Title

Individual Differences in Reading Skill Are Related to Trial-by-Trial Neural Activation Variability in the Reading Network

Journal

Journal of Neuroscience, 38(12)

ISSN

0270-6474

Authors

Malins, Jeffrey G
Pugh, Kenneth R
Buis, Bonnie
et al.

Publication Date

2018-03-21

DOI

10.1523/jneurosci.0907-17.2018

Peer reviewed

JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Individual Differences in Reading Skill are Related to Trial-by-Trial Neural Activation Variability in the Reading Network

Jeffrey G. Malins[a,b], Kenneth R. Pugh[a,c,d,e], Bonnie Buis[a], Stephen J. Frost[a], Fumiko Hoeft[a,f,g,h], Nicole Landi[a,c], W. Einar Mencl[a,d], Anish Kurian[a,c], Ryan Staples[a], Peter Molfese[a,c,i], Rose Sevcik[j] and Robin Morris[j]

[a]*Haskins Laboratories, New Haven, CT 06511*

[b]*Department of Pediatrics, Yale University School of Medicine, New Haven, CT 06520*

[c]*Department of Psychology, University of Connecticut, Storrs, CT 06269*

[d]*Department of Linguistics, Yale University, New Haven, CT 06511*

[e]*Department of Diagnostic Radiology, Yale University School of Medicine, New Haven, CT 06520*

[f]*Department of Psychiatry and Weill Institute for Neurosciences, University of California San Francisco (UCSF), San Francisco, CA 94143*

[g]*Dyslexia Center, UCSF, San Francisco, CA 94143*

[h]*UC6-Stanford Precision Learning Center Neurosciences, San Francisco, CA 94143*

[i]*Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institutes of Mental Health, National Institutes of Health, Bethesda, MD 20892*

[j]*Department of Psychology, Georgia State University, Atlanta, GA 30303*

DOI: 10.1523/JNEUROSCI.0907-17.2018

Received: 4 April 2017

Revised: 30 December 2017

Accepted: 9 January 2018

Published: 12 February 2018

Author contributions: J.G.M., K.R.P., S.J.F., F.H., N.L., W.E.M., A.K., P.J.M., R. Sevciknd R.M. designed research; J.G.M., K.R.P., B.B., S.J.F., F.H., N.L., W.E.M., A.K., R. Staples, P.J.M.nd R.M. performed research; J.G.M., W.E.M., A.K., R. Staplesnd P.J.M. contributed unpublished reagents/analytic tools; J.G.M., B.B., W.E.M.nd R. Staples analyzed data; J.G.M., K.R.P., S.J.F., F.H., N.L., W.E.M., A.K., R. Sevciknd R.M. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Numbers P01HD070837 to Georgia State University, P01HD001994 to Haskins Laboratoriesnd R01HD086168 to Haskins Laboratories. Additionally, FH was supported by NIH R01HD078351, R01HD067254, P50HD052120, R01HD044073, UCOP MRP-17-454926, Oak Foundation ORIO-16-012nd NSF 1540854. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Special thanks to the children and their parentss well as their schools and teachers, for participation in this study. We would also like to thank Candice Goerger for her extensive role in recruitment and data collection at the Atlanta site. The authors declare no competing financial interests.

Correspondence should be addressed to: Jeffrey G. Malins, Haskins Laboratories, 300 George Street Suite 900, New Haven, CT, 06511; Tel: 1-203-865-6163 x 237; E-mail: jeffrey.malins@yale.edu

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.0907-17.2018

Running Title: Reading Skill is Related to fMRI BOLD Variability

Individual Differences in Reading Skill are Related to Trial-by-Trial Neural Activation Variability in the Reading Network

Jeffrey G. Malins[a,b*], Kenneth R. Pugh[a,c,d,e], Bonnie Buis[a], Stephen J. Frost[a], Fumiko Hoeft[a,f-h], Nicole Landi[a,c], W. Einar Mencl[a,d], Anish Kurian[a,c], Ryan Staples[a], Peter Molfese[a,c,i], Rose Sevcik[j], & Robin Morris[j]

[a]Haskins Laboratories, New Haven, CT 06511
[b]Department of Pediatrics, Yale University School of Medicine, New Haven, CT 06520
[c]Department of Psychology, University of Connecticut, Storrs, CT 06269
[d]Department of Linguistics, Yale University, New Haven, CT 06511
[e]Department of Diagnostic Radiology, Yale University School of Medicine, New Haven, CT 06520
[f]Department of Psychiatry and Weill Institute for Neurosciences, University of California San Francisco (UCSF), San Francisco, CA 94143
[g]Dyslexia Center, UCSF, San Francisco, CA 94143
[h]UC6-Stanford Precision Learning Center Neurosciences, San Francisco, CA 94143
[i]Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institutes of Mental Health, National Institutes of Health, Bethesda, MD 20892
[j]Department of Psychology, Georgia State University, Atlanta, GA 30303

*Correspondence should be addressed to: Jeffrey G. Malins, Haskins Laboratories, 300 George Street Suite 900, New Haven, CT, 06511; Tel: 1-203-865-6163 x 237; E-mail: jeffrey.malins@yale.edu

Number of Tables: 5; Number of Figures: 3; Number of words in Abstract: 246; Number of words in Introduction: 649; Number of words in Discussion: 1500

Acknowledgments

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Numbers P01HD070837 to Georgia State University, P01HD001994 to Haskins Laboratories, and R01HD086168 to Haskins Laboratories. Additionally, FH was supported by NIH R01HD078351, R01HD067254, P50HD052120, R01HD044073, UCOP MRP-17-454926, Oak Foundation ORIO-16-012, and NSF 1540854. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Special thanks to the children and their parents, as well as their schools and teachers, for participation in this study. We would also like to thank Candice Goerger for her extensive role in recruitment and data collection at the Atlanta site. The authors declare no competing financial interests.

Abstract

Recent work has suggested that variability in levels of neural activation may be related to behavioral and cognitive performance across a number of domains, and may offer information that is not captured by more traditional measures that use the average level of brain activation. We examined the relationship between reading skill in school-aged children and neural activation variability during an fMRI reading task after taking into account average levels of activity. The reading task involved matching printed and spoken words to pictures of items. Single trial activation estimates were used to calculate the mean and standard deviation of children's responses to print and speech stimuli; multiple regression analyses evaluated the relationship between reading skill and trial-by-trial activation variability. The reliability of observed findings from the discovery sample (*N* = 44; ages 8-11; 18 female) was then confirmed in an independent sample of children (*N* = 32; ages 8-11; 14 female). Across the two samples, reading skill was positively related to trial-by-trial variability in the activation response to print in the left inferior frontal gyrus *pars triangularis*. This relationship held even when accounting for mean levels of activation. This finding suggests that intrasubject variability in trial-by-trial fMRI activation responses to printed words accounts for individual differences in human reading ability that are not fully captured by traditional mean levels of brain activity. Furthermore, this positive relationship between trial-by-trial activation variability and reading skill may provide evidence that neural variability plays a beneficial role during early reading development.

Keywords: Trial-by-trial variability; BOLD variability; Reading disability; Neural noise; Individual differences; Beta series; Event-related fMRI

Significance Statement

Recent work has suggested that neural activation variability, or moment-to-moment changes in the engagement of brain regions, is related to individual differences in behavioral and cognitive performance across multiple domains. However, differences in neural activation variability have not yet been evaluated in relation to reading skill. In the current study, we analyzed data from two independent groups of children who performed an fMRI task involving reading and listening to words. Across both samples, reading skill was positively related to trial-by-trial variability in activation to print stimuli in the left inferior frontal gyrus *pars triangularis*, even when accounting for the more conventional measure of mean levels of brain activity. This finding suggests that neural variability could be beneficial in developing readers.

