Published in final edited form as:

Neuroimage. 2021 November 15; 242: 118476. doi:10.1016/j.neuroimage.2021.118476.

Theory-driven classification of reading difficulties from fMRI data using Bayesian latent-mixture models

Noam Siegelman[a,*], **Mark R. van den Bunt**[a], **Jason Chor Ming Lo**[a], **Jay G. Rueckl**[a,b], **Kenneth R. Pugh**[a,b,c]

[a]Haskins Laboratories, USA

[b]University of Connecticut, USA

[c]Yale University, USA

Abstract

Decades of research have led to several competing theories regarding the neural contributors to impaired reading. But how can we know which theory (or theories) identifies the types of markers that indeed differentiate between individuals with reading disabilities (RD) and their typically developing (TD) peers? To answer this question, we propose a new analytical tool for theory evaluation and comparison, grounded in the Bayesian latent-mixture modeling framework. We start by constructing a series of latent-mixture classification models, each reflecting one existing theoretical claim regarding the neurofunctional markers of RD (highlighting network-level differences in either mean activation, inter-subject heterogeneity, inter-region variability, or connectivity). Then, we run each model on fMRI data alone (i.e., while models are blind to participants' behavioral status), which enables us to interpret the fit between a model's classification of participants and their behavioral (known) RD/TD status as an estimate of its explanatory power. Results from n=127 adolescents and young adults (RD: n=59; TD: n=68) show that models based on network-level differences in mean activation and heterogeneity failed to differentiate between TD and RD individuals. In contrast, classifications based on variability and connectivity were significantly associated with participants' behavioral status. These findings suggest that differences in inter-region variability and connectivity may be better network-level markers of RD than mean activation or heterogeneity (at least in some populations and tasks). More broadly, the results demonstrate the promise of latent-mixture modeling as a theory-driven tool for evaluating different theoretical claims regarding neural contributors to language disorders and other cognitive traits.

[*]Corresponding author at: Haskins Laboratories, 300 George St., New Haven, CT 06511, USA. noam.siegelman@yale.edu (N. Siegelman).

Credit authorship contribution statement

Noam Siegelman: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization. **Mark R. van den Bunt:** Methodology, Formal analysis, Writing – review & editing, Visualization. **Jason Chor Ming Lo:** Writing – review & editing. **Jay G. Rueckl:** Supervision, Funding acquisition, Writing – review & editing. **Kenneth R. Pugh:** Supervision, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118476.

Keywords

Reading disabilities; Reading; Bayesian modeling; Latent-mixture modeling; Neurofunctional markers; fMRI data analysis

1. Introduction

Reading disability (RD) is the most common neurodevelopmental disorder, with an estimated prevalence of 10-15% of children (Fletcher et al., 2007; Lyon, 1995). It is a life-long disorder, with detrimental effects lasting into adolescence and adulthood (Bruck, 1992; Shaywitz et al., 1999, 2003). Neuroscientists have long been interested in the neurofunctional markers of poor reading. Works comparing the brain activity during reading of participants with RD to typically developing (TD) readers can be traced back to early studies using EEG (Colon et al., 1979; Sklar et al., 1972), a line of work that became more prevalent with the advent of hemodynamic tools in the 1990′s (Rumsey et al., 1992; Salmelin et al., 1996; Shaywitz et al., 1998). Since then, a large and constantly growing number of studies followed up on these earlier observations with the aim of unveiling the neural contributors to impaired reading.[1]

There is no doubt that the increase in the amount of functional neuroimaging data on reading disabilities is laudable. Among other advantages, a large amount of data enables data accumulation across studies and highly powered meta-analyses (e.g. Paulesu et al., 2014; Richlan et al., 2009). Yet a large number of studies also brings a unique challenge: Decades of research resulted in numerous observations of differences between individuals with RD and their TD counterparts, leading to several different theoretical claims regarding the neural markers of impaired reading. Yet, there are reasons to believe that some of these observations may not be replicable, and consequently that the theories that were construed based on them have limited explanatory power. This is because a large number of studies increases the risk that at least some observations reflect Type-I errors (a problem that is further exacerbated by multiple comparisons per study, e.g., Lindquist and Gelman, 2009, and flexibility in analytical practices, e.g. Botvinik-Nezer et al., 2020; Hong et al., 2019).[2] Moreover, the generally limited sample sizes in neuroimaging studies means that many positive findings might not reflect true effects (e.g., Button et al., 2013; Szucs and Ioannidis, 2017). How can we then identify from all reported observations, and from the theories that were built in light of these findings, the markers that indeed differentiate between readers with and without RD?

The aim of the current paper is to provide a formal framework that directly evaluates the predictive power of different theoretical claims regarding neurofunctional markers. In a nutshell, we do so by adopting a Bayesian latent-mixture approach, a sub-type of generative modeling where classification models are constructed in a theory-driven manner

[1] In fact, a Web of Science search with the keywords “fMRI” AND (“dyslexia” OR “reading disability”) on publications from 1996 to 2019 shows that the average number of publications per year grew from 4.0 in 1996-2000, to 37.2 in 2015-2019.

[2] To clarify, we do not claim that all reasons behind the limited replicability in the field are statistical in nature. Other relevant factors include variability in design, diagnostic criterion, and potentially, the studied language.

(see details below). Here we adapt the latent-mixture modeling approach to build a series of classification models, each reflecting one theoretical claim regarding the markers that distinguish between individuals with and without RD in functional magnetic resonance imaging (fMRI) data of reading. That is, each of the models we build "translates" a theoretical claim regarding neurofunctional markers of RD into an explicit and formal generative model: By 'explicit' we mean a model where all assumptions are clearly stated, and by 'generative' we mean a model that specifies and estimates a set of latent parameters that presumably gave rise to the observed data. Then, each model is fitted to fMRI data from different individuals (some with and some without RD), while being blind to individuals' actual (i.e., behavioral) RD status. This procedure results in a group classification parameter for each individual, reflecting the model's certainty in classifying a person into the RD vs. TD group. We then compare the group classification parameter estimated by each model to the actual group classification of each participant, to examine which model(s) produces group classification (based on fMRI data only) that fits participants' actual group membership. This enables us to draw conclusions regarding the explanatory power (or lack thereof) of the different theoretical claims that each of these models represents.

Before diving into details, we wish to emphasize from the onset the major differences - as well as commonalities - between our approach and standard analytical frameworks. The vast majority of functional neuroimaging studies are set to test a specific hypothesis or theory, most commonly using univariate statistics (e.g., is activation in region X different between individuals with and without RD?). The univariate approach differs from our framework in two major aspects. First, univariate analysis is meant to test a single a priori hypothesis, and is therefore not suited for evaluating and contrasting multiple competing theories. In contrast, in our use of the latent-mixture modeling framework we evaluate and compare the predictive power of different theories. The second difference is that univariate statistics are more powerful when examining predictions that are confined to specific regions or a small number of regions (or else a correction for multiple corrections is needed, resulting in reduced power). The latent-mixture models we use, however, is particularly geared towards examining network-level (rather than region-specific) markers.

Other approaches to neurofunctional data analysis are multivariate and data-driven, including in particular Machine Learning algorithms (see Hoeft et al., 2011; Tamboer et al., 2016 for applications in the context of RD). What is shared between these data-driven approaches and our latent-mixture approach is that both focus on network-level differences. Indeed, we apply the latent-mixture approach to test claims about brain activity across wide networks, or even the whole brain (but see General Discussion for suggested extensions to examine specific regions-of-interests, ROIs). Crucially, the central difference between our approach and Machine Learning and other data-driven approaches (e.g., Connectome-based predictive analysis, Shen et al., 2017) are that the approaches from the latter class are not meant to test or evaluate theories: The signal that they look for as differentiating between groups of individuals (e.g., RD/TD) is not defined or constrained by theory. Instead, data-driven predictive frameworks aim to maximize classification performance by revealing the parts of the signal that best differentiates between the relevant groups of individuals - yet this signal may very well be non-transparent theoretically. In contrast, our approach does not aim to maximize classification performance, but instead to directly evaluate and compare

theories (i.e., to maximize theoretical transparency). Methodologically, this is reflected by the fact that our models classify individuals based on theorized neural differences between TD and RD readers while being blind to their actual (i.e., behavioral) group status.

2. Directly testing theories using a Bayesian latent-mixture modeling framework

The Latent-mixture approach we utilize here is a sub-type within the broader Bayesian generative modeling framework. As such, it shares many of its basic attributes with Bayesian modeling more generally (including the specification of prior distributions, the update of priors given data to estimate posterior distributions, and the interpretation of posterior distributions as reflecting researchers' current beliefs). For readers who are not familiar with the basics of Bayesian inference, we provide a brief overview of the approach in the Supplementary Materials S1. Importantly, we chose the Latent-mixture modeling approach because, as we explicate below (see Methods, sub-section Bayesian Latent-mixture models), it is particularly suited for evaluating competing theories regarding which parameters contribute to the classification of individuals into two groups of subjects (in the current case – individuals with and without RD). Also note that related implementations of this approach were previously used in different domains using behavioral data (e.g., Ortega et al., 2012; Siegelman et al., 2019).

In the current paper we applied the Bayesian latent-mixture framework to investigate the neurofunctional markers of reading disabilities. Thus, we built a series of latent-mixture models, each of them reflecting one theoretical claim regarding the markers that differentiate individuals with and without RD in fMRI data. After running each model on fMRI data acquired during a word recognition task from individuals with and without RD, we examined which latent-mixture model(s) classified individuals (based on their fMRI data) in a way that matched their actual (known, behavioral) RD status. Statistically, we did so by examining for each model whether a produced classification parameter, estimated by the model for each participant, was related to their actual behavioral status (see methodological details below). Importantly, since the models were blind to the subjects' actual TD/RD classification, this comparison evaluated directly which model could successfully classify participants into RD and TD sub-groups based on neuroimaging data alone. And since each model was built to reflect one theoretical claim regarding neurofunctional markers of RD, we could infer from these results which claim(s) indeed have explanatory power in differentiating between RD and TD participants.

3. Candidate theoretical claims

The first step in our approach is to refine from the literature candidate theories that are then translated into generative classification models. In this first paper, we chose to focus on a few candidate theories – yet we emphasize that these are not meant to provide an exhaustive list of all theoretical claims in the literature. Rather, they are meant to represent a few common and/or plausible theoretical claims, both as a way of estimating these theories' explanatory power in the context of RD and more broadly to evaluate and exemplify our novel approach.

All the candidate theories below share the notion that there are neurofunctional differences between individuals with and without RD, which should be captured by hemodynamic changes as measured in fMRI, consistent with observations of robust structural abnormalities in RD readers (e.g., Eckert et al., 2016; Richlan et al., 2013). The crucial difference between these candidate theories is that each of them highlights a different *type of signal* as the marker that differentiates between the two groups. Thus, broadly speaking, the candidate theoretical claims we examined can be categorized into four sets, according to the neurofunctional contributors they emphasize as differentiating between RD and TD individuals: (1) claims regarding differences in mean activation; (2) claims regarding differences in heterogeneity (i.e., variability across subjects); (3) claims regarding differences in intra-subject (inter-region) variability; and (4) claims regarding differences in functional connectivity. Note that these theoretical claims are not mutually exclusive (i.e., more than one model may successfully identify RD individuals) – and indeed below we present analysis of the added value provided by multiple successful models. In the rest of this section, we briefly review these four categories and the relevant literature that motivated our selection. We stress from the outset that some of the claims we test originated from mechanistic accounts, whereas others are motivated by empirical findings or more intuitive explanations; in this paper, we are agnostic to the claims' origins, but focus instead on their operational predictions.

3.1. Differences in mean activation

The majority of work into neurofunctional markers of RD examines differences in group-level (or mean) activation between individuals with impaired vs. typical reading. Thus, a large number of studies documented functional differences between these two groups of individuals during reading as reflected in fMRI data (see, e.g., D'mello and Gabrieli, 2018 for review). From this wide breadth of data, converging evidence points to *decreased* activation for individuals with RD in a network of left-hemispheric regions that are considered canonical hubs of typical reading, including occipito-temporal, temporo-parietal, and inferior frontal regions (e.g., Hoeft et al., 2007; Paulesu et al., 2014; Pugh et al., 2001; Shaywitz et al., 1998), while also suggesting that individuals with RD may show *increased* activation in right-hemispheric regions (e.g., Hoeft et al., 2011; Simos et al., 2002; Waldie et al., 2013). There are also reports of increased activation in RD in non-canonical regions that are not part of the typical reading network (Richlan et al., 2009; Shaywitz et al., 1998). Thus, while RD/TD group differences are consistently observed, whether the difference has to do with higher or lower activation in RD may vary by hemisphere and region. Nonetheless, the most frequent findings are reduced signal in RD for canonical and left-hemispheric regions, along with increased signal for right-hemispheric regions, and we test these claims here.[3]

To do so, we represented theoretical claims regarding differences in group-level activation in a series of 'mean-activation' models. These models classify participants to groups of RDs vs. TDs based on network-wide differences in mean activation over a network of regions

[3]Note that while we had a directional prediction in mind, our models were built in a way that makes them sensitive to activation differences in either direction. See below for full specification of models.

