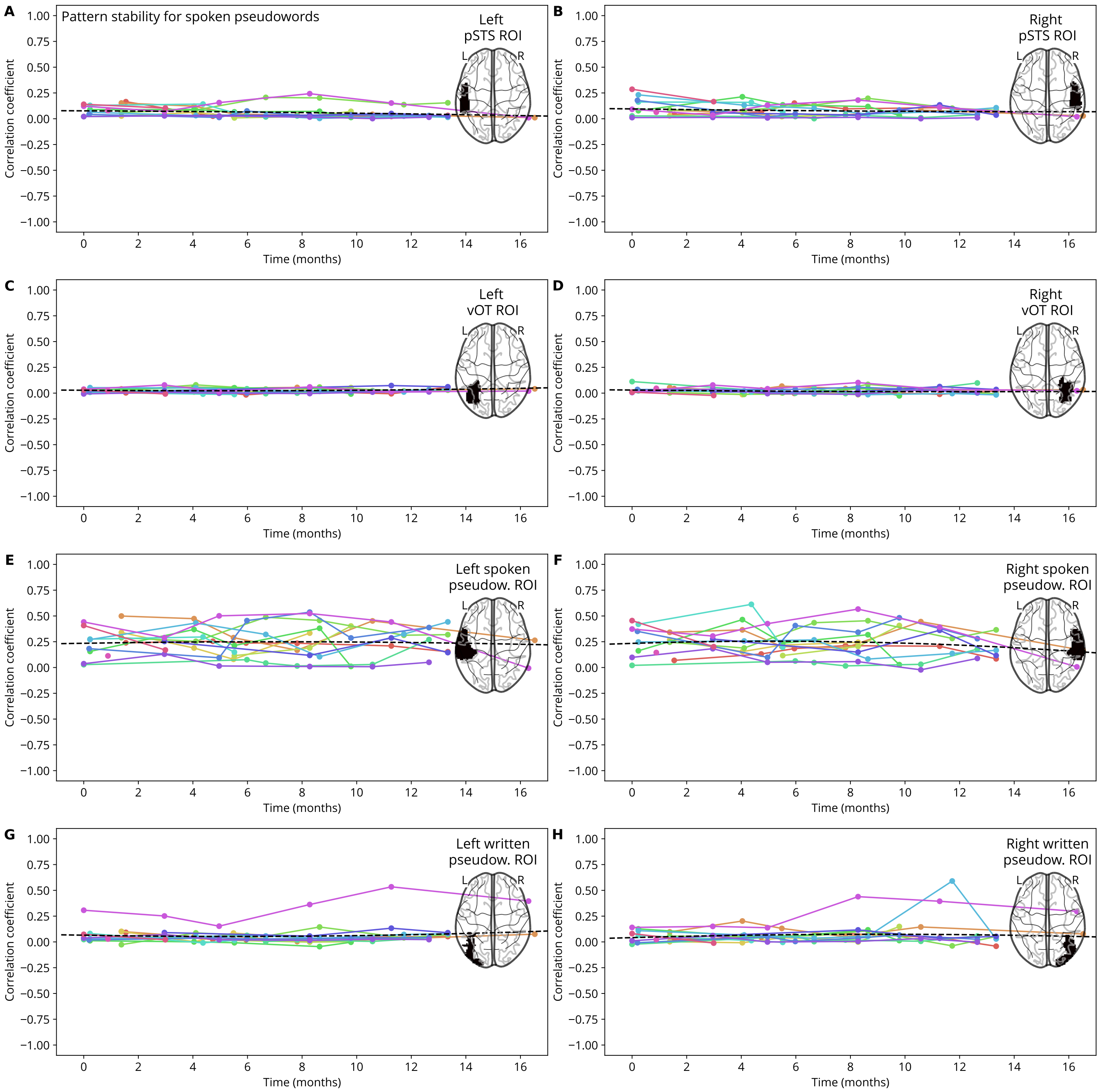

Figure 14. Pattern stability for spoken pseudowords.