Introduction

A growing body of neuroimaging research has linked reading skill to variation in structural and functional circuitry in the brain (Norton, Beach, & Gabrieli, 2015). Broadly, investigations concerning the functional neuroanatomy of reading have focused on mean levels of activation across trials while children are engaged in different reading tasks. Yet, an emerging literature suggests that mean differences in activation and connectivity reflect only part of the complex neural foundation of reading ability. Recent studies have linked reading skill to the stability of neural responses to speech sounds (Hornickel & Kraus, 2013). In addition, animal work has shown that expression of the rat homolog of the dyslexia susceptibility gene *KIAA0319* is linked to increased trial-by-trial variability in speech sound responses (Centanni, Booker, et al., 2014; Centanni, Chen, et al., 2014). Together, these studies have helped motivate the neural noise hypothesis of reading disability, which postulates that levels of neural noise can influence timing mechanisms that impact signal variability and thereby affect reading performance (Hancock, Pugh, & Hoeft, 2017).

This previous work leads to an expectation that reading skill in children is related to within-subject measures of neural activation variability. Yet, to date, neural activation variability has been evaluated with respect to reading skill in children only by examining brainstem electrophysiological responses to speech sounds (Hornickel & Kraus, 2013). In the current study, we instead asked children to perform a task involving word reading, and examined trialwise variability in cortical activation using fMRI, a technique that has been successfully used to examine the relationship between neural activation variability and behavioral performance in multiple domains outside of reading (Garrett et al., 2013).

Across these other domains, there exists some debate concerning whether increased variability in the blood oxygen level-dependent (BOLD) signal confers a positive or negative impact on behavior. The directionality of the effect appears in part to be related to the extent to which a task entails cognitive versus sensory processing. For example, increased BOLD signal variability has been associated with faster and more consistent reaction times in younger versus older adults during cognitive tasks including attentional cueing and delayed match-to-sample (Garrett, Kovacevic, McIntosh, & Grady, 2011), whereas in a study examining cognitive flexibility and stability (Armbruster-Genc, Ueltzhoffer, & Fiebach, 2016), the direction of the relationship between BOLD signal variability and cognitive performance has been characterized as positive or negative depending on the task. Conversely, for sensory processing, increased BOLD signal variability has been associated with increased behavioral variability in older compared to younger adults during audiovisual speech perception (Baum & Beauchamp, 2014), and has been considered maladaptive in adults with autism, who showed greater trial-to-trial variability compared to matched controls in primary sensory areas during a low-level sensory task (Haigh et al., 2016). Given that reading involves both sensory and cognitive components, the direction of the relationship between BOLD signal variability and reading skill therefore remains an open question.

In the current study, we address the following novel questions: (1) does trial-by-trial neural activation variability account for variance in reading skill in children above and beyond differences in mean activation; (2) if so, what is the direction of the relationship between neural activation variability and reading skill? To address these questions, we first conducted analyses on fMRI data from a discovery sample of children who performed a task in which they judged

whether printed or spoken words matched pictures of items. We then confirmed whether observed effects held in a separate, independent sample of children. Analyses focused on using single trial beta estimates to quantify mean activation across trials, as well as trial-by-trial variability, in the evoked response to print and speech within regions of the reading network. These mean and variability measures were then entered into multiple regression models characterizing the manner in which trial-by-trial activation variability is associated with reading skill after accounting for mean task-related activation as well as predictors of non-interest such as subject age.

Methods

Discovery Sample

Participants. Children were selected from a larger study examining response to intervention for reading disability; the data presented here correspond to baseline scans prior to the onset of any intervention. Of this larger sample of 82 children, 44 were selected who met the following inclusion criteria: in third or fourth grade (71/82; the other 11 participants belonged to a cohort of seventh and eighth graders who participated in the larger study), an average Euclidean movement of .25 mm or less (58/71), and at least 70% accuracy in each of the auditory and visual mismatch conditions (44/58). Euclidean movement was calculated per volume by first computing point-to-point change for each the six motion parameters (i.e., three translation and three rotation), and then taking the square root of the sum of squares of these measures; average Euclidean movement was calculated by taking the mean value of this measure across all volumes of data collection. The accuracy cutoffs were selected in order to have a sufficient

number of correct trials per participant to calculate dependable standard deviation (SD) measurements. The motion cutoff was selected in order to increase power to detect effects related to differences in trial-by-trial variability that are unrelated to motion, because we expected that intrasubject variability would be impacted by movement in the scanner (Lund et al., 2005). We acknowledge that the percent of data lost is larger than comparable fMRI studies with pediatric populations; however, data quality criteria were particularly stringent for this investigation given our concerns regarding participant motion as well as the requirement of having a sufficient number of trials to calculate valid SD measurements.

All children completed a battery of standardized cognitive assessments (Table 1). These included assessments of single word reading, pseudoword decoding, and passage comprehension from the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001); the Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 2007), which measures receptive vocabulary; the Comprehensive Test of Phonological Awareness (CTOPP-2; Wagner, Torgesen, Rashotte, & Pearson, 2013), which measures metalinguistic knowledge of the structure of speech, or phonological awareness, by assessing skills including phoneme elision, blending, and isolation; and the Wechsler Abbreviated Scale of Intelligence (WASI-II; Wechsler & Hsiao-Pin, 2011), which measures verbal and non-verbal intelligence. As can be observed in Table 1, the range of reading scores was very broad, and some children in the sample would be considered typically developing whereas others would be classified as having reading disability using traditional diagnostic criteria. However, we treated reading skill as a continuous dimension, in line with recent views concerning the multifactorial nature of

reading skill as well as the pitfalls of grouping children into diagnostic categories using cutoff scores (Pennington et al., 2012; Branum-Martin et al., 2013).

fMRI Task. Functional volumes were acquired while participants completed a task in which they judged whether picture cues matched auditory and visual target words (Frost et al., 2009; Jasińska et al., 2016; Landi et al., 2013; Preston et al., 2016). In this task, participants were presented with pictures of common items (e.g., "cake") that remained on the screen for 40-65 seconds – corresponding to between seven and eight trials – before being replaced by another picture. This procedure encouraged participants to generate strong expectations of target items, thereby maximizing responses to mismatches, and also obviated the need to associate targets with a new picture on every single trial, which could have been overly taxing. While each picture remained on the screen, participants were presented with target items in an event-related fashion; specifically, printed words appeared in a box below the picture (presented for 3000 ms in 40-point Arial font), or auditory words were presented via headphones. Importantly, in one sixth of trials, the printed or spoken word matched the picture, while in the other five sixths of trials the printed or spoken word mismatched the picture. Participants were asked to indicate via button press whether or not the printed or spoken word matched the picture. In total, participants completed 25 trials in each of the auditory (spoken) and visual (print) mismatch conditions. A sample trial sequence is illustrated in Figure 1.

Acquisition of MRI Data. Images were acquired using a 3T Siemens Trio scanner with a 12-channel head coil located at the GSU/GaTech Center for Advanced Brain Imaging in Atlanta, Georgia. T2*-weighted images were acquired in an axial-oblique orientation parallel to the intercommissural line (32 slices; 4 mm slice thickness; no gap) using single-shot echo planar

imaging (matrix size = 64 × 64; voxel size = 3.438 × 3.438 × 4 mm; FoV = 220 mm; TR = 2000 ms; TE = 30 ms; flip angle = 80°). To allow for stabilization of the magnetic field, the first four volumes within each run were discarded. Anatomical scans were collected in the same orientation as the functional volumes (MPRAGE; matrix size = 256 × 256; voxel size = 1 × 1 × 1 mm; FoV = 256 mm; TR = 2530 ms; TE = 2.77 ms; flip angle = 7°); these were acquired either following or between the functional runs. In total, participants completed two runs of the functional task, which had a combined duration of 7 minutes 32 seconds (226 volumes). Across all trials in the experiment, the time between trial onsets was jittered between 4 and 13 seconds; trial order and ITIs were optimized by an in-house Matlab program that balanced ITIs and null trials across conditions, and minimized the variability of the measured response in Monte Carlo simulations.