(e.g., RDs show less activation compared to TDs in left-hemispheric 'canonical' regions of reading but more activation in non-canonical regions; RD show less activation in the left hemisphere but increased activation in the right; etc.).

3.2. Differences in inter-subject heterogeneity

A different potential marker we examined does not have to do with differences in mean activation, but rather with the extent of heterogeneity (i.e., inter-subject variability) in groups of RD and TD participants. The motivation for this model traces back to behavioral studies, showing that RD individuals are characterized – despite being concentrated in a limited part of the lower tail of the reading skill distribution – by substantial heterogeneity in various reading and non-reading tasks and also by comorbid conditions that could impact reading (e.g., Pennington and Bishop, 2009; Zoubrinetzky et al., 2014). This observation raises the possibility that individuals with RD also show increased heterogeneity in terms of patterns of brain activation (i.e., more variability around a central tendency compared to TD individuals), potentially reflecting RD's greater likelihood to utilize highly varied and/or idiosyncratic networks. In this regard then, even the absence of a mean difference in activation between TD and RD in a region could reflect a pattern of highly variable circuit building in RD. Given the intuitive appeal of this notion that RD may fail to show activation in nomothetic analyses given greater variation, and despite the fact that previous studies have not directly examined variance differences, we investigate this possibility here. In sum, a potential marker of RD is a more heterogeneous brain activation than in TD (i.e., greater inter-individual variability in RD vs. TD), which we examine via a model that classifies individuals into two groups that are similar in their activation means but differ in inter-subject variability (i.e., a "heterogeneous" group, which supposedly include the RD readers, and a "homogeneous" group, which supposedly consist of TD individuals).

3.3. Differences in within-subject variability

In addition to inter-subject variability as captured by the model above, we also examined a candidate theory according to which RD and TD individuals differ in terms of *intra*-subject variability (Hancock et al., 2017; Hornickel and Kraus, 2013; Malins et al., 2018). This was motivated in part by EEG research which found that individuals with RD differ from TD in instant-to-instant variance during assessment of complex auditory brainstem measures (Hornickel and Kraus, 2013; Neef et al., 2017), as well as in cortical activity (Centanni et al., 2018). In line with these findings, a recent paper suggested that this increased instance-to-instance variation may reflect a putative neural noise deficit in RD, stemming from abnormal balance of excitatory and inhibitory expression (Hancock et al., 2017). With regard to fMRI findings, a recent study found that good and poor readers differ in intra-subject variance, with TD readers showing increased trial-by-trial variability in the pars triangularis sub-section of the left inferior frontal gyrus, perhaps reflecting a more adaptive or flexible state (Malins et al., 2018). In short, there are intriguing findings that point to the importance of models rounded in variance but evidence is limited and may differ across imaging modalities. This focus on within-subject variability has become prominent in other clinical domains (Dinstein et al., 2015; Easson and McIntosh, 2019) and merited analysis in the current paper. Following this line of reasoning, we simulated a model according to which RD and TD individuals differ in the extent of variability *across regions* (i.e., intra-subject

inter-region variability). This is because both the potential utilization of non-reading regions and neural noise (among other factors) may result in increased inter-region variability in RD. This model thus classifies individuals into two groups, one comprising individuals who are characterized by increased variability across regions (which we predicted should include RD readers), and the other by lesser inter-region variance (the presumably TD group).

3.4. Differences in functional connectivity

A last category of models we considered reflects common claims that RD is associated not (only) with differences in activation – either in terms of mean activation or between/within-subject variability – but rather with differences in functional *connectivity* between regions. Thus, findings demonstrate differences between RD and TD groups in functional connectivity between various hubs of the reading network, most typically in the left hemisphere (see, e.g., Koyama et al., 2013 for review). These include several reports of disrupted connectivity in RD individuals in connections to/from the angular gyrus (Horwitz et al., 1998; Pugh et al., 2000), the inferior frontal gyrus (Richards and Berninger, 2008), and the occipitotemporal region (Koyama et al., 2013; Shaywitz et al., 2003; van der Mark et al., 2011; and see Finn et al., 2014 for a whole-brain analysis). Based on these findings, we examined whether functional connectivity differentiates between individuals with and without RD by building a model that classifies individuals into two groups of subjects that show either greater or lesser network-level between-region connectivity.

4. The current study

The structure of the remainder of the paper is as follows. In the Methods section, we describe our general design and procedure (participants, task, acquisition parameters, and preprocessing procedure). Then, in the Results section, we go over each candidate theory, describe the specification of a Bayesian latent-mixture model built to reflect it, and the classification results of this model. In reviewing the results, we focus on the relations between a classification parameter obtained from each of the models we fitted to the fMRI data and participants' behavioral known RD/TD status. This enables us to examine which of the model(s) - and consequently, the theoretical claims it reflects - indeed differentiates between RD and TD individuals. At the end of the Results section, we present data on the added predictive value of each of two different models that were characterized by successful classification, to examine whether each highlights unique markers of RD in relation to the other (or whether their classification overlaps), as well as the results of analyses examining the predictive value of models fitted to activation to non-print stimuli. Then, in the General Discussion, we discuss the implications of our findings both from the narrower perspective of theories regarding neurofunctional contributors to RD, and from a broader methodological perspective regarding the utility of our theory-driven classification approach in neuroimaging research of language disorders.

5. Methods

5.1. Participants and behavioral RD/TD classification

The results we present below are based on data from a community sample of n=127 adolescents and young adults (age range: 13.5-25.2 years, mean age = 19.9; 72 males and 55 females). All participants provided informed consent and the ethics protocol was approved by Yale University's Institutional Review Board. A subset of this sample (n =59) was also used in an earlier publication on memory consolidation focusing on a different fMRI task not reported here (Landi et al., 2018). All these participants passed fMRI quality controls (see below), and completed behavioral assessment using the two sub-tests of the TOWRE-II (Torgesen et al., 2012): Sight Word Efficiency (timed test of word reading) and Phonemic Decoding Efficiency (pseudoword reading). Behavioral (i.e. actual) group membership was defined using a conventional in-study criterion of <90 standard score in either word or pseudoword naming sub-tests (see e.g., Arrington et al., 2019; Siegelman et al., 2020). Per this criterion, our sample included 59 participants who were defined as RD and 68 defined as TD. Table 1 presents basic characteristics of the two groups (reading skills, age, gender, and mean in-scanner motion). fMRI *task*. Functional volumes were acquired while participants completed a task in which they processed visual and auditory stimuli. The task consisted of four conditions: i) printed real words; ii) spoken real words; iii) printed symbol strings; and iv) noise-vocoded spoken words. This design has been shown to be sensitive to individual differences in reading skills (Chyl et al., 2018; Malins et al., 2016). In each trial, subjects were presented with four stimuli in rapid succession in one of the four conditions. For the visual conditions, items within tetrads were present on the screen for 250 ms, with an ISI of 200 ms. Auditory items had a mean duration of 536 ms (SD = 110.2 ms) and were presented within tetrads with an SOA of 800 ms. At the beginning of the session, subjects were instructed to attend to the stimuli and told they would be given a short recognition memory test at the end of each run to motivate paying attention. Across all trials in the experiment, the time between trial onsets was jittered between 4 and 13 s. The task was performed in two runs, each lasting 5 minutes and 2 seconds. All conditions were presented in each run, with 48 trials per run in a pseudorandom order. No condition could repeat more than three times in a row. In total, this resulted in 24 trials (each trial being a tetrad of stimuli) per condition. Stimuli were presented using E-Prime.

5.2. Acquisition of MRI data

Images were acquired using a Siemens TIM-Trio 3T magnetic resonance imaging system (Siemens AG, Erlangen, Germany) with a 12-channel head coil. Prior to functional imaging, sagittal localizers were run (matrix size = 240 × 256; voxel size = 1 × 1 × 4 mm; FoV = 240/256 mm; TR = 20 ms; TE = 6.83 ms; flip angle = 25°). Next, anatomical scans were acquired for each participant in an axial-oblique orientation parallel to the intercommissural line (MPRAGE; matrix size = 176 × 256 × 256; voxel size = $1mm^3$; FoV = 256 mm; TR = 2530 ms; TE = 3.66 ms; flip angle = 7°). Following this, T2*-weighted images were collected in the same orientation as the anatomical volumes (32 slices; 4 mm slice thickness; no gap) using single-shot echo planar imaging (matrix size = 64 × 64; voxel size = 3.4375 × 3.4375 × 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.

Participants completed two runs in the functional task, which had a combined duration of 10 minutes and 4 seconds. fMRI *processing*. Data were preprocessed using AFNI (Cox, 1996; RRID: SCR_005927). Prior to running the *afni_proc.py* pipeline on the data of each subject, the *@SSwarper* program was run for brain extraction of the anatomical image and to apply the nonlinear warp of the anatomy to MNI space. Functional images were preprocessed by first correcting for slice acquisition time (*3dTshift*). Then, functional images were aligned with the anatomical images using the warps computed by the *@SSwarper* program (using the *tlrc_NL_warped_dsets* and *volreg_tlrc_warp* options). These 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 full width at half maximum of 8 mm (i.e., twice the between-plane distance of 4 mm; Skudlarski et al., 1999) and data were scaled (*3dcalc*) so that each voxel's time series had a mean of 100 for each run allowing the interpretation of EPI values as a percentage of the mean. During this scaling step, values in excess of 200 (meaning a > 100% signal increase) were clipped; this is the default value for scaling in AFNI and was selected to retain the precision of scaled short values.

Single trial estimates were obtained using a General Linear Model including nuisance regressors for the six motion parameters. This model was specified using the *stim_times* flag for *3dDeconvolve* in AFNI. The regression used a generalized least-squares time-series fit, with a restricted maximum likelihood estimation of the temporal auto-correlation structure (*3dREMLfit*). The hemodynamic response function was approximated using a gamma function. When performing the GLM, any volume that exceeded the thresholds of 0.3 mm Euclidean movement and/or if more than 10% of the voxels were flagged as outliers (using the *regress_censor_outliers* flag) and were censored from further analysis.

5.3. ROI selection for the canonical and non-canonical networks

As described below, our models used a categorization of ROIs into four different networks; left canonical; left non-canonical; right canonical and right non-canonical. Left canonical ROIs were based on a recent meta-analysis that reports peak activation coordinates for reading in adults (Martin et al., 2015). Left non-canonical areas were selected from a functional brain topography based on resting-state connectivity (Power et al., 2011). All left hemispheric coordinates that did not overlap with the canonical regions and were not within 2 mm of the central sulcus were included as left non-canonical ROIs. Note that this set of non-canonical regions include a large number of (cortical and sub-cortical) ROIs, many of them not specific to reading or language. The networks in the right hemisphere were mirror images of the left networks. As a result of this procedure, canonical networks consisted of 11 ROIs per hemisphere (see Table 2 for coordinates) and the non-canonical networks of 120 ROIs per hemisphere. We then created spheres with a radius of 3mm centered on these coordinates as ROIs, and obtained for each participant the mean beta across all trials in a given condition for each of these ROIs.

5.4. Input matrix (for mean-, heterogeneity-, and variability-based models)

The preprocessing procedure yielded a 127 × 262 matrix: Each value in this matrix was the mean estimated beta across trials in a given condition for each of the 127 subjects in each of the 262 ROIs (11 canonical + 120 non-canonical in each hemisphere). Our central analysis

below focuses on the *print* condition. In the input to these analyses, we subsetted the columns of this matrix to include only ROIs that showed significant group-level responses to print ($p < 0.01$). This was done to increase the signal-to-noise ratio in the input and improve model convergence (preliminary analyses revealed that many of the models failed to converge without such censoring). As a result, the print-activation input matrix included 140 ROIs: 10 canonical regions in each hemisphere, 57 left-hemispheric non-canonical, and 63 right-hemispheric non-canonical (see Figure in Supplementary Materials S2 for location of ROIs). All values in these remaining ROIs were scaled and centered within ROI (again to facilitate convergence, and to have a more interpretable scale on which to define prior distributions[4]). As a last step, we removed outliers with mean beta values farther than 3 SDs from the mean of each ROI. The resulting matrix served as the input to models 1-3 below (i.e., mean activation models, heterogeneity model, and intra-subject variability model).[5] Additional models were run on parallel matrices reflecting activation in two other task conditions - the false-font and the spoken-word condition; the goal of these additional analyses was to examine whether models' performance was specific to print processing or could be generalized to other types of materials (i.e., non-print visual stimuli and/or auditorily presented words). The procedure of the creation of these matrices was identical. However, note that the subsetting procedure was always based on condition-specific values, and therefore the matrices for different conditions included different ROIs (see more below).