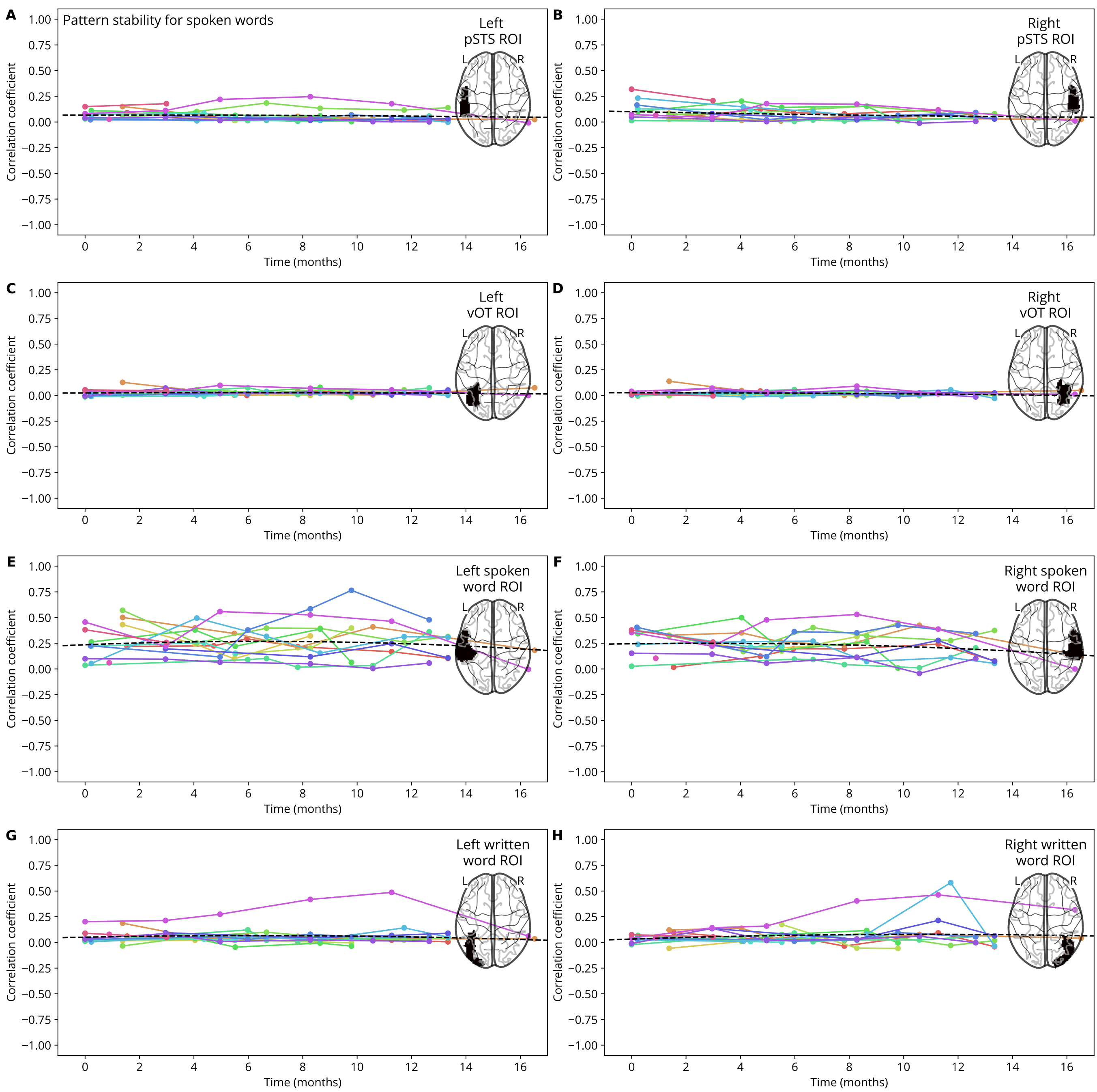

Figure 15. Pattern stability for spoken words.