Analysis Pipeline. ***Preprocessing.*** Data were analyzed using AFNI (Cox, 1996; RRID:SCR_005927). Functional images were pre-processed by first correcting for slice acquisition time (*3dTshift*). Following this, functional images were aligned with anatomical images, were corrected for motion using a six-parameter rigid-body transform (*3dvolreg*), and were normalized to the Colin27 brain in Talairach space using an affine transform (*@auto_tlrc*). These three steps were combined into a single transform that also forced a 3 mm isotropic voxel size on the data. All images were then smoothed (*3dmerge*) using a Gaussian kernel with a FWHM of 8 mm (i.e., twice the between-plane distance of 4 mm; Skudlarski, Constable, & Gore, 1999), and data were scaled (*3dcalc*) so that each voxel's time series had a mean of 100 for each run. During this scaling step, values in excess of 200 were clipped; this is the default value for scaling in AFNI, and was selected in order to retain the precision of scaled short values.

We elected to use the Talairach atlas for normalization because Burgund et al. (2002) have shown that relative to the resolution of fMRI data, there are minimal anatomical differences between children ages 7 and 8 compared to adults. Given that the children in the current study were even older than the children in the Burgund et al. (2002) study (i.e., between 8 and 12 years of age), our view is that use of the Talairach atlas should allow for broader comparability between our study and others, including developmental investigations with adult samples.

GLM Analysis. Single trial beta estimates were obtained using a single GLM including nuisance regressors for the six motion parameters as well as a separate regressor for each trial (least-squares all, or LS-A; Mumford, Turner, Ashby, & Poldrack, 2012; Rissman et al., 2004). This model was specified using the *-stim_times_IM* flag for *3dDeconvolve* in AFNI. The HRF was approximated using a gamma function. Because we were interested in intrinsic neural variability as opposed to variability related to individual differences in behavioral performance on the task, we included reaction times for each trial as duration modulators in the GLM (Grinband, Wager, Lindquist, Ferrera, & Hirsch, 2008; Yarkoni, Barch, Gray, Conturo, & Braver, 2009). For trials in which a participant either did not respond, responded with an RT less than 200 ms (i.e., invalid anticipation), or responded with an RT greater than 1.5 times the interquartile range above the third quartile for a participant's distribution of reaction times, overall mean RT for that participant was used as a duration modulator; however, these trials were not considered in further analyses.

Beta estimates corresponded to the amplitude assigned to each regressor in the GLM, and the set of beta estimates across trials for a given voxel constituted that voxel's beta series. When performing the GLM, any volume that exceeded the thresholds of .3 mm Euclidean movement

and/or 10% outliers were censored from further analysis, resulting in an average loss of less than one trial in each of the auditory and visual mismatch conditions. It should be noted that this approach gave rise to some extreme outlier beta values due to rare spikes that were still present in the data even after censoring these volumes. To handle these, outlier beta values were identified for each participant using the program *3dToutcount* in AFNI, which flags outliers using an algorithm based on median absolute deviation. Trials with outliers in greater than 10% of voxels in the brain were censored from analysis; this occurred for an average of two trials in each of the auditory and visual mismatch conditions. In all other trials, outlier values were replaced with zeroes and ignored when calculating average beta values within regions of interest (the mean number of voxels with outlier values across all trials and participants was less than one in both the auditory and visual mismatch conditions in each of the ROIs detailed in the next section).

ROI Selection. Given that mean activation was one of the predictors we aimed to include in the multiple regression models evaluating relationships with reading skill, we elected not to analyze mean activation at the whole brain level because this would have biased selection of ROIs. Moreover, we did not predict perfect concordance between areas in which the reading task resulted in overall levels of activation and areas in which the task resulted in increased levels of variance in activation. Therefore, we instead defined ROIs using a recent meta-analysis that took the results of 20 different imaging studies of reading in children and combined them to identify a set of coordinates which showed convergence across studies (Martin, Schurz, Kronbichler, & Richlan, 2015). Because this meta-analysis combined results across tasks examining different aspects of reading, our view was that by using these co-ordinates, we were

more likely to include regions that may show a relationship between neural activation variability and reading skill, even if these regions do not appear in a map of mean activation for the current task; with that said, we acknowledge the limitation that the meta-analysis also used mean activation to define ROIs.

We created spheres with a radius of 6 mm (two voxels) centered on the Martin et al. (2015) coordinates for the following regions: left inferior frontal gyrus (IFG) *pars opercularis*, left IFG *pars triangularis*, left middle temporal gyrus, left superior temporal gyrus, left superior parietal lobule, and left inferior temporal gyrus. In addition, we also included an ROI for left thalamus, given extant findings indicating that the thalamus contributes to the reading network (Galaburda, Sherman, Rosen, Aboitiz, & Geschwind, 1985; Pugh et al., 2013). The full set of ROIs selected for analysis are detailed in Table 3 and displayed in Figure 2. For reference, we overlaid these ROIs on a conjunction map that shows the extent of overlap in task-related activation between the Discovery Sample and the Confirmation Sample (further details below). This map was created by running a standard GLM for each participant with a single regressor per condition; groupwise evoked response maps across all task conditions were then generated for both the Discovery and Confirmation samples using the program *3dANOVA2* (corrected at FDR < .01). The conjunction map was created using step functions (*3dcalc*) and adding together resultant maps.

Analysis of Trialwise Variability. For analysis of trialwise variability, we considered correct trials in the auditory and visual word mismatch conditions with RTs within the acceptable range. Our rationale for analyzing both print and speech trials was that even though we were specifically interested in responses to printed words, the paradigm included spoken words, and analyzing

neural responses in this condition afforded us the ability to examine whether any potential relationships between neural activation variability and reading skill were print-specific or were instead more general for language. Our rationale for analyzing only the mismatch conditions was that these were the predominant conditions in the experiment in terms of the overall number of trials; related to this, there were too few trials in the match conditions to calculate valid SD measurements. For each trial and each ROI, we calculated the average beta weight across the voxels in the ROI (*3dROIstats*), ignoring outlier voxels that had been replaced with zero.

Next, we calculated the mean and SD of the beta series in each ROI. Intrasubject SD measures were calculated by using leave-one-out jack-knife estimation in version 2015.2 of the package 'bootstrap' (Tibshirani & Leisch, 2015) in the R Project for Statistical Computing (RRID:SCR_001905) and taking the mean across estimates. Jack-knife estimation was used to mitigate bias of SD estimates, especially given the relatively small number of measurements from which these SDs were derived (Efron, 1981). Then, in each ROI, we ran separate multiple regression models for the auditory and visual mismatch conditions with reading skill as the dependent variable, which was quantified using raw scores for Letter-Word Identification (LWID) from the Woodcock-Johnson III Tests of Achievement. We started with a full model that included the mean and SD of the beta series in either the auditory or visual mismatch condition as well as the following predictors of non-interest: age in months (Garrett et al., 2011; McIntosh, Kovacevic, & Itier, 2008); amount of subject motion, defined as the average point-to-point Euclidean movement across all volumes of data collection (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012); and the number of trials used to calculate the mean and SD of the beta series

(i.e., the number of correct trials following removal of trials that exceeded motion, outlier, or RT thresholds). Using the program *dropterm* in version 7.43-45 of the 'MASS' package in R (Venables & Ripley, 2002), we removed, in a stepwise fashion, any of the three predictors of non-interest that did not account for significant variance in reading skill (an alpha criterion of .05 was used for backward selection; at each step, the predictor with the largest associated *p*-value was removed). Then, for the resulting models, change in AIC, change in BIC, and change in adjusted R^2 were quantified for both the mean and SD of the beta series by comparing final models with models in which each of these respective terms were removed.