5.5. Processing and input matrix for connectivity-based model

For the connectivity-based analysis, we had to create another input matrix, this time including connectivity values (rather than mean activation). There are multiple options for how to calculate functional connectivity, including resting state and task-based connectivity (e.g., psychophysiological interaction, Friston et al., 1997), and beta-series correlations (Rissman et al., 2004) methods, all with its own (dis)advantages. In this study we decided to use the full time-course during the task, for two main reasons: i) the number of trials (maximum of 24 per condition, and often less because of movement) was relatively limited for purely task-based connectivity analyses, and ii) recent work showing that cognitive tasks that are related to the skill of interest (here reading/language) amplify trait-relevant individual differences in functional connectivity patterns during the entire experiment (Greene et al., 2018).

To obtain connectivity metrics we examined the EPI values of the entire time-course of the experiment (i.e., across all conditions of the experiment, not just print trials), correlating each ROI with each other ROI, resulting in a 262 × 262 matrix per participant. Then, for each ROI we averaged the *r*-to-*z* transformed correlation with other ROIs that are in the same subnetwork resulting in a 127 (participants) by 262 (mean connectivity for each ROI

[4]Although the motivation for the scaling was methodological, we note that applying it may have led to over-weighing of ROIs that are less active on average. In other words, because input matrices included scaled values, less activated ROIs may have contributed more to the classification than they would when using classification based on raw values. We leave it to future work to explore these alternative procedures, while noting that using raw values requires careful consideration of models' convergence and use of proper prior distributions for different ROIs.

[5]Note that our input matrix included estimated activation for print overall (i.e., printed words vs. fixation contrast) rather than more specific contrasts (e.g., printed words vs. printed symbol strings; printed words vs. spoken words). We opted for this more general contrast in order to maximize the signal strength and the number of ROIs included in the analysis: More specific contrasts were associated with weaker signal overall and significant activation in a smaller number of regions.

with other ROIs in its sub-network) matrix. Note that to increase the comparability with the activation-based models fitted to print data (which becomes relevant when comparing the added classification value across these models, see below), in the analysis below we again subsetted the columns of this matrix to include the same 140 ROIs used in the print-models.[6] We again scaled values within columns and removed outliers farther than 3 SDs from each column's mean.

5.6. Processing and input matrix for connectivity-based model

In the latent-mixture modeling approach, two competing (i.e., latent) models are specified (Groups 1 and 2) and are pitted against each other, using a larger model that contains the two competing models and a binary classification parameter (z_i). This classification parameter examines, for each individual, whether their data are more likely under the specification of Group 1 or under Group 2.[7] Operationally, the model is structured in a way that in each Monte Carlo Markov Chain (MCMC) iteration z_i can be either 0 or 1: z_i = 0 reflects classification to Group 1 (i.e., data more likely under sub-model 1), and z_i = 1 reflects classification to Group 2. Then, similarly to all Bayesian models, the distribution of the classification parameter z_i across iterations can be taken as a proxy of the posterior distribution of this parameter (Kruschke, 2014; Lee and Wagenmakers, 2013). Specifically, in the latent-mixture case, the posterior distribution reflects the certainty in classifying subject *i* as following sub-model 2 (compared to sub-model 1): The mean of z_i across iterations is the model's certainty in classifying subject *i* as a member of Group 2 (e.g., $\bar{z}_i = 1$ reflects full certainty in classification as Group 2; $\bar{z}_i = 0$ reflects full certainty in classification as Group 1). Note that latent-mixture models allow for individual differences or mixture in the data (hence their name): that is, a situation where some individuals in the sample are classified as members of Group 1, whereas others of Group 2.

We re-iterate that in all analyses presented below, models were only fitted to fMRI data; only after the z_i parameter was estimated for each subject (under each model) we compared it to the actual group membership of participants. This means that in contrast to common approaches (e.g., Machine Learning), our approach does not require cross-validation. In other words, in contrast to typical approaches where models are trained on both predictors (in the current case, fMRI data) and outcome (RD/TD status), in our approach parameters are estimated only on the predictors, which eliminates the risk of overfitting and hence obviates the need for cross-validation.

In the Results section, we describe the performance of a series of Latent-mixture models each built to reflect one of the theories laid out in the Introduction. For readability, we describe each model's specification along with its classification performance. Note that

[6] In this connectivity matrix values represented the mean connectivity between each of the 140 ROIs and all other regions in the same sub-network, regardless of whether they were included in the subset of 140 ROIs. A slightly different approach is to confine connectivity estimates only to the subset of 140 ROIs (i.e., calculate the connectivity between each of the 140 ROIs and all other regions in the same sub-network that were also part of the 140-ROI subset). Running the models described below on this modified input matrix resulted in qualitatively similar results.

[7] The latent-mixture modeling approach requires a binary criterion. In all our analysis below, we therefore use a binary split of participants into RD and TD sub-groups. Our choice here is driven by methodological considerations and should not be taken to reflect a theoretical claim regarding whether individuals with RD constitute a qualitatively distinct subgroup or simply the lower end of the reading skill continuum.

the different models were built to maximize their similarity to one another, with minimal changes implemented to reflect the critical difference(s) regarding the source of RD/TD classification (i.e., all models share many assumptions, except for those related to the type of signal they view as the source of RD/TD classification). This was done to ensure that all models have a similar potential to pick up on meaningful individual differences.

5.7. Specification and estimation of Bayesian models

MCMC samples were run using JAGS (Depaoli et al., 2016), version 43.0, and the *rjags* package in R (Plummer, 2016), version 4-10. In all estimations we used three separate MCMC chains with random starting points. Each chain included 2000 iterations after 5000 burn-in iterations; the goal of the burn-in iterations was to ensure that samples were taken from the posterior distribution only after the MCMC procedure was sufficiently stable. To check whether the 3 chains converged to similar posterior distributions we used the Gelman-Rubin diagnostic measure (Gelman and Rubin, 1992). Lower values reflect high agreement across chains, with values under 1.1 generally interpreted as good model convergence. Convergence estimates for the models below were generally under this threshold, both for group-level parameters and for the majority of subject-level classification parameters (see Supplementary Materials S3 for full information). Full codes with th6e specification of Bayesian models, the data fed to the models (i.e., input matrices), and already-fitted posterior distributions (which were used in the results below) are available via the project's OSF page, at: https://osf.io/2vrwa/.

6. Results

6.1. Analysis of basic print activation

Before turning to the main analysis (using the Bayesian latent-mixture models), we first examined the activation of the print vs. fixation cross contrast in our sample (across all subjects). Fig. 1 shows the voxel-by-voxel activation map for this contrast, at a threshold of $p < .001$. As can be seen, we observed strong bilateral print-related activation across reading-related areas including the fusiform gyri, superior temporal, inferior parietal and frontal gyri, extending into sub-cortical structures such as the thalamus and putamen. This is expected for this print contrast and in line with previous findings with this task (Chyl et al., 2018; Malins et al., 2016). In addition, the mean beta for ROIs in the left ($\beta = 0.135$) and right ($\beta = 0.128$) canonical reading networks were stronger than in the left ($\beta = 0.040$) and right ($\beta = 0.041$) non-canonical networks. These results corroborated that we were adequately measuring activation for print and were distinguishing well between canonical and non-canonical regions.

6.2. Bayesian latent-mixture models: specification of models and classification performance

As mentioned above, our central focus in this paper is the evaluation of a series of latent-mixture classification models built in light of four sets of theories regarding the classification of RD vs. TD individuals from fMRI data. In this section we review, for each model, the major assumptions that went into it – in the main text we provide a more intuitive explanation, yet detailed descriptions are available in the Supplementary Materials S4. Then,

we report each model's classification performance, focusing on whether the classification produced by the model matched individuals' actual (i.e., behavioral) group membership.

6.2.1. Mean-activation models

Model specification.: Here we describe a series of models that classify individuals into two groups based on differences in mean activation (see Fig. 2, Panel A, for the general architecture). Building upon observations of mean hemispheric differences between TD and RD individuals (see Introduction), the first mean-activation model we specified was a model we refer to as *Model 1A: left-right model*. The priors of this model are shown in Fig. 2 panel B.

Recall that the *left-right model* classifies individuals into two groups – Group 1 has a greater mean activation in the left hemisphere compared to the Group 2 (in both canonical and non-canonical regions), and Group 2 has a greater mean activation in the right hemisphere than Group 1 (again, in both canonical and non-canonical sub-networks). This is reflected in the constraints on the models' parameters that express the population means in each of the four sub-networks – see Fig. 2 and its caption, and details in Supplementary Materials S4. Importantly, the latent-mixture model estimates, for each individual, whether they belong to Group 1 or Group 2. This is done based on the crucial classification parameter z_i, a dichotomous parameter that in each iteration of the model can be either 0 or 1. If z_i=0, then the subject's mean activation parameters are taken from a distribution under Group 1; while if z_i=1 they follow a distribution under Group 2. The mean of z_i across iterations thus reflects the model's certainty in classifying the subject *i* to Group 2 compared to Group 1.

In addition to the left-right mean-activation model, we defined two other mean-activation models to further explore classification based on network-level mean activation differences. One was a *left-only model (model 1B);* Under this model, Group 1 has a greater mean activation than Group 2 in the left hemisphere (in both canonical and non-canonical networks; same as in model 1A above), but there is no difference between the two groups in mean activation in the two sub-networks of the right hemisphere. The other mean-activation model is a *left-canonical-only model (model 1C)*, according to which the difference between the two groups is specific to the left-canonical sub-network (with greater activation in Group 1 than 2), while the population means of the two groups are equal in the remaining three sub-networks. See Supplementary Materials S4 for formal specification.

Classification Results.: The results of *model 1A: Left-right model* are presented in Fig. 3. Panel A presents the histogram of the mean group classification parameter (z_i) across subjects (as well as the estimated population-level means in the two Groups); Panel B presents the mean classification parameter across subjects in the TD and RD groups (based on their behavioral status); Panel C presents individual-level distribution of the group classification parameter along with their actual group membership (i.e., TD/RD status); and Panel D presents the dichotomous classification performance of the model. The later was examined by categorizing participants according to whether their mean z_i was above or below the median group classification parameter of the sample, and cross-tabulating this information with participants– behavioral group membership.

As can be seen in Panel 3A, the model produced a bimodal distribution of mean z_i parameters across participants, suggesting that indeed it discovered two groups of subjects in the data based on differences in mean activation (i.e., two groups of participants such that participants belonging to Group 1 had greater mean activation in the left hemisphere than those in Group 2, and participants in Group 2 had greater mean activation in the right hemisphere than those in Group 1). The estimated population means presented in Panel 3A suggests however that the distinction between the two groups was driven mostly by differences in right-hemispheric activation (i.e., the population means of the two groups in the left hemisphere were estimated to be almost identical). Importantly, Panel 3B shows that the classification of this model did not match the actual RD/TD group membership of the subjects in our sample. Thus, the two groups only minimally differed in their mean classification parameter across participants (TD: $\bar{z}_i = 0.47$, SD = 0.45; RD: $\bar{z}_i = 0.46$, SD = 0.45), a difference which was not statistically significant (Mann-Whitney-U(125)=1992, $p = 0.95$;[8] see also Panel 3C for individual-level data). In the same vein, the dichotomous classification performance was not significantly different than a chance-level of 50% (64/127 (50.3%) participants classified correctly, $p = \sim 1$; see Panel 3D).

We also fitted to our data the two additional mean-based activation models: the *left-only model (1B)* and *the left-canonical-only model (1C)*. For brevity, we do not report the results of these models here – the full results are presented in the Supplementary Materials S5. In a nutshell, these models again did not produce a classification that matched the actual group membership of the participants in our sample.

6.2.2. Heterogeneity-based model

Model specification.: The next model we specified was a *heterogeneity-based model (Model 2)*. This model classifies individuals into two groups not based on differences in mean activation, but instead based on inter-subject variability. Thus, it aims to discover two sub-groups of individuals that differ in their heterogeneity (but not in their mean activation). Intuitively, since the model assumes that there are two latent populations with identical means but differences in inter-subject variability, individuals with more extreme values (i.e., further away from the population means in the different sub-networks) are more likely to be classified into the more heterogeneous group (and vice-versa for subjects closer to the sub-groups' means). Formally, the architecture of this model is similar to that of the mean-activation models (Fig. 2A), but differs in the a priori constraints it applies on the unobserved parameters. Thus, the heterogeneity-based model assumes that the two groups differ in their inter-subject variability (i.e., between-subject standard deviations): Group 1 is defined as the more heterogeneous group, and Group 2 as the more homogeneous group (see Supplementary Materials S4 for details). Note that under this model, the two groups have identical population means in all four sub-networks.