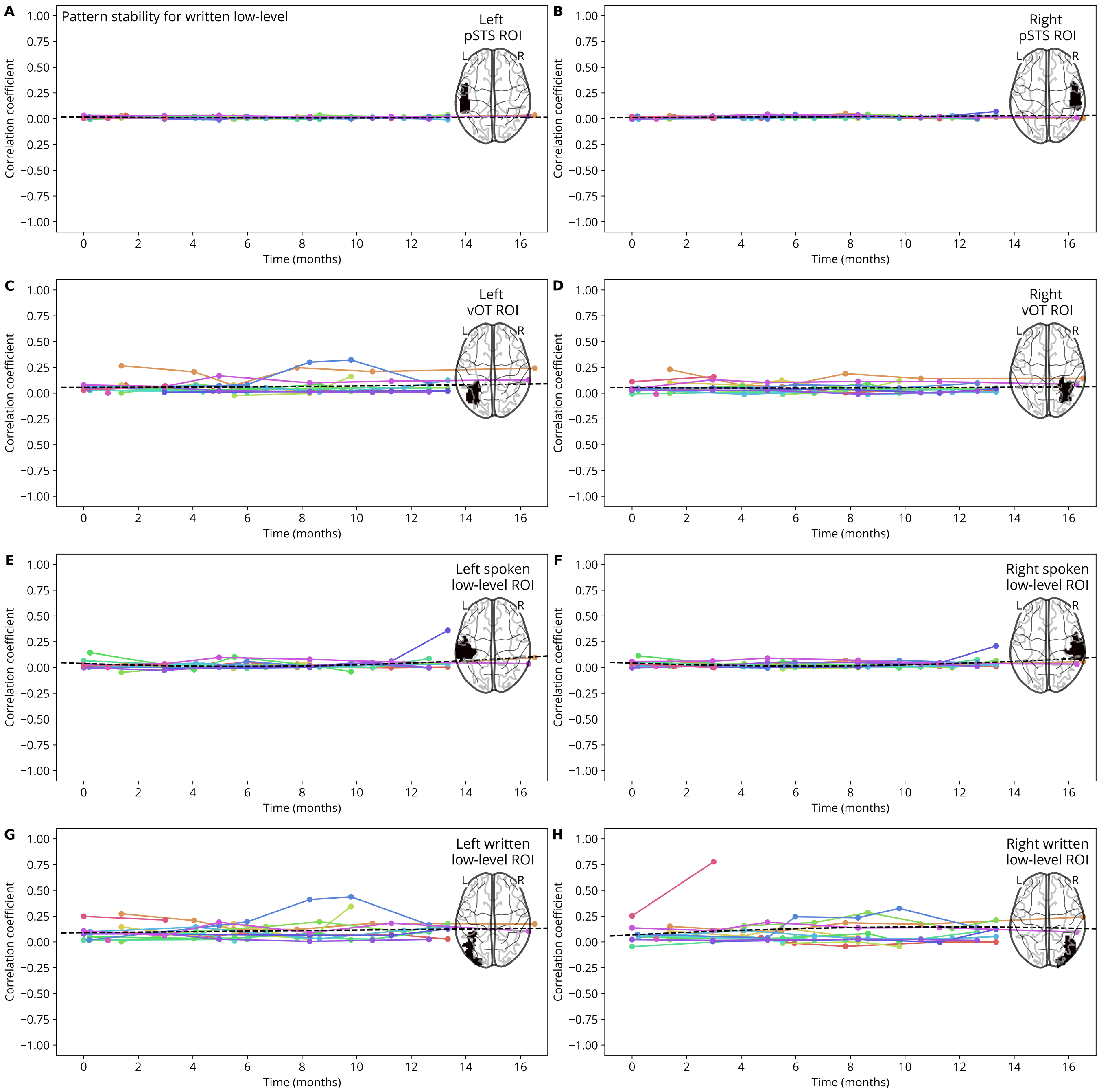

Figure 16. Pattern stability for written low-level contros.