Confirmation Sample

Participants. Children were selected for this analysis from a large dataset that has been the subject of other reports (Frost et al., 2009; Jasińska et al., 2016; Landi et al., 2013; Preston et al., 2016). From this larger sample of 122 children, we first selected participants whose average Euclidean movement was .25 mm or less (81/122). Next, because the distribution of reading ability in the larger sample differed from the Discovery Sample, which was weighted toward the lower end of the reading skill distribution, we selected a subset of children who were matched to the Discovery Sample in age and raw single word reading scores (WJ-III LWID) using version 3.0.1 of the R package 'MatchIt' (Ho, Imai, King, & Stuart, 2011). From this subset of children, we then selected those who attained at least 70% accuracy in each of the auditory and visual mismatch conditions, which resulted in 32 children in the Confirmation Sample (14 female). Assessment scores for the Confirmation Sample are listed in Table 1; as can be noted from the table, mean raw reading scores and mean age were not significantly different across the two samples (WJ-III LWID raw scores: $t(55) = -1.51$, $p = .14$; age: $t(47) = -.30$, $p = .77$), although

standard single word reading scores, phonological awareness, vocabulary, and IQ were lower in the Discovery Sample compared to the Confirmation Sample (WJ-III LWID standard scores: $t(64) = -2.26$, $p = .03$; CTOPP phonological awareness composite standard scores: $t(57) = -5.51$, $p < .001$; PPVT standard scores: $t(72) = -3.39$, $p = .001$; WASI FSIQ-2 standard scores: $t(68) = -3.15$, $p < .01$). We would argue that such sample differences provide for increased generalizability of results.

fMRI Task. Functional volumes were acquired while participants completed the same picture cue-target word identification task as the children in the Discovery Sample. However, the task in this sample included a larger number of conditions. More specifically, for both the auditory and visual modalities, mismatches were either real words or pseudowords; in addition, for the visual modality, some mismatches were either semantically related words or meaningless consonant strings. As a result of this different design, the match to mismatch ratio was 1:4 instead of 1:5; in addition, printed words were presented for a duration of 2000 ms and in 18-point Verdana font. To keep the confirmatory analyses as comparable as possible to the analyses employed for the Discovery Sample, only the real word conditions in both modalities were considered.

Acquisition of MRI Data. Images were acquired using a 1.5T Siemens Sonata scanner with a one-channel head coil located at the Yale Magnetic Resonance Research Center in New Haven, Connecticut. T2*-weighted images were acquired in an axial-oblique orientation parallel to the intercommissural line (20 slices; 6 mm slice thickness; no gap) using single-shot echo planar imaging (matrix size = 64 × 64; voxel size = 3.125 × 3.125 × 6 mm; FoV = 200 mm; TR = 2000 ms; TE = 50 ms; flip angle = 80°). To allow for stabilization of the magnetic field, the first four

volumes within each run were discarded. Anatomical scans were collected in a sagittal orientation (MPRAGE; matrix size = 256 × 256; voxel size = 1 × 1 × 1 mm; FoV = 256 mm; TR = 2000 ms; TE = 3.65 ms; flip angle = 8°); these were acquired either following or between the functional runs. Participants completed between seven and ten functional runs each 3:46 (113 volumes) in length, which corresponded to up to 40 trials in each of the auditory and visual mismatch conditions (i.e., four trials in each condition in each run). Across all trials in the experiment, the time between trial onsets was jittered between 4 and 13 seconds.

Analysis Pipeline. All analyses were conducted in the exact same fashion as they were for the Discovery Sample. Removal of volumes that exceeded the thresholds of .3 mm Euclidean movement and/or 10% outliers prior to the GLM resulted in an average loss of less than one trial in each of the auditory and visual mismatch conditions. Following the GLM, removal of trials with outliers in greater than 10% of voxels in the brain resulted in a further loss of three trials on average in the auditory mismatch condition and two trials on average in the visual mismatch condition; subsequently, the mean number of voxels with outlier values across all trials and participants was less than one in both the auditory and visual mismatch conditions in each ROI. Analyses focused solely on the ROIs and experimental conditions for which we observed effects in the Discovery Sample for either the mean or SD of the beta series. When performing these confirmatory analyses, we opted to use a Bonferroni-corrected alpha threshold for significance of .0125, which was calculated by dividing .05 by 4, the total number of models tested.

Results

Behavioral performance for the in-scanner task for both samples is summarized in Table 2, along with data concerning average amount of movement in the scanner and the number of trials in the beta series in each condition. Average values across participants for the mean and SD of the beta series within each ROI are presented in Table 3; results for the multiple regression analysis are detailed in Table 4 for the Discovery Sample and Table 5 for the Confirmation Sample.

Across the two samples, we observed a positive relationship between reading skill and trial-by-trial neural activation variability for printed words in the left IFG *pars triangularis* (Figure 3). In this region, trial-by-trial neural activation variability for printed words not only accounted for significant variance in reading skill above and beyond mean levels of activation, but actually accounted for a greater proportion of variance in reading skill than did mean activation. The relationship between reading skill and neural activation variability in this region appears to be fairly selective for print, as we did not observe a significant relationship between reading skill and trial-by-trial variability in neural activation for spoken words in this region. For spoken words, the only relationship we observed between reading skill and trial-by-trial neural activation variability was a negative association in the left STG in the Discovery Sample; however, this finding did not hold in the Confirmation Sample. These results also appear to be fairly selective for reading ability, as we ran a secondary analysis with performance IQ as the dependent variable to assess whether neural activation variability was related to general cognitive ability. These analyses did not reveal any significant relationship between performance IQ and neural activation variability for print or speech in any region of interest. In

addition, we also re-ran the analysis with single word reading skill as the dependent variable, this time including performance IQ as a covariate in the multiple regression models. The positive relationship between reading skill and neural activation variability for print in the left IFG *pars triangularis* was still marginally significant in both the Discovery and Confirmation samples, even when accounting for individual differences in performance IQ (Discovery Sample: $p = .055$; Confirmation Sample: $p = .017$).

The above results stand in contrast to those for mean activation, which only showed a significant relationship between reading skill and activation for printed and spoken words in the left thalamus in the Discovery Sample but not the Confirmation Sample. We should also note that the alternative predictors did not show systematic patterns across both samples; however, we did observe that average Euclidean motion and the number of trials within the beta series accounted for significant variance in reading skill in a number of regions. To further test for the influence of subject motion and the number of trials in the beta series on the observed results, we ran a secondary analysis in which we relaxed the subject inclusion criteria to a maximum of .40 mm average Euclidean motion and a minimum of 50% accuracy for both the auditory and visual mismatch conditions. This resulted in a sample size of 50 for the Discovery Sample and 43 for the Confirmation Sample. The relationship between reading skill and neural activation variability for print in the left IFG *pars triangularis* was still significant in the same direction in both samples (Discovery Sample: $p = .012$; Confirmation Sample: $p = .002$).

Discussion

Our aim was to assess the relationship between reading skill in school-aged children and trial-by-trial variability in fMRI activation for print or speech. This stemmed from recent advances concerning individual differences in neural response stability in relation to reading skill (Hornickel & Kraus, 2013), as well as the potential impact of neural noise on the timing and systematic variability of processes important for reading (Hancock, Pugh, & Hoeft, 2017). Based on this previous work, we hypothesized that reading skill would be related to trial-by-trial variability in neural activation even after accounting for intrasubject differences in mean levels of task-related activation. However, the direction of this relationship remained an open question, as domains outside of reading have shown different relationships between behavioral performance and variability in the fMRI BOLD response, whether measured from moment-to-moment within blocks (Garrett et al., 2011), or from trial-to-trial in experiments employing event-related designs (Armbruster-Genc et al., 2016; Baum & Beauchamp, 2014; Haigh et al., 2016).

Trial-by-Trial Activation Variability versus Mean Activation

For each of two samples of children who performed an fMRI picture-word matching task, we entered intrasubject means and SDs of single trial beta estimates for print and speech trials into multiple regression models predicting reading skill. We observed that in the left inferior frontal gyrus *pars triangularis*, the SD of the beta series for printed words not only accounted for additional variance in reading skill that was not captured by the mean of the beta series, but actually accounted for a greater proportion of variance in reading skill than did mean levels of activation. This effect held across the two samples, despite differences in participants, scanners,

and slight differences in trial context. Moreover, these effects were fairly selective for print, as we did not observe a relationship between reading skill and activation variability for spoken words in this region; the only effect we observed for spoken words was a negative relationship between reading skill and activation variability in the left STG in the Discovery Sample that was not observed in the Confirmation Sample. In contrast to the findings for activation variability, for mean activation, the only relationships we observed were positive correlations between reading skill and mean activation for both printed and spoken words in the left thalamus in the Discovery Sample, supporting previous studies documenting the important contributions of the thalamus to reading (Galaburda et al., 1985; Pugh et al., 2013); however, these relationships were not observed in the Confirmation Sample.