Classification performance.: The results of the *heterogeneity-based model* are presented in Fig. 4. Panel 4A shows that the model identified two groups of subjects differing in

[8]We used here and below the non-parametric Mann-Whitney test rather than a t-test because the classification parameter (z_i) was not distributed normally across participants. The results were qualitatively similar when using a two-sample t-test (all significant tests remained significant; all insignificant tests remained insignificant).

their inter-subject variability: Group 1 being the more heterogonous group and Group 2 being the more homogenous group. Panel 4B shows, however, that there was no evidence for a difference in the group classification parameter between RD and TD individuals in our sample (TD: $\bar{z}_i = 0.47$, SD = 0.40; RD: $\bar{z}_i = 0.49$, SD = 0.38; Mann-Whitney-U(125) = 2048, $p = .84$). The dichotomous classification success of this model was similarly around chance-level (successful classification of 62/127 participants, or 48.9%; Panel 4D).

6.2.3. Variability-based model

Model specification.: This model classifies subjects into two groups based on their extent of (intra-subject) inter-region variability. It uses a modified architecture and priors shown in Fig. 5. Most of the model's specification is similar to that of the models above. It again uses higher-order parameters to estimate population means and (inter-subject) standard deviations, which in this case are equal in the two groups (i.e., the two groups have the same mean activation and inter-subject heterogeneity). Crucially, this model differs from the previous models in how it estimates the parameters reflecting *intra-subject inter-region* variability. Thus, the estimation of these parameters depends on each subject's group membership (using the same latent-mixture strategy as the models above): With subjects showing more variability across ROIs (in all four sub-networks) classified into Group 1 and those showing less variability classified into Group 2 (please refer to Supplementary Materials S4 for details).

Classification performance.: Similarly to previous models, the variability model was able to identify two groups of subjects with generally high certainty (Fig. 6, Panel A). Critically – and in contrast to other models reviewed so far - this assignment (based on fMRI data) was related to participants' actual (i.e., behavioral) group membership. Thus, there was a significant difference between the RD and TD groups (as defined by their behavioral reading performance) in the mean classification parameter: TD: $\bar{z}_i = 0.56$, SD = 0.43 RD: $\bar{z}_i = 0.38$, SD = 0.41; Mann-Whitney-U(125)=1591, $p = 0.04$ (Panel 6B). Note that RD participants were more likely to be classified to the more variable group, in line with the underlying theory's predictions (see also Panel 6C for individual-level distribution). The dichotomous classification performance of the model was also significantly greater than chance, with a successful classification of 78/127=61.4% participants ($p = 0.01$; see Fig. 6 Panel D).[9]

One possible concern is that the differences in intra-subject variability between the two groups were related to other confounding factors, not inherent to the individuals' RD/TD status. Such factors include the extent of motion during the scan session, or differences in gender and/or age distribution between the two groups (age in particular is a potential confound given the significant age difference between the TD and RD groups in our sample, see Table 1 above). To rule out this possibility, we ran a logistic multiple regression model with behavioral group classification as the dependent variable, with mean

[9]Given the successful classification of the variability-based model, we went back to the raw data and calculated the mean intra-subject inter-region variability in the two groups. In line with the results of the model, RD participants had greater variability than TD participants, consistently observed in the four sub-networks. At the same time, these numerical differences were not deemed as significance in standard analyses, which suggests that the Bayesian models we used are more sensitive than these techniques. See Supplementary Materials S6 for details.

z_i, age, gender, and percent TRs censored due to motion (proxy of extent of motion) as predictors. The results of this analysis, presented in Table 3, revealed that the link between the estimated group classification parameter and participants' actual group membership remained significant ($p = 0.008$) also when controlling for these possible confounds.

6.2.4. Connectivity-based model

Model specification.: The architecture of the connectivity model is identical to that of the mean-activation (and heterogeneity) models above (Fig. 2). The crucial difference is in the input to those models – which uses connectivity values (a matrix with mean r-to-z transformed correlations between each region and all other regions in the same sub-network; see Methods) instead of activation values. In terms of priors, the connectivity-based model employs constraints on population connectivity means, such that the mean connectivity of Group 1 is constrained to be higher than that of Group 2 (in the same direction in all four sub-networks). All other parameters are similar to those in the mean-based activation models (only now they represent connectivity values rather than activation). Overall, then, the connectivity-based model classifies individuals into two groups, one that has greater mean connectivity (across sub-networks) compared to the other.

Classification performance.: The results of the connectivity-based model are presented in Fig. 7. Similar to the models above, this model produced a bimodal distribution of the group classification parameter (Panel 7A). Importantly, the model's classification was associated with individuals' actual group membership: RD individuals had a higher mean classification parameter on average compared to TD individuals (TD: $\bar{z}_i = 0.44$, SD = 0.46; RD: $\bar{z}_i = 0.63$, SD = 0.47; Mann-Whitney-U(125)=2489.5, $p = .01$; Panel B). This means that RD individuals were more likely to be classified as belonging to the group showing *less* inter-region connectivity (Group 2), in line with the prediction of the model's underlying theory (see also Panel 7C). The dichotomous classification performance of this model was also above-chance, with 77/127 (60.6%) individuals classified correctly ($p = .02$; Panel D). Note that again, the group classification parameter predicted behavioral group membership also when controlling for age, gender, and motion (Table 4).

6.3. Added classification value of the two successful models

So far, our results show that two models – the intra-subject variability-based model and the connectivity-based model – produced successful classification reflecting participants' actual group membership. A follow-up question is whether the two models have any added value on top of the other. Specifically, one may posit that the extent of connectivity and inter-region variability are related, as greater inter-region variability in activation for printed words may reflect poorer connectivity between regions (computed across all trials). We therefore ran additional analyses to assess the relation between the group classification parameters produced by the two models, and examine whether each of the two models predicted individuals' actual RD/TD status beyond the information produced by the other.

First, we estimated the correlation between the group classification parameter estimates under the two models, revealing that they were weakly and not significantly associated ($r = .03$, $p = .70$). This already suggests that the two classification models utilized non-

overlapping signal. Next, we ran a logistic model, where we predicted the behavioral group membership (TD/RD) from the group classification (z_i) parameters produced by both models, while also including motion, gender, and age as controls. The results, presented in Table 5, showed that the group classification estimates produced by each of the models still significantly predicted individuals' behavioral group membership above and beyond the other, again pointing to the utilization of non-overlapping signal, which is differentially captured by measures of variability and connectivity. That is, individuals with greater z_i value per the variability model (classified as the *less variable group*), and smaller z_i per the connectivity model (classified as the group with the *greater connectivity*), were more likely to have an actual TD (rather than RD) diagnosis. Lastly, we ran a classification analysis based on the dichotomous classification produced by the two models, where for each individual we examined whether they were classified into the presumably RD group according to both models, only according to the variability-based model, only according to the connectivity-based model, or in neither case (i.e., classified as TD in both). The results are presented in Table 6, showing that the actual group membership for individuals who were classified as either TD or RD under *both* models was highly likely to fit with the models' classification: That is, the conjunction of the two models' classification had a success rate of 71.9% (46/64) in participants for whom there was an agreement between the two. The fact that this classification rate was higher than in each of the models alone (see above), again suggests that the two successful models tap into non-overlapping parts of the signal, and thus that each theoretical claim carries unique explanatory power. We return to this point in the General Discussion below.

6.4. Classification based on non-print activation

In analyses reported so far, models fitted to activation matrices (i.e., mean activation, heterogeneity, and variability-based models) used input matrices reflecting activation to printed words. We next examined whether the results of these models are specific to activation to print, or whether they generalize to responses to other types of materials used in the task. We did so by fitting the same models to two additional input matrices, each including beta values reflecting activation to one of two conditions: False-font visual stimuli, and spoken words. The full results are reported in Supplementary Materials S7. In a nutshell, the results of the false font condition were essentially identical to that from the print-based models above: That is, at-chance classification based on mean-activation and heterogeneity, along with successful (i.e., above-chance) classification based on inter-region variability. Moreover, the estimated z_i parameters from the print and false-font models correlated strongly ($r = 0.51$, $p < .001$), and their predictive value of RD/TD over-lapped substantially (i.e., no significant independent predictive value of either when both are included in the same logistic model). This suggests that inter-region variability in false-font activation patterns is as informative regarding TD/RD status as variability for print. In contrast, we found that classification based on speech matrices was at chance for all models (including the variability model). We return to discuss the implications of these findings below.

7. General discussion

The goal of the current paper is to propose a novel framework for evaluating different theoretical claims regarding the neurobiological markers of a given behavioral trait. To this aim, we adopt the Bayesian latent-mixture modeling method, which was successfully implemented in previous behavioral work in multiple domains (see, e.g., Siegelman et al., 2019; Steingroever et al., 2019 for recent applications), and apply it for the first time to neuroimaging data. The strength of this approach, we argue, is that it is geared specifically for the problem of evaluating competing theories regarding classification of individuals, as it combines features from different common techniques of neuroimaging data analysis that are particularly important for this purpose. Thus, on the one hand, our approach is theory-driven – much like often-used univariate statistical methods – which enables us to use our models' classification performance as a proxy for the explanatory power of the theories that each of them reflects. On the other hand, our approach shares some features with data-driven approaches (such as Machine Learning algorithms), particularly in how it identifies markers at the network-level. In a sense, then, our latent-mixture modeling approach complements existing neuroimaging data analysis procedures, that either test singular theories that are often region-specific, or detect network-level patterns without regard to whether or not they are theoretically transparent.

In this first paper, we applied the latent-mixture approach to fMRI data to unveil the neurofunctional markers of impaired reading. We compared four classes of existing and plausible theoretical claims regarding classification of RD and TD individuals from fMRI data, each underlining a different type of signal as the one that differentiates between these two groups of individuals. Namely, we tested theories that stress differences in mean activation, inter-individual variance (i.e., heterogeneity), intra-individual (inter-region) variance, and functional connectivity. To re-iterate, all models were fitted to fMRI data alone, and therefore they could classify individuals in concordance with their actual behavioral status only if the theories they were built to reflect capture relevant neurobiological markers. We found that the models built to reflect theories regarding differences in intra-individual variability and functional connectivity produced classifications that were significantly associated with participants' actual (i.e., behavioral) RD/TD status (while models positing global mean activation and heterogeneity differences failed to do so). This result suggests that key neural correlates of reading (dis)abilities are found beyond differences in mean activation. As such, it strengthens previous systems-level oriented reports of associations of reading skills with metrics of functional connectivity (e.g., Finn et al., 2014; Pugh et al., 2000; Wang et al., 2013), and the emergent literature on the role of neural variability in reading development (Hancock et al., 2017; Hornickel & Kraus, 2013; Malins et al., 2018).

At the same time, it is notable that despite the significant associations between the variability- and connectivity-based models' estimated classifications and individuals' behavioral status, their overall performance was somewhat limited. Concretely, our models' classification performance was notably lower than that reported in papers using data-driven methods (compare for example the binary classification success rate of around 61% in both our successful models, to performance of about 80% in work using Multivoxel Pattern

Analysis by Tanaka et al., 2011). We note that the disparity between data-driven models and our approach is expected: As mentioned in the Introduction, our models do not aim to maximize classification performance but rather theoretical transparency, whereas data-driven approaches can pick up on any part of the input signal that contributes to successful classification (i.e., their classification rate is not constrained by a theory). Importantly, the higher performance achieved by the data-driven approaches compared to our theory-driven models suggest that there are in fact other informative parts in the fMRI signal that are associated with TD/RD status, which are not captured by current theories (or at least not by the 'global' models examined here). In general, the precision of available theories presents an upper-bound for the classification success of our latent-mixture models: A theory-driven model can only be as precise as the theory it reflects. As theories become more precise and complete, we expect the classification rate of theory-driven models to increase as well, eventually reaching values similar to those obtained by data-driven methods.

Importantly, we stress that even our "theory-constrained" classification rates became higher when information from *both* types of successful models were combined. Thus, group classification parameters produced by each of the models predicted behavioral group membership beyond that produced by the other, and binary classification rate based on the conjunction of the two models was high among individuals who were classified into the same group by both models. This result suggests that the two types of signals used for classification by the two successful models tap into non-overlapping information: That is, that differences in inter-ROI variability cannot be reduced to differences in (the more frequently studied measure of) functional connectivity, or vice versa. More broadly, this result underlines the importance of the development of mechanistic accounts that consider multiple types of signals, explain how they are related to each other, and how they eventually lead to a behavioral deficit (as an example, consider models tying suboptimal balance of neurometabolites to increased neural noise, which then contributes to RD's auditory processing deficits and inefficient print-speech binding; Del Tufo et al., 2018; Hancock et al., 2017; Pugh et al., 2014).