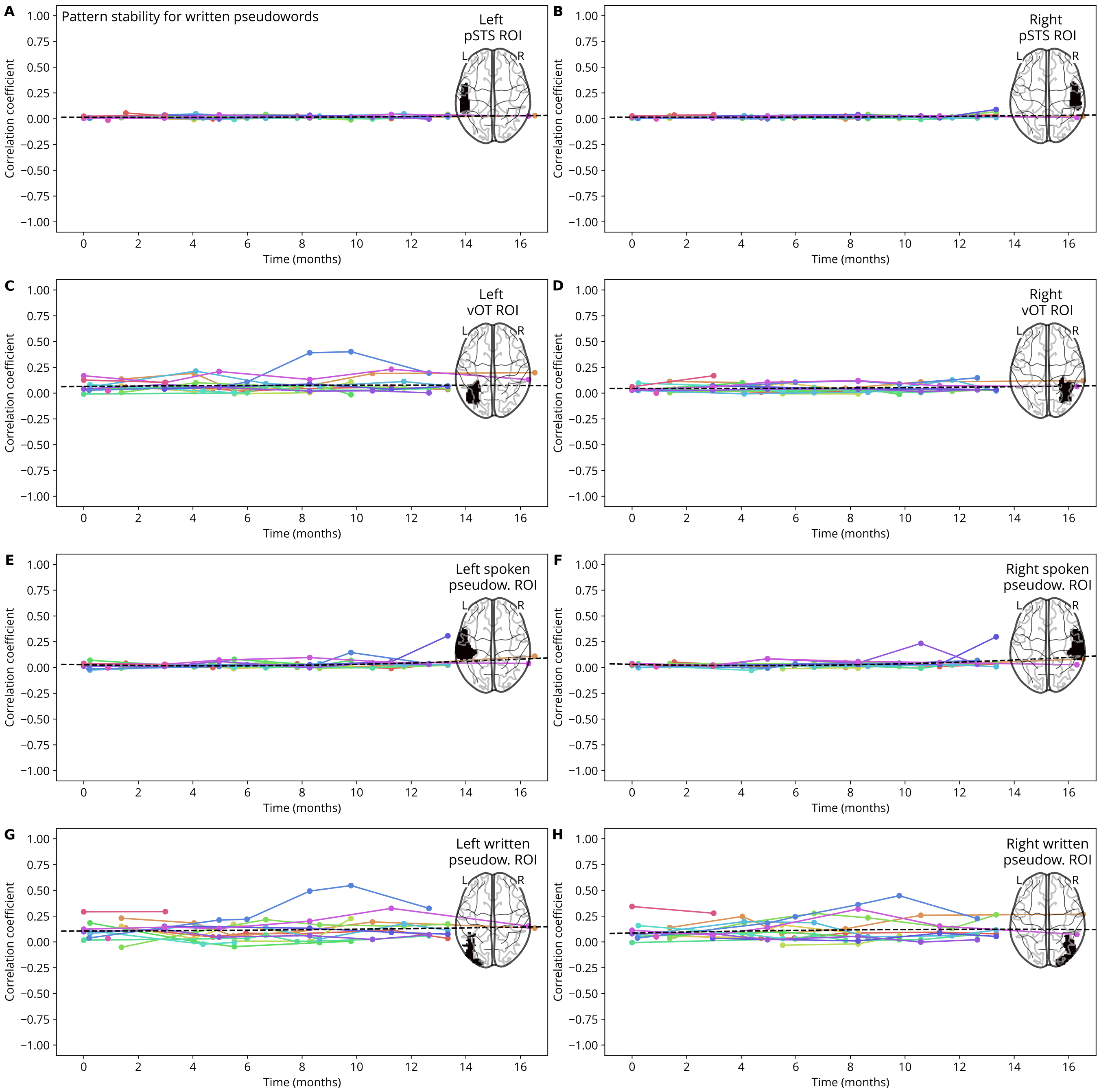

Figure 17. Pattern stability for written pseudowords.