The left IFG has long been implicated as a critical part of the skilled reading network in adults, with more anterior and lateral subregions of IFG such as *pars triangularis* thought to be involved in semantic processing (Bookheimer, 2002; Poldrack et al., 1999). Moreover, activation of the left IFG, as well as connectivity between the left IFG and certain components of the reading network, has been associated with age-related increases over the course of reading development, and this region has been linked to processes such as phonological segmentation and covert articulation (Bitan et al., 2007; Schlaggar et al., 2002; Turkeltaub et al., 2003). However, the current study is the first time that individual differences in reading ability in children have been associated with variability in fMRI activation for printed words in this region.

The Direction of the Relationship between Reading Skill and BOLD Signal Variability

In the left IFG *pars triangularis*, trial-by-trial variability in neural activation for print was positively related to reading skill. A positive relationship between BOLD signal variability and behavioral performance has been previously observed by Garrett et al. (2011), who found that increased levels of BOLD signal variability were associated with faster and more consistent behavioral performance in younger versus older adults across a range of cognitive tasks, as well as by Armbruster-Genc et al. (2016), who found that increased levels of BOLD signal variability were associated with greater cognitive flexibility in adults, which manifested as reduced behavioral switching costs in a task-switching paradigm.

Based on EEG measures in children, McIntosh et al. (2008) suggest that increased neural variability reflects a greater dynamic range of cognitive states as well as a greater ability to transition between them, which perhaps translates to a greater ability to adapt to the environment. It is possible that the increased neural variability observed in the better readers in the current study could reflect greater neural adaptability; however, we interpret this with caution given that we assessed trial-by-trial variability as opposed to moment-to-moment variability at a finer within-trial timescale. Furthermore, increased neural variability may not always have a positive effect on behavioral performance. Armbruster-Genc et al. (2016) observed that one of the brain regions that showed a positive association between neural variability and cognitive flexibility – that is, the left inferior frontal junction – actually showed a negative association with cognitive stability, which manifested as more extensive behavioral costs for distractor inhibition. This dissociation in terms of directionality may depend on the level of hierarchical organization in which a brain region is situated as well as the extent to

461 which a task is weighted towards sensory versus cognitive processing. For example, individuals
462 with autism – a neurodevelopmental disorder that can co-occur with reading disability – have
463 shown increased fMRI BOLD signal variability in primary sensory areas in response to low-level
464 sensory stimulation, and this finding has been used to explain why individuals with autism may
465 experience difficulties in highly sensory environments (Haigh et al., 2016). This distinction
466 between sensory and cognitive processing may help reconcile the current results with the
467 observation that low-level neural responses to speech sounds show greater variability in
468 children with reading disability compared to typically developing children (Hornickel & Kraus,
469 2013). In addition, these findings – as well as future experiments that more directly tease apart
470 sensory versus cognitive processing – may inform the neural noise hypothesis put forth by
471 Hancock et al. (2017) by elucidating the conditions that promote greater versus lesser neural
472 variability in developing readers as well as how these relationships pattern across different
473 brain regions as a function of reading experience. The differentiation of the role of random
474 neural "noise" versus systemic components that drive "dynamic range" or "adaptability" within
475 such greater neural variability indices may be a critical conceptual and analytic challenge.

476 In the electrophysiological literature, He and Zempel (2013) have asserted that a certain
477 amount of neural variability is beneficial, but if the level of variability is too high, brain activity
478 could be scattered across too wide a range, which could be detrimental. Thus, it is possible that
479 the relationship between neural variability and behavioral performance is non-monotonic, and
480 that we are only observing the ascending portion of an inverted U-shaped curve. Future studies
481 could address this possibility by including children who are more severely impaired than the
482 poorest readers in the present sample.

Possible Mechanisms of Neural Variability

The increased neural variability observed in the more skilled readers in the current study could be the result of spontaneous fluctuations in the BOLD signal that are intrinsically generated in the brain and not attributable to specific inputs or outputs (Fox & Raichle, 2007; Fox, Snyder, Vincent, & Raichle, 2007). These spontaneous fluctuations may serve to coordinate neuronal activity between distal brain regions, and may be the product of changes in the power of high-frequency electrical activity such as the gamma band (Leopold, Murayama, & Logothetis, 2003). These changes in gamma oscillation frequency may in turn be associated with differences in GABA concentrations and their resulting influence on the balance of neural excitation and inhibition (Muthukumaraswamy, Edden, Jones, Swettenham, & Singh, 2009). It is also possible that the balance of excitation and inhibition could have been influenced by glutamatergic inputs, especially given recent findings documenting an association between glutamate concentrations and reading skill (Pugh et al., 2014), as well as similar support from animal models (Che, Truong, Fitch, & LoTurco, 2015). Based on these findings, future investigations should target the neural mechanisms of BOLD signal variability and their links to other neurobiological indices, including neural oscillations, neurochemistry, indices of neural noise, and neuroanatomical measures (Becker, Reinacher, Freyer, Villringer, & Ritter, 2011).

Conclusions and Future Directions

Overall, this investigation lends support to work advocating for the added value of evaluating intrasubject variability in brain signals compared to solely evaluating mean levels of neural activity (Faisal, Selen, & Wolpert, 2008; Garrett et al., 2011, 2010; Garrett et al., 2013; Pernet, Sajda, & Rousselet, 2011), and highlights the importance of considering individual difference

dimensions beyond subject age as contributors to, or reflections of, individual differences in neural activation variability (Grady & Garrett, 2014; McIntosh et al., 2008). Furthermore, the current findings constitute a critical first step in considering the role of adaptability in developing brain systems involved in reading, and motivate future investigations concerning the mechanistic link between neural activation variability, neural noise, and reading skill. Finally, from an applied standpoint, these results beg the tantalizing question of whether trial-by-trial activation variability in reading-related brain areas could serve as a useful biomarker for clinically relevant phenotypes such as response to intervention for reading disability.

References

Armbruster-Genc, D. J. N., Ueltzhoffer, K., & Fiebach, C. J. (2016). Brain signal variability differentially affects cognitive flexibility and cognitive stability. *Journal of Neuroscience*, *36*(14), 3978–3987. http://doi.org/1.1523/JNEUROSCI.2517-14.2016

Baum, S. H., & Beauchamp, M. S. (2014). Greater BOLD variability in older compared with younger adults during audiovisual speech perception. *PLoS ONE*, *9*(10). http://doi.org/1.1371/journal.pone.0111121

Becker, R., Reinacher, M., Freyer, F., Villringer, A., & Ritter, P. (2011). How ongoing neuronal oscillations account for evoked fMRI variability. *Journal of Neuroscience*, *31*(30), 11016–11027. http://doi.org/1.1523/JNEUROSCI.0210-11.2011

Bitan, T., Cheon, J., Lu, D., Burman, D. D., Gitelman, D. R., Mesulam, M. M., & Booth, J. R. (2007). Developmental changes in activation and effective connectivity in phonological processing. *NeuroImage*, *38*(3), 564-575.

Bookheimer, S. (2002). Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, *25*, 151–88. http://doi.org/1.1146/annurev.neuro.25.112701.142946

Branum-Martin, L., Fletcher, J. M., & Stuebing, K. K. (2013). Classification and identification of reading and math disabilities: The special case of comorbidity. *Journal of Learning Disabilities*, *46*(6), 490-499.

Burgund, E. D., Kang, H. C., Kelly, J. E., Buckner, R. L., Snyder, A. Z., Petersen, S. E., & Schlaggar, B. L. (2002). The feasibility of a common stereotactic space for children and adults in fMRI studies of development. *NeuroImage*, *17*(1), 184–200. http://doi.org/10.1006/nimg.2002.1174

Centanni, T. M., Booker, A. B., Sloan, A. M., Chen, F., Maher, B. J., Carraway, R. S., … Kilgard, M. P. (2014). Knockdown of the dyslexia-associated gene *Kiaa0319* impairs temporal responses to speech stimuli in rat primary auditory cortex. *Cerebral Cortex*, *24*(7), 1753–1766. http://doi.org/1.1093/cercor/bht028

Centanni, T. M., Chen, F., Booker, A. M., Engineer, C. T., Sloan, A. M., Rennaker, R. L., … Kilgard, M. P. (2014). Speech sound processing deficits and training-induced neural plasticity in rats with dyslexia gene knockdown. *PloS ONE*, *9*(5), e98439. http://doi.org/1.1371/journal.pone.0098439

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–73. http://doi.org/1.1006/cbmr.1996.0014

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test*. Pearson Assessments.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, *68*(3), 589-599.