Another intriguing finding worth emphasizing is the similarity in classification performance between models fitted to different types of visual stimuli (and the dissimilarity of these conditions from that of the speech condition). Thus, regardless of whether models were fitted to input matrices with activation to printed words or to false-font stimuli, they resulted in very similar performance, with successful classification based on inter-region variability in activation to the two types of stimuli. At face value, this finding may be taken to suggest that whatever deficit increased inter-region variability reflects, it generalizes beyond printed words and similarly applies to the processing of other visual stimuli, in line with visual deficit theories of reading disabilities (Eden et al., 1996; Lobier et al., 2012). Yet given the similarity between false font and printed words, it is entirely possible that the processing of false font stimuli is the consequence of the organization of the reading system given an individual's exposure to print, rather than reflecting differential processing of visual materials more generally. In other words, processing of false font may reflect how a well-established print system "attempts" to code such stimuli, leading to the strong overlap and classification similarity between the print and false font conditions. This interpretation is also consistent with the high proportion of regions involved in processing false font

stimuli that were also activated when reading printed words (out of the 106 showing group-level activation at the false font condition, 94 also had significant activation for print; see Supplementary Materials S7). With the current data, we cannot adjudicate between these two accounts; future research can do so by comparing classification performance based on variability in activation across other visual conditions that carry less resemblance to printed words.

In addition to the implications of these positive findings, we wish to clarify what can – and cannot – be concluded from the *un*successful classification produced by some of the models we tested (i.e., the null findings produced by the mean activation models: Both the left-right model reported above, and related models reported in the Supplementary Materials S5). The conclusion that can be drawn from these null findings is that global mean-activation differences are not sensitive enough to distinguish between individuals with and without RD, at least not in the current sample and design. We note that there may be, however, developmental changes in TD/RD differences in mean activation. In fact, work by Shaywitz and colleagues (2002, 2007) showed that while TD individuals show relatively stable patterns of activation over age (see also Church et al., 2008), individuals with RD exhibit a substantial increase in activation over development in large parts of the brain. It is therefore possible that although the RD and TD individuals in our sample (who already had years of exposure to print) did not show global differences in mean activation, such differences may be more diagnostic among younger populations (and see Maurer et al., 2011 for longitudinal evidence). Furthermore, the task used may modulate the informativeness of such mean-activation differences. In fact, it was shown that activation changes differently over trials within a task in RD and TD; With repetition TD readers reduce BOLD signal while RD increase it producing a crossover interaction in the same regions (Pugh et al., 2008), implying that static group contrasts of activation are context sensitive. It is crucial that future studies further map the factors that determine when and to what extent mean differences are diagnostic of RD/TD status, which can be done by applying our method to data from different tasks and developmental stages.

In addition to these important factors, it is possible that RD and TD individuals do differ in activation in one or in some small set of ROIs (consistent with meta-analytic findings; Paulesu et al., 2014; Richlan et al., 2009). Our current specification of models – which searches for global differences in mean activation over networks of regions – cannot capture such region-specific differences. That being said, the latent-mixture models we use can be adapted to reflect theories that focus on one or on a specific set of ROIs (for instance, claims regarding differences in activation between TD and RD individuals in the Visual Word Form Area, Dehaene and Cohen, 2011; van der Mark et al., 2009). Such adaptations can be made not only in mean activation models, but also in models that did already result in above-chance classification (based on connectivity and intra-subject variability): Our specification of models was only meant to serve as a coarse-grained representation of current theories, testing for overall differences in connectivity/variability across the brain, not to assess more spatially-specific claims (e.g., the importance of connectivity to and from the occipitotemporal region; Koyama et al., 2013; Shaywitz et al., 2003; van der Mark et al., 2011). We leave it for future work to examine whether models positing region-

specific differences –in activation and other types of markers - do in fact result in improved classification.

Throughout this work, we promoted the use of the latent-mixture approach in the analysis of neuroimaging data. Indeed, we are hopeful that the tool we offer will contribute to advances in identifying and refining theories in the field of reading and language disorders, as well as other cognitive deficits. But our embrace of this approach should not be taken as a criticism of other methods. As mentioned above, we see our approach as one that is meant to complement – not override – other techniques. Thus, standard univariate approaches are still patently useful when it comes to examining one hypothesis in limited sample sizes, especially when it comes to region-specific predictions (although global differences can be examined via extensions to univariate mixed-effect models, see Chen et al., 2019). Standard approaches also offer the option of relating brain signal to behavior continuously, in contrast to our approach which at least in the current specification requires a dichotomous behavioral outcome. In parallel, data-driven methods are valuable because they can help estimate the upper-bound classification levels one can expect from theory-driven methods, and because they are key in exploratory research whose goal is to identify *novel* candidate theories. Once such candidate theories are identified, they should be incorporated and tested using theory-driven tools, such as the one we present here, which are built to evaluate existing theories.

Undeniably, the Bayesian modeling approach (and any other type of generative modeling) requires researchers to be explicit regarding how a series of latent parameters gave rise to the observed data. We see this as an advantage of our approach – using this method, we hope, will encourage researchers to be confronted with their (often implicit) assumptions. At the same time, methods that require explicit specification always incorporate a series of non-trivial assumptions, and the models we used here present no exception to this rule. In the current case these include, for example, assumptions regarding the distribution of beta values over subjects and ROIs (assumed to be normal within populations of subjects/networks) and about the assignment of ROIs into sub-networks (with ROIs labeled as belonging to four sub-networks of canonical and non-canonical regions in the two hemispheres). It is inevitable that the accounts of other researchers will vary to some extent from the one reflected by the assumptions we incorporated in our models. Importantly, the Bayesian framework provides a clear way of incorporating and testing different assumptions in a formal manner: All assumptions in our models are explicitly stated and can be easily changed; once such modifications are made, a model's output can be re-examined to check how the change in assumptions impacts a model's classification performance (i.e., whether the success/failure of a model is contingent on specific assumptions). In this first paper, we did not attempt to cover different possible models' architectures, but our approach provides a straightforward framework for proposing alternative models and revisit our assumptions. We are certain that our models are bound to become more precise with increasingly more sophisticated formulations of brain organization proposed by the research community.

We end by returning to the starting point of this paper ' the question of what are the neurofunctional markers of impaired reading. Our work, we believe, already shows promise for advancing theory-grounded research into this question – highlighting the types of

markers that indeed differentiate between RD and TD individuals at the network-level. At the same time, we wish to underline some open questions that should be examined by future research. The first question has to do with development. We already discussed above how developmental changes may have contributed to the (null) findings in the mean activation models. Similarly, questions remain about how the contributions of connectivity and variability to TD/RD differences change over age and experience. Much like in the case of mean activation differences, it is crucial that future research examine whether these differences are present already from a young age (as suggested, for example, by studies showing prospective correlations of connectivity before literacy onset and later reading skills, Jasińska et al., 2020), or whether the profiles we see in adults and adolescents is the result of a differential growth in these metrics over age and/or exposure to print in TD and RD populations (Morken et al., 2017). A second open question has to do with the multi-dimensional nature of RD. In the current (and first) specification of the models, classification was based on variation along a single axis. However, key accounts of RD suggest that there may be multiple deficits contributing to reading disorders, where multiple risk factors accumulate until the threshold of categorical diagnosis is met (Pennington, 2006; Snowling and Hulme, 2020). Such a multiple deficit can be incorporated (and tested) in modified generative models, where classification is informed by multiple dimensions, reflected in different types of neural signatures (potentially, in different regions) that all contribute to the categorical TD/RD parameter. Note that a successful classification based on multiple signals may be present even in the absence of increased inter-subject variability along a single axis, which was directly tested in our heterogeneity model (and in fact, the added predictive value of the connectivity and variability models, even in the absence of increased network-level heterogeneity in activation, is consistent with this notion). Lastly, another open question has to do with imaging modality: Whereas we only focused on functional MR data, neural correlates of reading skills are well-documented in other imaging modalities, including in various measures of neuroanatomy (e.g., Tamboer et al., 2016; Wai et al., 2008; but see Ramus et al., 2018 for a more critical review), and in other neurofunctional techniques (including EEG, e.g., Ackerman et al., 1994; Maurer et al., 2007; Sklar et al., 1972; and more recently fNIRS, Jasińska et al., 2020). Future work should therefore adapt the current models to accommodate different types of inputs – with the eventual goal of examining theories spanning brain structure and function and being informed by different types of data. Going forward, the computational framework used here can serve as a foundation for these and other extensions, providing researchers with a tool for evaluating different theoretical accounts in explicit, quantifiable terms. This approach should therefore prove valuable in advancing fleshed out accounts of reading difficulties, as well as of other language- and cognitive impairments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under award P20HD091013, R01HD086168, and R37HD090153. The data included was collected under support by award R01HD065794. The content is solely the responsibility of the

authors and does not necessarily represent the official views of the National Institutes of Health. N.S. received funding from the Israeli Science Foundation (ISF), grant number 48/20.

Data and code availability statement

Please refer to the project's OSF page for raw data and code: https://osf.io/2vrwa/.

References

Ackerman PT, Dykman RA, Oglesby DM, Newton JE, 1994. EEG power spectra of children with dyslexia, slow learners, and normally reading children with ADD during verbal processing. J. Learn. Disabil 27 (10), 619–630. doi:10.1177/002221949402701002. [PubMed: 7844478]

Arrington CN, Malins JG, Winter R, Mencl WE, Pugh KR, Morris R, 2019. Examining individual differences in reading and attentional control networks utilizing an oddball fMRI task. Dev. Cognit. Neurosc 38, 100674. doi:10.1016/j.dcn.2019.100674.

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA, Avesani P, Baczkowski BM, Bajracharya A, Bakst L, Ball S, Barilari M, Bault N, Beaton D, Beitner J, … Schonberg T, 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582 (7810), 84–88. doi:10.1038/s41586-020-2314-9. [PubMed: 32483374]

Bruck M, 1992. Persistence of dyslexics' phonological awareness deficits. Dev. Psychol 28 (5), 874–886. doi:10.1037/0012-1649.28.5.874.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR, 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci 14 (5), 365–376. doi:10.1038/nrn3475. [PubMed: 23571845]

Centanni TM, Pantazis D, Truong DT, Gruen JR, Gabrieli JDE, Hogan TP, 2018. Increased variability of stimulus-driven cortical responses is associated with genetic variability in children with and without dyslexia. Dev. Cognit. Neurosci 34, 7–17. doi:10.1016/j.dcn.2018.05.008. [PubMed: 29894888]

Chen G, Xiao Y, Taylor PA, Rajendra JK, Riggins T, Geng F, Redcay E, Cox RW, 2019. Handling multiplicity in neuroimaging through bayesian lenses with multilevel modeling. Neuroinformatics 17 (4), 515–545. doi:10.1007/s12021-018-9409-6. [PubMed: 30649677]

Church JA, Coalson RS, Lugar HM, Petersen SE, Schlaggar BL, 2008. A developmental fMRI study of reading and repetition reveals changes in phonological and visual mechanisms over age. Cereb. Cortex 18 (9), 2054–2065. doi:10.1093/cercor/bhm228. [PubMed: 18245043]

Chyl K, Kossowski B, Dębska A, Łuniewska M, Banaszkiewicz A, Żelechowska A, Frost SJ, Mencl WE, Wypych M, Marchewka A, Pugh KR, Jednoróg K, 2018. Prereader to beginning reader: changes induced by reading acquisition in print and speech brain networks. J. Child Psychol. Psychiatry 59 (1), 76–87. doi:10.1111/jcpp.12774. [PubMed: 28691732]

Colon EJ, Notermans SLH, de Weerd JPC, Kap J, 1979. The discriminating role of EEG power spectra in dyslexic children. J. Neurol 221 (4), 257–262. doi:10.1007/BF00314642. [PubMed: 92551]

Cox RW, 1996. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res 29 (3), 162–173. doi:10.1006/cbmr.1996.0014. [PubMed: 8812068]

D'mello AM, Gabrieli JDE, 2018. Cognitive neuroscience of dyslexia. Lang. Speech Hear. Serv. Sch 49 (4), 798–809. doi:10.1044/2018_LSHSS-DYSLC-18-0020. [PubMed: 30458541]

Dehaene S, Cohen L, 2011. The unique role of the visual word form area in reading. Trends Cogn. Sci 15 (6), 254–262. doi:10.1016/j.tics.2011.04.003. [PubMed: 21592844]

Del Tufo SN, Frost SJ, Hoeft F, Cutting LE, Molfese PJ, Mason GF, Rothman DL, Fulbright RK, Pugh KR, 2018. Neurochemistry predicts convergence of written and spoken language: a proton magnetic resonance spectroscopy study of cross-modal language integration. Front. Psychol 9, 1507. doi:10.3389/fpsyg.2018.01507. [PubMed: 30233445]

Depaoli S, Clifton JP, Cobb PR, 2016. Just another gibbs sampler (JAGS): flexible software for MCMC implementation. J. Educ. Behav. Stat 41 (6), 628–649. doi:10.3102/1076998616664876 .

Dinstein I, Heeger DJ, Behrmann M, 2015. Neural variability: friend or foe? Trends Cogn. Sci 19 (6), 322–328. doi:10.1016/j.tics.2015.04.005. [PubMed: 25979849]

Easson AK, McIntosh AR, 2019. BOLD signal variability and complexity in children and adolescents with and without autism spectrum disorder. Dev. Cognit. Neurosci 36, 100630. doi:10.1016/j.dcn.2019.100630. [PubMed: 30878549]

Eckert MA, Berninger VW, Vaden KI, Gebregziabher M, Tsu L, 2016. Gray matter features of reading disability: a combined meta-analytic and direct analysis approach. ENeuro (1) 3. doi:10.1523/ENEURO.0103-15.2015.