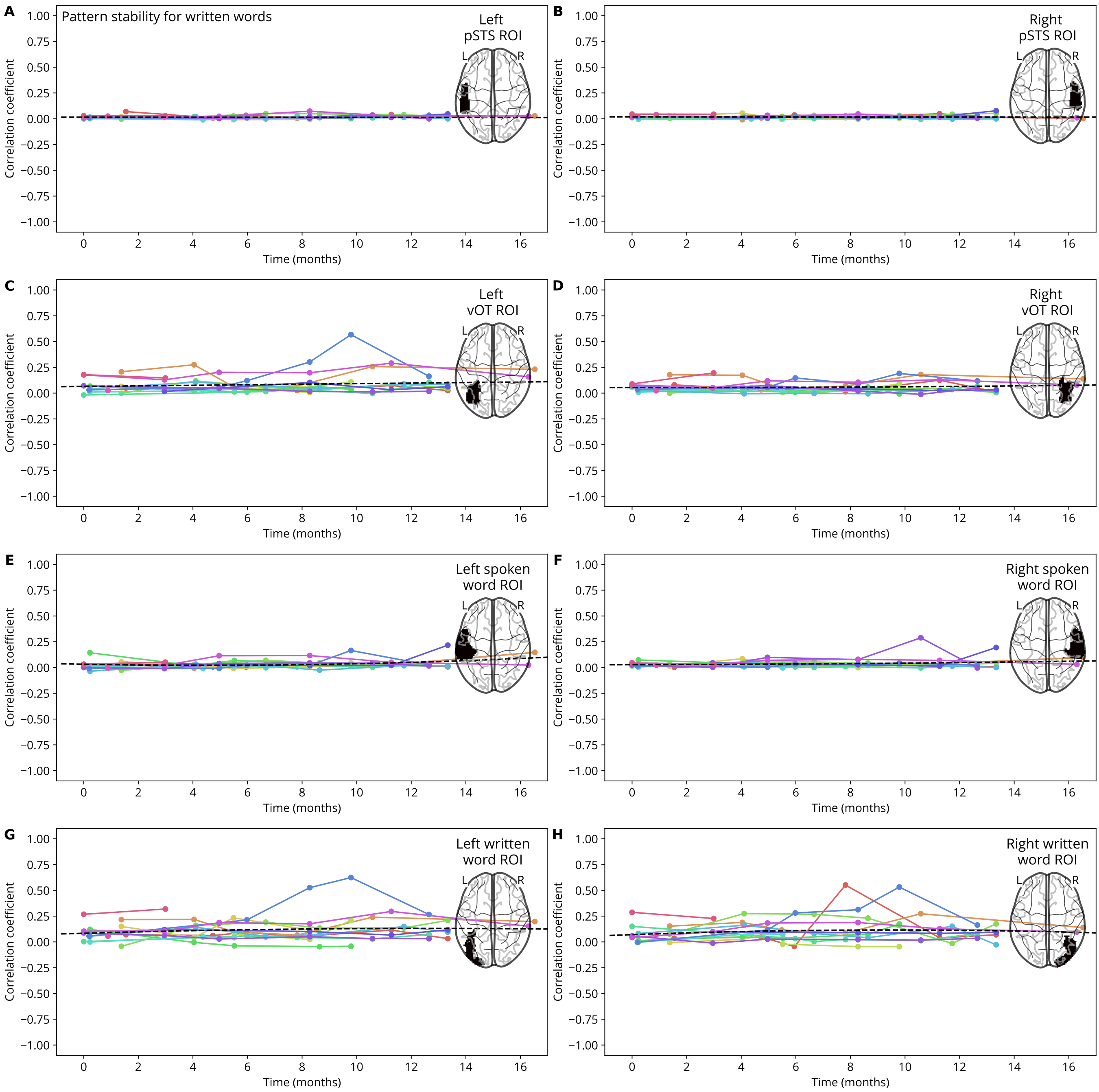

Figure 18. Pattern stability for written words.