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*, 292–303. http://doi.org/1.1038/nrn2258

Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, *8*(9), 700–711. http://doi.org/1.1038/nrn2201

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of*

America, *102*(27), 9673–8. http://doi.org/1.1073/pnas.0504136102

Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron*, *56*(1), 171–184. http://doi.org/1.1016/j.neuron.2007.08.023

Frost, S. J., Landi, N., Mencl, W. E., Sandak, R., Fulbright, R. K., Tejada, E. T., … Pugh, K. R. (2009). Phonological awareness predicts activation patterns for print and speech. *Annals of Dyslexia*, *59*(1), 78–97. http://doi.org/1.1007/s11881-009-0024-y

Galaburda, A. M., Sherman, G. F., Rosen, G. D., Aboitiz, F., & Geschwind, N. (1985). Developmental dyslexia: four consecutive cases with cortical anomalies. *Annals of Neurololgy*, *18*, 222–233. http://dx.doi.org/1.1002/ana.410180210

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2011). The Importance of Being Variable. *Journal of Neuroscience*, *31*(12), 4496–4503. http://doi.org/1.1523/JNEUROSCI.5641-1.2011

Garrett, D. D., Kovacevic, N., McIntosh, a. R., & Grady, C. L. (2010). Blood oxygen level-dependent signal variability is more than just noise. *Journal of Neuroscience*, *30*(14), 4914–4921. http://doi.org/1.1523/JNEUROSCI.5166-09.2010

Garrett, D. D., Samanez-Larkin, G. R., MacDonald, S. W. S., Lindenberger, U., McIntosh, A. R., & Grady, C. L. (2013). Moment-to-moment brain signal variability: A next frontier in human brain mapping? *Neuroscience and Biobehavioral Reviews*, *37*(4), 610–624. http://doi.org/1.1016/j.neubiorev.2013.02.015

Grady, C. L., & Garrett, D. D. (2014). Understanding variability in the BOLD signal and why it matters for aging. *Brain Imaging and Behavior*, *8*(2), 274–283. http://doi.org/1.1007/s11682-013-9253-0

Grinband, J., Wager, T. D., Lindquist, M., Ferrera, V. P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *NeuroImage*, *43*(3), 509–52. http://doi.org/1.1016/j.neuroimage.2008.07.065

Haigh, S. M., Gupta, A., Barb, S. M., Glass, S. A. F., Minshew, N. J., Dinstein, I., … Behrmann, M. (2016). Differential sensory fMRI signatures in autism and schizophrenia: Analysis of amplitude and trial-to-trial variability. *Schizophrenia Research*, *175*(1–3), 12–19. http://doi.org/1.1016/j.schres.2016.03.036

Hancock, R., Pugh, K. R., & Hoeft, F. (2017). Neural noise hypothesis of developmental dyslexia. *Trends in Cognitive Sciences*, *21*(6), 434-448.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1-28.

Hornickel, J., & Kraus, N. (2013). Unstable representation of sound: a biological marker of dyslexia. *Journal of Neuroscience*, *33*(8), 3500–3504. http://doi.org/1.1523/JNEUROSCI.4205-12.2013

Jasińska, K. K., Molfese, P. J., Kornilov, S. A., Mencl, W. E., Frost, S. J., Lee, M., Pugh, K. R., Grigorenko, E. L., & Landi, N. (2016). The BDNF Val66Met polymorphism influences reading ability and patterns of neural activation in children. PLoS ONE, 11(8), e0157449.

Landi, N., Frost, S. J., Mencl, W. E., Preston, J. L., Jacobsen, L. K., Lee, M., … Grigorenko, E. L. (2013). The COMT Val/Met polymorphism is associated with reading-related skills and consistent patterns of functional neural activation. *Developmental Science*, *16*(1), 13–23. http://doi.org/1.1111/j.1467-7687.2012.0118.x

Leopold, D. A., Murayama, Y., & Logothetis, N. K. (2003). Very slow activity fluctuations in

605 monkey visual cortex: implications for functional brain imaging. *Cerebral Cortex*, *13*,
606 422-433.
607 Lund, T. E., Nørgaard, M. D., Rostrup, E., Rowe, J. B., & Paulson, O. B. (2005). Motion or
608 activity: Their role in intra- and inter-subject variation in fMRI. *NeuroImage*, *26*(3),
609 960-964.
610 Martin, A., Schurz, M., Kronbichler, M., & Richlan, F. (2015). Reading in the brain of children
611 and adults: A meta-analysis of 40 functional magnetic resonance imaging studies.
612 *Human Brain Mapping*, *36*(5), 1963–1981. http://doi.org/1.1002/hbm.22749
613 McIntosh, A. R., Kovacevic, N., & Itier, R. J. (2008). Increased brain signal variability
614 accompanies lower behavioral variability in development. *PLoS Computational Biology*,
615 *4*(7). http://doi.org/1.1371/journal.pcbi.1000106
616 Mumford, J. a., Turner, B. O., Ashby, F. G., & Poldrack, R. a. (2012). Deconvolving BOLD
617 activation in event-related designs for multivoxel pattern classification analyses.
618 *NeuroImage*, *59*(3), 2636–2643. http://doi.org/1.1016/j.neuroimage.2011.08.076
619 Muthukumaraswamy, S. D., Edden, R. a E., Jones, D. K., Swettenham, J. B., & Singh, K. D.
620 (2009). Resting GABA concentration predicts peak gamma frequency and fMRI
621 amplitude in response to visual stimulation in humans. *Proceedings of the National*
622 *Academy of Sciences of the United States of America*, *106*(20), 8356–8361.
623 http://doi.org/1.1073/pnas.0900728106
624 Norton, E. S., Beach, S. D., & Gabrieli, J. De. (2015). Neurobiology of dyslexia. *Current Opinion*
625 *in Neurobiology*, *30C*, 73–78. http://doi.org/1.1016/j.conb.2014.09.007
626 Pennington, B. F., Santerre–Lemmon, L., Rosenberg, J., MacDonald, B., Boada, R., Friend, A.,
627 Leopold, D. R., Samuelsson, S., Byrne, B., Willcutt, E. G., & Olson, R. K. (2012). Individual
628 prediction of dyslexia by single versus multiple deficit models. *Journal of Abnormal*
629 *Psychology*, *121*(1), 212.
630 Pernet, C. R., Sajda, P., & Rousselet, G. A. (2011). Single-trial analyses: Why bother?
631 *Frontiers in Psychology*, *2*(November), 1–2. http://doi.org/1.3389/fpsyg.2011.00322
632 Poldrack, R. A, Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D.
633 (1999). Functional specialization for semantic and phonological processing in the left
634 inferior prefrontal cortex. *NeuroImage*, *10*(1), 15–35.
635 http://doi.org/1.1006/nimg.1999.0441
636 Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious
637 but systematic correlations in functional connectivity MRI networks arise from subject
638 motion. *NeuroImage*, *59*(3), 2142–2154.
639 http://doi.org/1.1016/j.neuroimage.2011.1.018
640 Preston, J. L., Molfese, P. J., Frost, S. J., Mencl, W. E., Fulbright, R. K., Hoeft, F., … Pugh, K. R.
641 (2015). Print-Speech Convergence Predicts Future Reading Outcomes in Early Readers.
642 *Psychological Science*, 1-1. http://doi.org/1.1177/0956797615611921
643 Pugh, K. R., Frost, S. J., Rothman, D. L., Hoeft, F., Del Tufo, S. N., Mason, G. F., … Fulbright, R. K.
644 (2014). Glutamate and choline levels predict individual differences in reading ability in
645 emergent readers. *Journal of Neuroscience*, *34*(11), 4082–9.
646 http://doi.org/1.1523/JNEUROSCI.3907-13.2014
647 Pugh, K. R., Landi, N., Preston, J. L., Mencl, W. E., Austin, A. C., Sibley, D., … Frost, S. J. (2013).
648 The relationship between phonological and auditory processing and brain
649 organization in beginning readers. *Brain and Language*, *125*(2), 173–83.
650 http://doi.org/1.1016/j.bandl.2012.04.004