Eden GF, Vanmeter JW, Rumsey JM, Zeffird'O TA, 1996. The visual deficit theory of developmental dyslexia. Neuroimage 4 (3), S108–S117. doi:10.1006/nimg.1996.0061. [PubMed: 9345535]

Finn ES, Shen X, Holahan JM, Scheinost D, Lacadie C, Papademetris X, Shaywitz SE, Shaywitz BA, Constable RT, 2014. Disruption of functional networks in dyslexia: a whole-brain, data-driven analysis of connectivity. Biol. Psychiatry 76 (5), 397–404. doi:10.1016/j.biopsych.2013.08.031. [PubMed: 24124929]

Fletcher JM, Lyon GR, Fuchs LS, Barnes MA, 2007. Learning Disabilities: from Identification to Intervention. Guilford Press.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ, 1997. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6 (3), 218–229. doi:10.1006/nimg.1997.0291. [PubMed: 9344826]

Gelman A, Rubin DB, 1992. Inference from iterative simulation using multiple sequences. Statistical science 7 (4), 457–472.

Greene AS, Gao S, Scheinost D, Constable RT, 2018. Task-induced brain state manipulation improves prediction of individual traits. Nat. Commun 9 (1), 1–13. doi:10.1038/s41467-018-04920-3. [PubMed: 29317637]

Hancock R, Pugh KR, Hoeft F, 2017. Neural noise hypothesis of developmental dyslexia. Trends Cogn. Sci 21 (6), 434–448. doi:10.1016/j.tics.2017.03.008. [PubMed: 28400089]

Hoeft F, McCandliss BD, Black JM, Gantman A, Zakerani N, Hulme C, Lyytinen H, Whitfield-Gabrieli S, Glover GH, Reiss AL, Gabrieli JDE, 2011. Neural systems predicting long-term outcome in dyslexia. PNAS 108 (1), 361–366. doi:10.1073/pnas.1008950108. [PubMed: 21173250]

Hoeft F, Meyler A, Hernandez A, Juel C, Taylor-Hill H, Martindale JL, McMillon G, Kolchugina G, Black JM, Faizi A, Deutsch GK, Wai TS, Reiss AL, Whitfield-Gabrieli S, Gabrieli JDE, 2007. Functional and morphometric brain dissociation between dyslexia and reading ability. PNAS 104 (10), 4234–4239. doi:10.1073/pnas.0609399104. [PubMed: 17360506]

Hong YW, Yoo Y, Han J, Wager TD, Woo CW, 2019. False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. Neuroimage 195, 384–395. doi:10.1016/j.neuroimage.2019.03.070. [PubMed: 30946952]

Hornickel J, Kraus N, 2013. Unstable representation of sound: a biological marker of dyslexia. J. Neurosci 33 (8), 3500–3504. doi:10.1523/JNEUROSCI.4205-12.2013. [PubMed: 23426677]

Horwitz B, Rumsey JM, Donohue BC, 1998. Functional connectivity of the angular gyrus in normal reading and dyslexia. PNAS 95 (15), 8939–8944. doi:10.1073/pnas.95.15.8939. [PubMed: 9671783]

Jasińska KK, Shuai L, Lau A, Frost S, Landi N, Pugh KR, 2020. Functional connectivity in the developing language network in 4-year-old children predicts future reading ability. Dev. Sci doi:10.1111/desc.13041.

Koyama MS, Di Martino A, Kelly C, Jutagir DR, Sunshine J, Schwartz SJ, Castellanos FX, Milham MP, 2013. Cortical signatures of dyslexia and remediation: an intrinsic functional connectivity approach. PLoS ONE 8 (2), e55454. doi:10.1371/journal.pone.0055454. [PubMed: 23408984]

Kruschke JK, 2014. Doing bayesian data analysis: a tutorial with R, JAGS, and Stan. Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan, second editionsecond ed. doi:10.1016/B978-0-12-405888-0.09999-2.

Landi N, Malins JG, Frost SJ, Magnuson JS, Molfese P, Ryherd K, Rueckl JG, Mencl WE, Pugh KR, 2018. Neural representations for newly learned words are modulated by

overnight consolidation, reading skill, and age. Neuropsychologia 111, 133–144. doi:10.1016/j.neuropsychologia.2018.01.011. [PubMed: 29366948]

Lee MD, Wagenmakers EJ, 2013. Bayesian Cognitive Modeling: a Practical Course. Cambridge University Press doi:10.1017/CBO9781139087759.

Lindquist MA, Gelman A, 2009. Correlations and multiple comparisons in functional imaging: a statistical perspective (Commentary on Vul et al., 2009). Perspect. Psychol. Sci 4 (3), 310–313. doi:10.1111/j.1745-6924.2009.01130.x. [PubMed: 26158969]

Lobier M, Zoubrinetzky R, Valdois S, 2012. The visual attention span deficit in dyslexia is visual and not verbal. Cortex 48 (6), 768–773. doi:10.1016/j.cortex.2011.09.003. [PubMed: 21982580]

Lyon GR, 1995. Research initiatives in learning disabilities: Contributions from scientists supported by the National Institute of Child Health and Human Development. J. Child Neurol 10 (1), S120–S126. [PubMed: 7751548]

Malins JG, Gumkowski N, Buis B, Molfese P, Rueckl JG, Frost SJ, Pugh KR, Morris R, Mencl WE, 2016. Dough, tough, cough, rough: a “fast” fMRI localizer of component processes in reading. Neuropsychologia 91, 394–406. doi:10.1016/j.neuropsychologia.2016.08.027. [PubMed: 27592331]

Malins JG, Pugh KR, Buis B, Frost SJ, Hoeft F, Landi N, Mencl WE, Kurian A, Staples R, Molfese PJ, Sevcik R, Morris R, 2018. Individual differences in reading skill are related to trial-by-trial neural activation variability in the reading network. J. Neurosci 38 (12), 2981–2989. doi:10.1523/JNEUROSCI.0907-17.2018. [PubMed: 29440534]

Martin A, Schurz M, Kronbichler M, Richlan F, 2015. Reading in the brain of children and adults: a meta-analysis of 40 functional magnetic resonance imaging studies. Hum. Brain Mapp 36 (5), 1963–1981. doi:10.1002/hbm.22749. [PubMed: 25628041]

Maurer U, Brem S, Bucher K, Kranz F, Benz R, Steinhausen HC, Brandeis D, 2007. Impaired tuning of a fast occipito-temporal response for print in dyslexic children learning to read. Brain 130 (12), 3200–3210. doi:10.1093/brain/awm193. [PubMed: 17728359]

Maurer U, Schulz E, Brem S, der Mark S, Bucher K, Martin E, Brandeis D, 2011. The development of print tuning in children with dyslexia: evidence from longitudinal ERP data supported by fMRI. Neuroimage 57 (3), 714–722. doi:10.1016/j.neuroimage.2010.10.055. [PubMed: 21040695]

Morken F, Helland T, Hugdahl K, Specht K, 2017. Reading in dyslexia across literacy development: a longitudinal study of effective connectivity. Neuroimage 144 (A), 92–100. doi:10.1016/j.neuroimage.2016.09.060. [PubMed: 27688204]

Neef NE, Müller B, Liebig J, Schaadt G, Grigutsch M, Gunter TC, Wilcke A, Kirsten H, Skeide MA, Kraft I, Kraus N, Emmrich F, Brauer J, Boltze J, Friederici AD, 2017. Dyslexia risk gene relates to representation of sound in the auditory brainstem. Dev. Cognit. Neurosci 24, 63–71. doi:10.1016/j.dcn.2017.01.008. [PubMed: 28182973]

Ortega A, Wagenmakers EJ, Lee MD, Markowitsch HJ, Piefke M, 2012. A bayesian latent group analysis for detecting poor effort in the assessment of malingering. Arch. Clin. Neuropsychol 27 (4), 453–465. doi:10.1093/arclin/acs038. [PubMed: 22543568]

Paulesu E, Danelli L, Berlingeri M, 2014. Reading the dyslexic brain: multiple dys-functional routes revealed by a new meta-analysis of PET and fMRI activation studies. Front. Hum. Neurosci, 8, 830. doi:10.3389/fnhum.2014.00830 . [PubMed: 25426043]

Pennington BF, 2006. From single to multiple deficit models of developmental disorders. Cognition 101 (2), 385–413. doi:10.1016/j.cognition.2006.04.008. [PubMed: 16844106]

Pennington BF, Bishop DVM, 2009. Relations among speech, language, and reading disorders. Annu. Rev. Psychol 60, 283–306. doi:10.1146/annurev.psych.60.110707.163548. [PubMed: 18652545]

Plummer M, 2016. rjags: Bayesian graphical models using MCMC. R Package Version 4-6 http://cran.r-project.org/package=rjags .

Power JD, Cohen AL, Nelson SM , Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL, Petersen SE, 2011. Functional network organization of the human brain. Neuron 72 (4), 665–678. [PubMed: 22099467]

Pugh KR, Frost SJ, Rothman DL, Hoeft F, Del Tufo SN, Mason GF, Molfese PJ, Einar Mencl W, Grigorenko EL, Landi N, Preston JL, Jacobsen L, Seidenberg MS, Fulbright RK, 2014. Glutamate

and choline levels predict individual differences in reading ability in emergent readers. J. Neurosci 34 (11), 4082–4089. doi:10.1523/JNEUROSCI.3907-13.2014. [PubMed: 24623786]

Pugh KR, Frost SJ, Sandak R, Landi N, Rueckl JG, Constable RT, Seidenberg MS, Fulbright RK, Katz L, Mencl WE, 2008. Effects of stimulus difficulty and repetition on printed word identification: an fMRI comparison of nonimpaired and reading-disabled adolescent cohorts. J. Cogn. Neurosci 20 (7), 1146–1160. doi:10.1162/jocn.2008.20079. [PubMed: 18284344]

Pugh KR, Mencl WE, Jenner AR, Katz L, Frost SJ, Lee JR, Shaywitz SE, Shaywitz BA, 2000. Functional neuroimaging studies of reading and reading disability (developmental dyslexia). Mental Retard. Dev. Disabil 6, 207–213. doi:10.1002/1098-2779(2000)6:3<207::AID-MRDD8>3.0.CO;2.

Pugh KR, Mencl WE, Jenner AR, Lee JR, Katz L, Frost SJ, Shaywitz SE, Shaywitz BA, 2001. Neuroimaging studies of reading development and reading disability. Learn. Disabil. Res. Pract 16 (4), 240–249. doi:10.1111/0938-8982.00024.

Ramus F, Altarelli I, Jednoróg K, Zhao J, Scotto di Covella L, 2018. Neuroanatomy of developmental dyslexia: pitfalls and promise. Neurosci. Biobehav. Rev 84, 434–452. doi:10.1016/j.neubiorev.2017.08.001. [PubMed: 28797557]

Richards TL, Berninger VW, 2008. Abnormal fMRI connectivity in children with dyslexia during a phoneme task: before but not after treatment. J. Neurolinguistics 21 (4), 294–304. doi:10.1016/j.jneuroling.2007.07.002. [PubMed: 19079567]

Richlan F, Kronbichler M, Wimmer H, 2009. Functional abnormalities in the dyslexic brain: a quantitative meta-analysis of neuroimaging studies. Hum. Brain Mapp 30 (10), 3299–3308. doi:10.1002/hbm.20752. [PubMed: 19288465]

Richlan F, Kronbichler M, Wimmer H, 2013. Structural abnormalities in the dyslexic brain: a meta-analysis of voxel-based morphometry studies. Hum. Brain Mapp 34 (11), 3055–3065. doi:10.1002/hbm.22127. [PubMed: 22711189]

Rissman J, Gazzaley A, D'Esposito M, 2004. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23 (2), 752–763. doi:10.1016/j.neuroimage.2004.06.035. [PubMed: 15488425]

Rumsey JM, Aquino T, King AC, Hamburger SD, Rapoport JL, Andreason P, Zametkin AJ, Cohen RM, Pikus A, 1992. Failure to activate the left temporoparietal cortex in dyslexia: an oxygen 15 positron emission tomographic study. Arch. Neurol 49 (5), 527–534. doi:10.1001/archneur.1992.00530290115020. [PubMed: 1580816]

Salmelin R, Service E, Kiesilä P, Uutela K, Salonen O, 1996. Impaired visual word processing in dyslexia revealed with magnetoencephalography. Ann. Neurol 40 (2), 157–162. doi:10.1002/ana.410400206. [PubMed: 8773596]

Shaywitz BA, Shaywitz SE, Pugh KR, Mencl WE, Fulbright RK, Skudlarski P, Constable RT, Marchione KE, Fletcher JM, Lyon GR, Gore JC, 2002. Disruption of posterior brain systems for reading in children with developmental dyslexia. Biol. Psychiatry 52 (2), 101–110. doi:10.1016/S0006-3223(02)01365-3. [PubMed: 12114001]

Shaywitz BA, Skudlarski P, Holahan JM, Marchione KE, Constable RT, Fulbright RK, Zelterman D, Lacadie C, Shaywitz SE, 2007. Age-related changes in reading systems of dyslexic children. Ann. Neurol 61 (4), 363–370. doi:10.1002/ana.21093. [PubMed: 17444510]

Shaywitz SE, Fletcher JM, Holahan JM, Shneider AE, Marchione KE, Stuebing KK, Francis DJ, Pugh KR, Shaywitz BA, 1999. Persistence of dyslexia: the Connecticut longitudinal study at adolescence. Pediatrics 104 (6). doi:10.1542/peds.104.6.1351 .