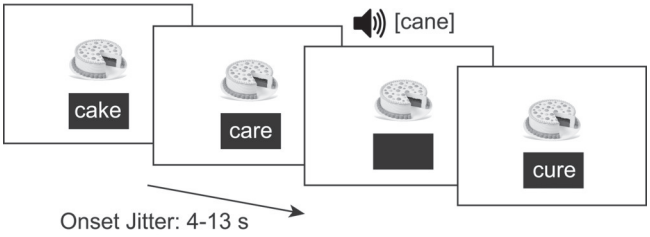
651 Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during
652 distinct stages of a cognitive task. *NeuroImage*, *23*(2), 752–63.
653 http://doi.org/1.1016/j.neuroimage.2004.06.035
654 Schlaggar, B. L., Brown, T. T., Lugar, H. M., Visscher, K. M., Miezin, F. M., & Petersen, S. E.
655 (2002). Functional neuroanatomical differences between adults and school-age
656 children in the processing of single words. *Science*, *296*(5572), 1476-1479.
657 Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). ROC analysis of statistical methods used
658 in functional MRI: individual subjects. *NeuroImage*, *9*(3), 311–29.
659 http://doi.org/1.1006/nimg.1999.0402
660 Tibshirani, R., & Leisch, F. (2015). bootstrap: Functions for the Book "An Introduction to the
661 Bootstrap". R package version 2015.2.
662 Turkeltaub, P. E., Gareau, L., Flowers, D. L., Zeffiro, T. A., & Eden, G. F. (2003). Development
663 of neural mechanisms for reading. *Nature Neuroscience*, *6*(6), 767-773.
664 Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition.
665 Springer, New York.
666 Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of*
667 *Phonological Processing: CTOPP2*. Austin, TX: Pro-Ed.
668 Wechsler, D., & Hsiao-pin, C. (2011). *WASI-II: Wechsler Sbbreviated Scale of Intelligence*.
669 Pearson.
670 Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of*
671 *Achievement*. Itasca, IL: Riverside Publishing.
672 Yarkoni, T., Barch, D. M., Gray, J. R., Conturo, T. E., & Braver, T. S. (2009). BOLD correlates of
673 trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI
674 analysis. *PloS ONE*, *4*(1), e4257. http://doi.org/1.1371/journal.pone.0004257
675
676

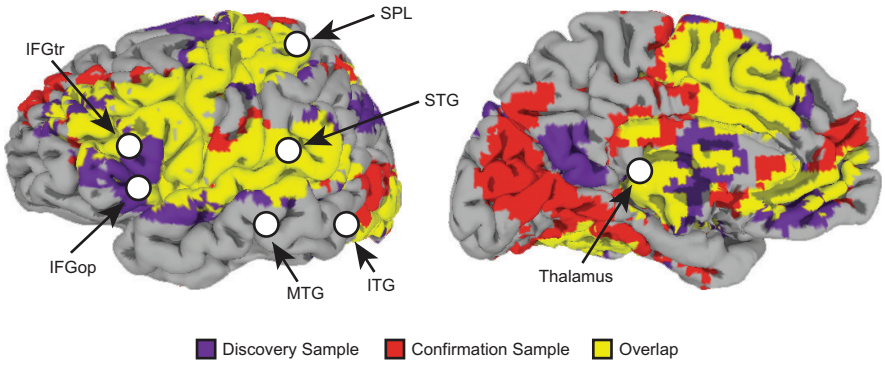
Figure Captions

Figure 1. A sample trial sequence for the fMRI task.

Figure 2. Location of each of the ROIs tested in the multiple regression analyses. These ROIs, shown in white, are peak coordinates from the Martin et al. (2015) meta-analysis, and are overlaid on a conjunction map that shows the overlap in task-related activation between the Discovery and Confirmation samples (each corrected at FDR < .01; for more details on how this map was constructed, refer to the Methods section). IFGtr = inferior frontal gyrus *pars triangularis*; IFGop = inferior frontal gyrus *pars opercularis*; SPL = superior parietal lobule; STG = superior temporal gyrus; MTG = middle temporal gyrus; ITG = inferior temporal gyrus

Figure 3. Partial correlation results for the multiple regression analyses performed for the left inferior frontal gyrus *pars triangularis* ROI. The *x*-axis specifies adjusted trial-by-trial neural activation variability, which corresponds to residuals from a model where the standard deviation (SD) of the beta series is the dependent variable and the mean of the beta series as well as any significant predictors of non-interest (i.e., age, average Euclidean motion, and/or the number of trials in the beta series) are the regressors; the *y*-axis specifies adjusted reading scores, which are the size of the residuals from a model where Letter Word Identification raw scores are the dependent variable, and the regressors are the mean of the beta series as well as the same predictors of non-interest. The regression line is superimposed on the plot; the shaded region represents the 95% confidence interval.





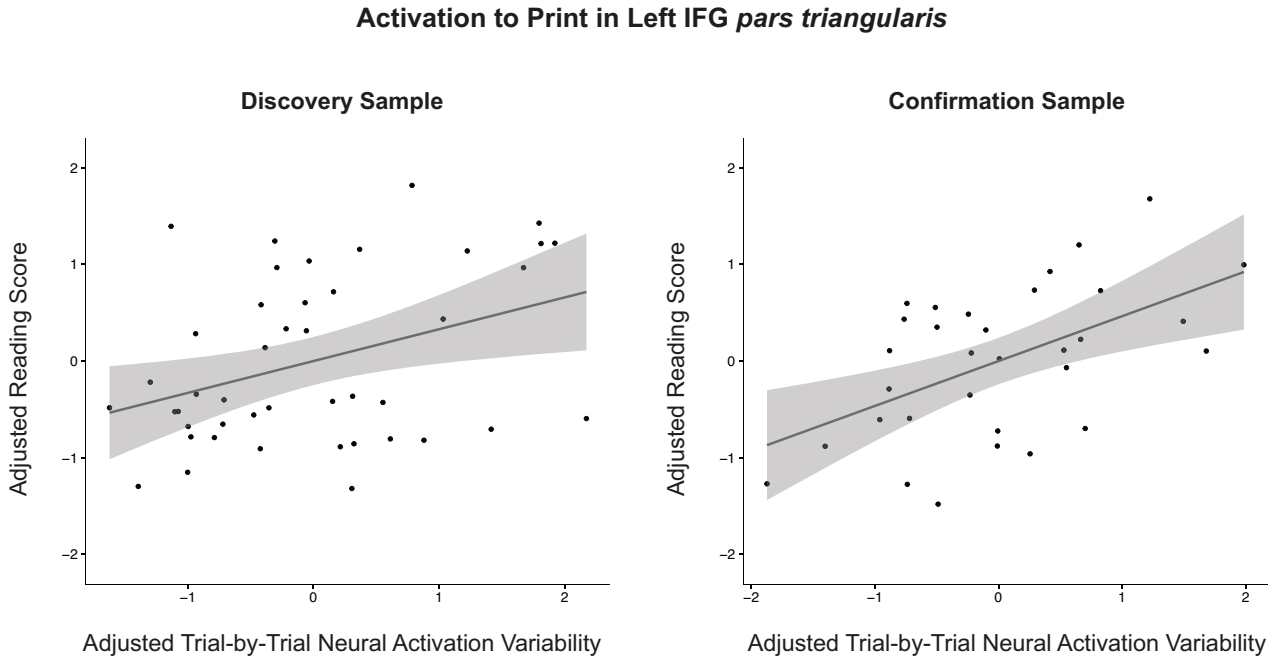


Table 1. Descriptive information concerning the two groups of children who performed the fMRI experiment.

Assessment	Measure	Discovery Sample (*N* = 44; 18 female)			Confirmation Sample (*N* = 32; 14 female)		
		Mean	SD	Range	Mean	SD	Range
–	Age	9.3	0.6	7.8–11.3	9.4	1.1	7.5-11.3
WJ-III Letter Word ID – raw score	Single word reading	43.0	9.0	31-61	46.8	12.1	23-69
WJ-III Letter Word ID – standard score	Single word reading	95.5	13.9	67-124	103.1	14.9	76-133
WASI Full Scale IQ-2[1]	Intelligence	99.6	15.8	80-140	110.8	14.6	76-138
CTOPP phonological awareness – composite score[2]	Phonological awareness	85.5	13.4	65-112	104.4	15.2	67-145
PPVT – standard score[3]	Receptive vocabulary	103.3	16.9	73-135	114.9	12.5	84-135

Note. WJ-III = Woodcock-Johnson Tests of Cognitive Abilities; WASI = Wechsler Abbreviated Scale of Intelligence; CTOPP = Comprehensive Test of Phonological Processing; PPVT = Peabody Picture Vocabulary Test. For the Discovery Sample, the following versions were used: WASI-II, CTOPP-2, PPVT-4. For the Confirmation Sample, the following versions were used: WASI-I, CTOPP-1, PPVT-3.

[1]Full scale IQ-4 was measured instead of Full Scale IQ-2 for one participant in the Discovery Sample. Furthermore, Full Scale IQ-2 is missing from one participant in the Confirmation Sample.

[2]CTOPP scores are missing from two participants in the Confirmation Sample.

[3]PPVT scores are missing from one participant in the Discovery Sample and one participant in the Confirmation Sample.

Table 2. MRI quality control parameters and performance for the in-scanner picture cue-target word identification task for the two groups of children who performed the fMRI experiment.

Measure	Discovery Sample			Confirmation Sample		
	Mean	SD	Range	Mean	SD	Range
Average motion per brain volume (mm/TR)	.12	.05	.04-.23	.13	.05	.05-.25
Visual mismatch condition (Print)						
Number of trials in the beta series	19.6	2.8	12-25	26.7	3.9	19-37
Percent accuracy	90.8	7.3	76-100	91.3	7.4	75-100
Mean reaction time for correct trials	1670	348	1159-3013	1479	355	989-2125
SD of reaction time for correct trials	506	203	181-1001	442	166	169-700
Auditory mismatch condition (Speech)						
Number of trials in the beta series	21.0	2.1	17-24	27.3	3.7	20-36
Percent accuracy	96.5	4.7	80-100	93.0	4.9	81-100
Mean reaction time for correct trials	1698	296	1121-2335	1419	237	1072-1933
SD of reaction time for correct trials	456	194	135-1015	373	123	171-633

Table 3. Location of each of the ROIs selected for the multiple regression analysis as well as average values across participants for the mean and standard deviation (SD) of the beta series within each ROI.

					Discovery Sample				Confirmation Sample			
		Centre of Mass (Talairach)[1]			Print		Speech		Print		Speech	
Region	Abbreviation	x	y	z	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Left inferior frontal gyrus	IFGop	-49	22	3	1.11	2.83	1.24	2.56	.41	3.30	.88	3.02
pars opercularis												
Left inferior frontal gyrus	IFGtr	-51	24	15	.58	1.83	.61	1.70	.49	2.10	.54	1.96
pars triangularis												
Left superior parietal lobule	SPL	-22	-48	50	.16	1.13	.12	1.05	.03	1.18	.07	1.16
Left middle temporal gyrus	MTG	-57	-25	-4	.24	1.69	.47	1.52	-.17	1.84	.37	1.79
Left inferior temporal gyrus	ITG	-52	-59	-9	.66	2.28	.07	2.44	.55	3.18	.19	2.97
Left superior temporal	STG	-55	-30	16	.45	1.91	1.61	1.89	.11	1.89	1.53	1.82
gyrus												
Left thalamus	–	-12	-25	10	.59	1.29	.45	1.23	.27	1.15	.45	1.13

Note: Each ROI was 891 mm^3 in size (33 voxels; 3 mm isotropic).
[1]LPI orientation.

Table 4. Multiple regressions for the Discovery Sample. Each model quantifies the contribution of mean activation and trial-by-trial activation variability to WJ-III Letter Word ID raw scores.

		Print							Speech						
Region	Regressor	β	*SE*	*ΔAIC*	*ΔBIC*	ΔR^2_{adj}	*F*	*p*	β	*SE*	*ΔAIC*	*ΔBIC*	ΔR^2_{adj}	*F*	*p*
Left IFGop	Mean activation	-.072	.146	1.73	3.52	-.015	.246	.623	-.043	.158	1.92	3.71	-.021	.075	.786
	Activation variability	.177	.146	.405	2.19	.009	1.48	.231	.208	.179	.536	2.32	.008	1.35	.252
Left IFGtr	Mean activation	.042	.130	1.88	3.67	-.015	.106	.746	-.043	.150	1.91	3.69	-.021	.083	.775
	Activation variability	.329	.129	-4.66	-2.87	.095	6.53	.014	.150	.150	.904	2.69	<.001	1.01	.321
Left SPL	Mean activation	-.006	.136	2.00	3.78	-.019	.002	.964	-.011	.150	1.99	3.78	-.023	.006	.939
	Activation variability	.221	.141	-.644	1.14	.028	2.48	.123	.086	.166	1.71	3.49	-.017	.267	.608
Left MTG	Mean activation	.046	.137	1.88	3.66	-.017	.111	.741	.057	.153	1.85	3.63	-.020	.138	.712
	Activation variability	.211	.137	-.537	1.25	.026	2.37	.131	.107	.153	1.47	3.25	-.012	.490	.488
Left ITG	Mean activation	-.093	.138	1.51	3.29	-.011	.449	.507	.103	.171	1.61	3.40	-.015	.364	.549
	Activation variability	.102	.141	1.42	3.21	-.009	.527	.472	-.187	.171	.739	2.52	.005	1.19	.281
Left STG	Mean activation	-.039	.140	1.92	3.70	-.018	.077	.782	.224	.179	.353	2.14	.013	1.56	.218
	Activation variability	.102	.141	1.42	3.21	-.009	.530	.471	-.399	.179	-3.01	-1.23	.088	4.95	.032
Left thalamus	Mean activation	.413	.119	-9.52	-7.74	.160	12.0	.001	.304	.147	-2.39	-.604	.072	4.30	.044
	Activation variability	.160	.120	.095	1.88	.011	1.77	.191	-.156	.147	.802	2.59	.003	1.13	.294

Note: In all models, removal of the activation variability term did not impact the significance of the mean activation term.

Table 5. Multiple regressions for the Confirmation Sample. Each model quantifies the contribution of mean activation and trial-by-trial activation variability to WJ-III Letter Word ID raw scores.

Experimental Condition	Region	Regressor	β	*SE*	*ΔAIC*	*ΔBIC*	ΔR^2_{adj}	*F*	p^1
Print	Left IFGtr	Mean activation	-.062	.139	1.76	3.23	-.014	.200	.658
		Activation variability	.465	.142	-8.70	-7.23	.171	10.7	.003
	Left thalamus	Mean activation	-.003	.141	2.00	3.47	-.021	.001	.982
		Activation variability	-.037	.145	1.92	3.39	-.019	.065	.801
Speech	Left STG	Mean activation	.214	.152	-.189	1.28	.022	1.98	.170
		Activation variability	-.179	.152	.461	1.93	.009	1.38	.250
	Left thalamus	Mean activation	.178	.156	.545	2.01	.007	1.30	.263
		Activation variability	-.121	.156	1.32	2.79	-.009	.598	.446

Note: In all models, removal of the activation variability term did not impact the significance of the mean activation term.
[1] For this confirmatory analysis, we adopted a Bonferroni-corrected alpha threshold for significance of .0125 (.05 divided by 4, the total number of models tested).