Shaywitz SE, Shaywitz BA, Fulbright RK, Skudlarski P, Mencl WE, Constable RT, Pugh KR, Holahan JM, Marchione KE, Fletcher JM, Lyon GR, Gore JC, 2003. Neural systems for compensation and persistence: young adult outcome of childhood reading disability. Biol. Psychiatry 54 (1), 25–33. doi:10.1016/S0006-3223(02)01836-X . [PubMed: 12842305]

Shaywitz SE, Shaywitz BA, Pugh KR, Fulbright RK, Constable RT, Mencl WE, Shankweiler DP, Liberman AM, Skudlarski P, Fletcher JM, Katz L, Marchione KE, Lacadie C, Gatenby C, Gore JC, 1998. Functional disruption in the organization of the brain for reading in dyslexia. PNAS 95 (5), 2636–2641. doi:10.1073/pnas.95.5.2636. [PubMed: 9482939]

Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, Constable RT, 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc 12 (3), 506–518. doi:10.1038/nprot.2016.178. [PubMed: 28182017]

Siegelman N, Bogaerts L, Armstrong BC, Frost R, 2019. What exactly is learned in visual statistical learning? Insights from Bayesian modeling. Cognition 192, 104002. doi:10.1016/j.cognition.2019.06.014 . [PubMed: 31228679]

Siegelman N, Rueckl JG, Steacy LM, Frost SJ, van den Bunt M, Zevin JD, Seidenberg MS, Pugh KR, Compton DL, Morris RD, 2020. Individual differences in learning the regularities between orthography, phonology and semantics predict early reading skills. J. Memory Lang 114, 104145. doi:10.1016/j.jml.2020.104145.

Simos PG, Fletcher JM, Bergman E, Breier JI, Foorman BR, Castillo EM, Davis RN, Fitzgerald M, Papanicolaou AC, 2002. Dyslexia-specific brain activation profile becomes normal following successful remedial training. Neurology 58 (8), 1203–1213. doi:10.1212/WNL.58.8.1203 . [PubMed: 11971088]

Sklar B, Hanley J, Simmons WW, 1972. An EEG experiment aimed toward identifying dyslexic children. Nature 240 (5381), 414–416. doi:10.1038/240414a0. [PubMed: 4564321]

Skudlarski P, Constable RT, Gore JC, 1999. ROC analysis of statistical methods used in functional MRI: individual subjects. Neuroimage 9 (3), 311–329. doi:10.1006/nimg.1999.0402. [PubMed: 10075901]

Snowling MJ, Hulme C, 2020. Annual research review: reading disorders revisited – the critical importance of oral language. J. Child Psychol. Psychiatry 62 (5), 635–653. doi:10.1111/jcpp.13324. [PubMed: 32956509]

Steingroever H, Jepma M, Lee MD, Jansen BRJ, Huizenga HM, 2019. Detecting strategies in developmental psychology. Comput. Brain Behav 2 (2), 128–140. doi:10.1007/s42113-019-0024-x.

Szucs D, Ioannidis JPA, 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 15 (3), e2000797. doi:10.1371/journal.pbio.2000797. [PubMed: 28253258]

Tamboer P, Vorst HCM, Ghebreab S, Scholte HS, 2016. Machine learning and dyslexia: classification of individual structural neuro-imaging scans of students with and without dyslexia. NeuroImage 11, 508–514. doi:10.1016/j.nicl.2016.03.014. [PubMed: 27114899]

Tanaka H, Black JM, Hulme C, Stanley LM, Kesler SR, Whitfield-Gabrieli S, Reiss AL, Gabrieli JDE, Hoeft F, 2011. The brain basis of the phonological deficit in dyslexia is independent of IQ. Psychol. Sci 22 (11), 1442–1451. doi:10.1177/0956797611419521. [PubMed: 22006060]

Torgesen JK, Wagner R. Rashotte C. 2012. TOWRE 2: Test of Word Reading Efficiency. Pro-Ed Inc.

van der Mark S, Bucher K, Maurer U, Schulz E, Brem S, Buckelmüller J, Kronbichler M, Loenneker T, Klaver P, Martin E, Brandeis D, 2009. Children with dyslexia lack multiple specializations along the visual word-form (VWF) system. Neuroimage 47 (4), 1940–1949. doi:10.1016/j.neuroimage.2009.05.021. [PubMed: 19446640]

van der Mark S, Klaver P, Bucher K, Maurer U, Schulz E, Brem S, Martin E, Brandeis D, 2011. The left occipitotemporal system in reading: Disruption of focal fMRI connectivity to left inferior frontal and inferior parietal language areas in children with dyslexia. Neuroimage 54 (3), 2426–2436. doi:10.1016/j.neuroimage.2010.10.002. [PubMed: 20934519]

Wai TS, Niu Z, Jin Z, Perfetti CA, Li HT, 2008. A structural-functional basis for dyslexia in the cortex of Chinese readers. PNAS 105 (14), 5561–5566. doi:10.1073/pnas.0801750105. [PubMed: 18391194]

Waldie KE, Haigh CE, Badzakova-Trajkov G, Buckley J, Kirk IJ, 2013. Reading the wrong way with the right hemisphere. Brain Sci. 3 (3), 1060–1075. doi:10.3390/brainsci3031060. [PubMed: 24961521]

Wang JX, Bartolotti J, Amaral LAN, Booth JR, 2013. Changes in task-related functional connectivity across multiple spatial scales are related to reading performance. PLoS ONE 8 (3), e59204. doi:10.1371/journal.pone.0059204. [PubMed: 23544057]

Zoubrinetzky R, Bielle F, Valdois S, 2014. New insights on developmental dyslexia subtypes: Heterogeneity of mixed reading profiles. PLoS ONE 9 (6), e99337. doi:10.1371/journal.pone.0099337. [PubMed: 24918441]

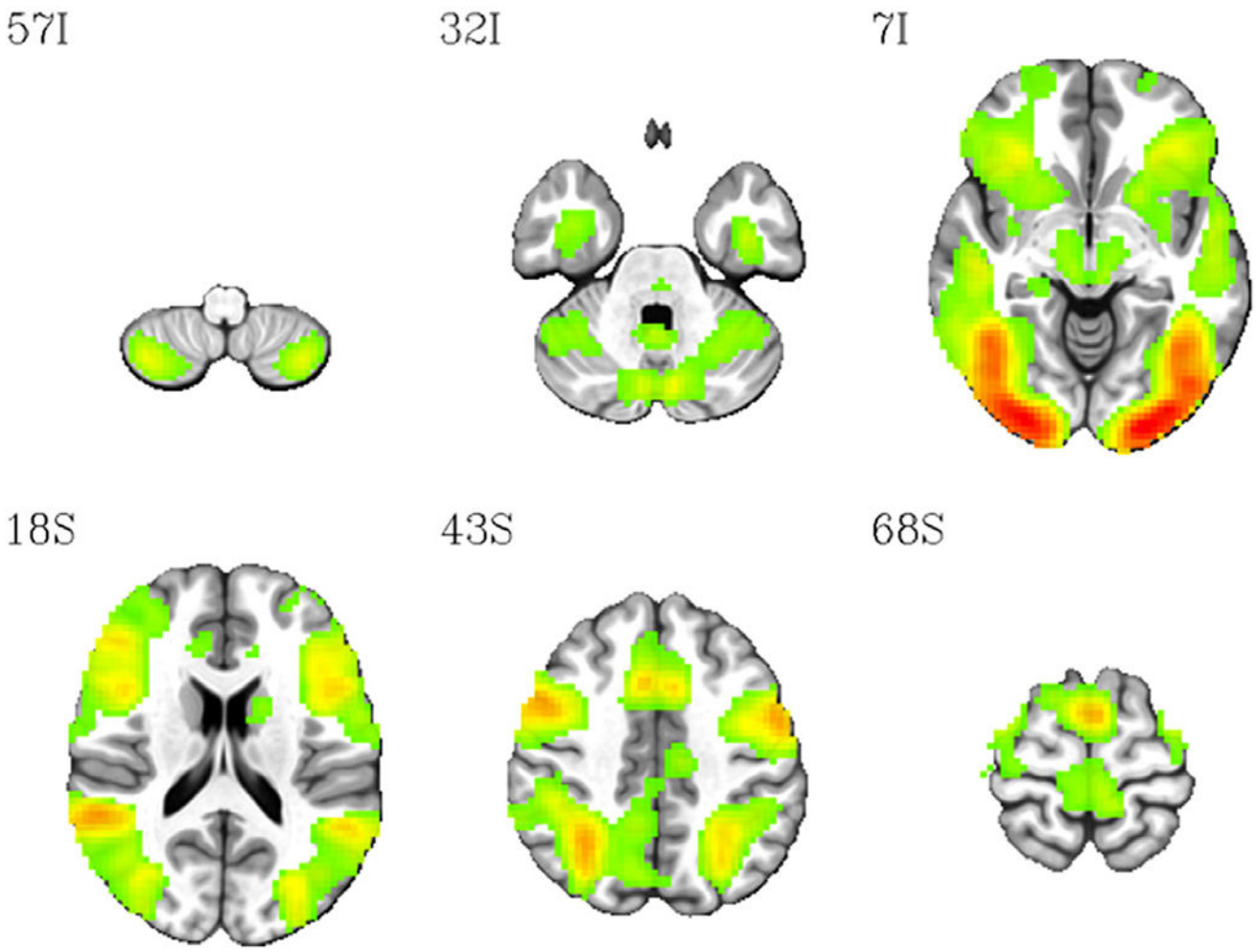

Fig. 1.
Voxel-by-voxel significant activation ($p < .001$) for the print condition.

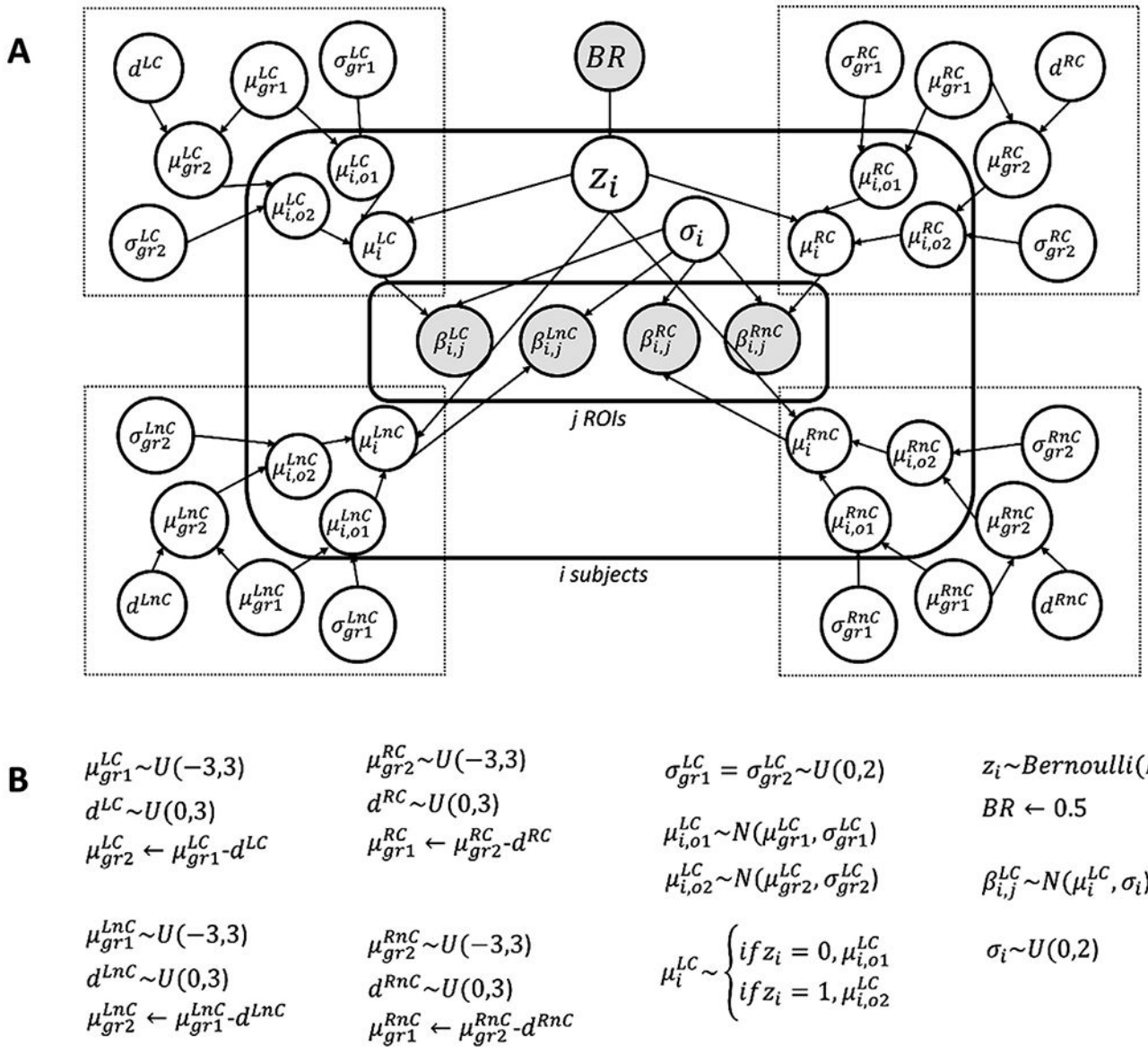

Fig. 2.

Visual depiction of activation-based models. Panel A: models' architecture. The notation follows that of Lee and Wagenmakers (2013): White nodes: unobserved (i.e., to-be-estimated) parameters; grey nodes: observed data or known values; rounded rectangles: loops (running over subjects/trials). The subscript *i* to refer to subjects and *j* to refer to ROIs (e.g., β_{ij} - activation for the subject *i* in ROI *j*) . The four sub-networks of regions (i.e., the left canonical network, left non-canonical network, right canonical network, and right non-canonical networks) are shown in different quartiles within dashed boxes and are marked with different superscripts (Left side: left hemisphere, marked ^{L}; Right side: right hemisphere, marked with ^{R}; Top: canonical regions, marked with ^{C}; Bottom: non-canonical regions, marked with ^{nC}). Panel B: central priors in the *left-right mean activation model* (model 1A). In the left-canonical and left non-canonical networks, Group

1′s population means can a priori be any standard value between −3 and +3 ($\mu_{gr1}^{LC} \sim U(-3, 3)$, $\mu_{gr1}^{LnC} \sim U(-3, 3)$), a range which presumably includes any reasonable true value (i.e., an uninformative prior). Group 2′s means in the same networks are expected to be smaller than that of Group 1, and are thus defined as the population mean of Group 1 minus a difference parameter ($\mu_{gr2}^{LC} \leftarrow \mu_{gr1}^{LC} - d^{LC}$; $\mu_{gr2}^{LnC} \leftarrow \mu_{gr1}^{LnC} - d^{LnC}$). The difference parameters can a priori take any positive value up to 3 standard deviations ($d^{LC} \sim U(0, 3)$, $d^{LnC} \sim U(0, 3)$). In the right hemisphere, in contrast, the population means for Group 2 are higher than for Group 1 (in both canonical and non-canonical ROIs): Group 2′s means follow an uninformative prior ($\mu_{gr2}^{RC} \sim U(-3, 3)$, $\mu_{gr2}^{RnC} \sim U(-3, 3)$), and the population mean for Group 1 is *smaller* by a difference parameter from these values (i.e. $d^{RC} \sim U(0, 3)$, $d^{RnC} \sim U(0, 3)$; and $\mu_{gr1}^{RC} \leftarrow \mu_{gr2}^{RC} - d^{RC}$, $\mu_{gr1}^{RnC} \leftarrow \mu_{gr2}^{RnC} - d^{RnC}$). Priors that are not listed for the left non-canonical, right-canonical, and right non-canonical sub-networks are identical in their specification to the left-canonical network. See Supplementary Materials S4 for further details, and the project's OSF page for full codes.

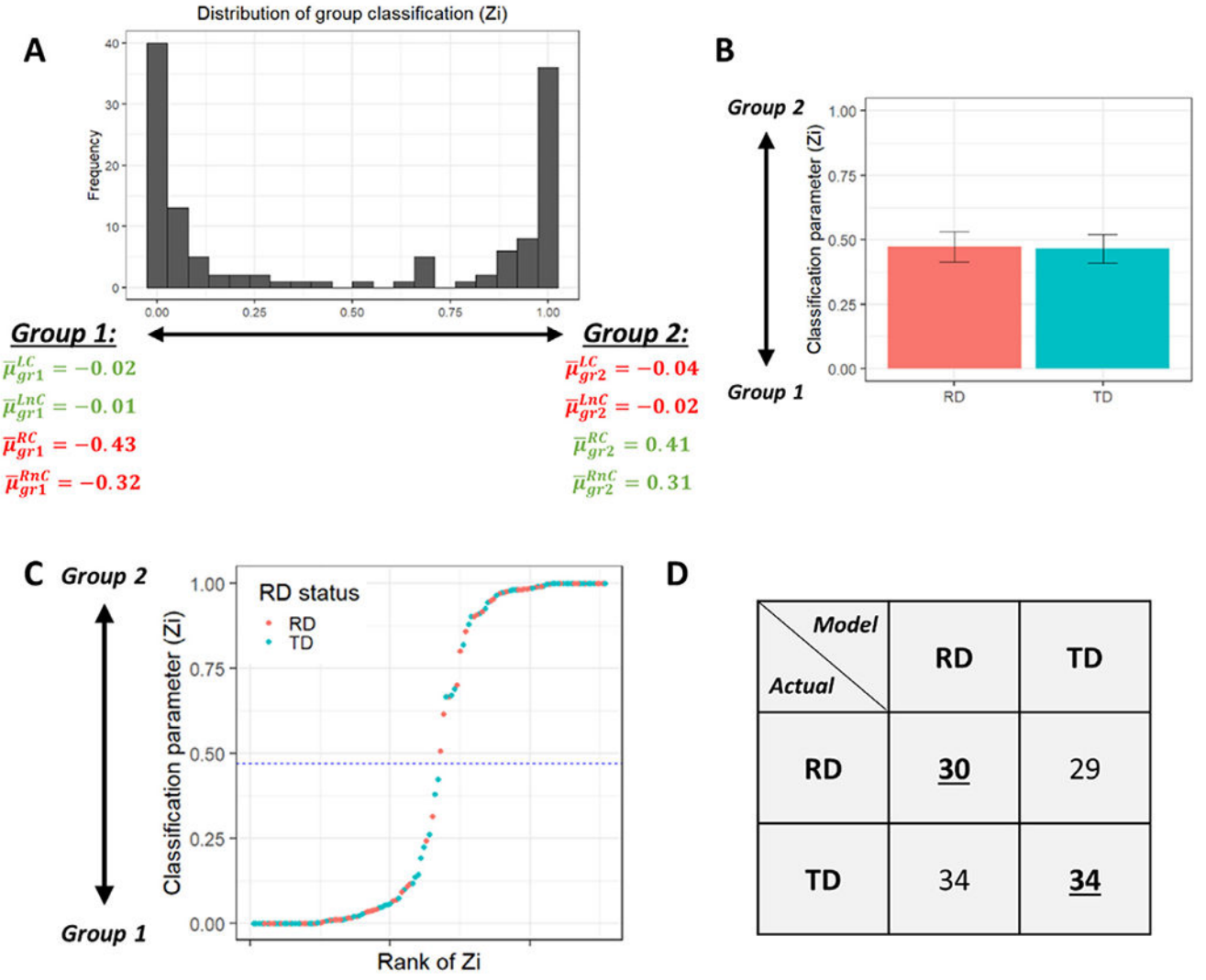

Fig. 3.

Results of the *left-right mean activation model (model 1A)*. Panel A: histogram of the group classification parameter (z_i) and estimated population means (means of posterior distributions). Population means' estimates in green were constrained to be larger than those in red. Panel B: mean group classification parameter in the two subject groups (RD and TD individuals). Panel C: individual-level distribution of the group classification parameter (y-axis: estimated z_i values; x-axis: ranks of z_i) and behavioral RD/TD status (in color). The horizontal line presents the median estimated z_i across individuals, used as a threshold for the dichotomous classification. Panel D: cross-tabulation of dichotomous classification success. Counts of successful classification (on the diagonal) are in bold/underline font. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

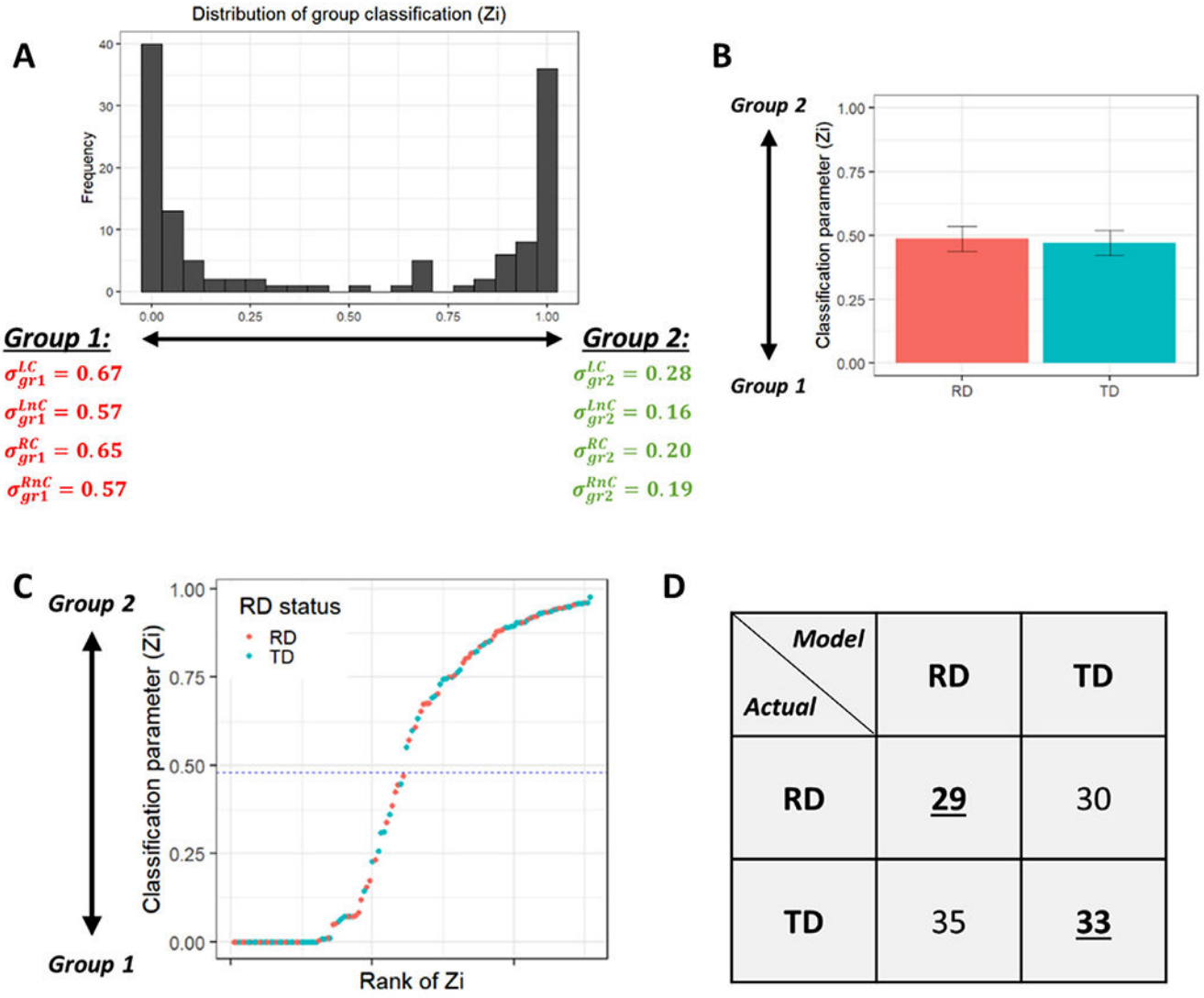

Fig. 4.

Results of the *heterogeneity-based model (model 2)*. Panel A: histogram of the group classification parameter (z_i) and estimated population standard deviations (means of posterior distributions). Standard deviation parameters in green were constrained to be smaller (i.e. more homogeneous) than those in red. Panel B: mean group classification parameter in the two subject groups (RD and TD individuals). Panel C: individual-level distribution of the group classification parameter (y-axis: estimated z_i values; x-axis: ranks of z_i) and behavioral RD/TD status (in color). The horizontal line presents the median estimated z_i across individuals, used as a threshold for the dichotomous classification. Panel D: cross-tabulation of dichotomous classification success. Counts of successful classification (on the diagonal) are in bold/underline font. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

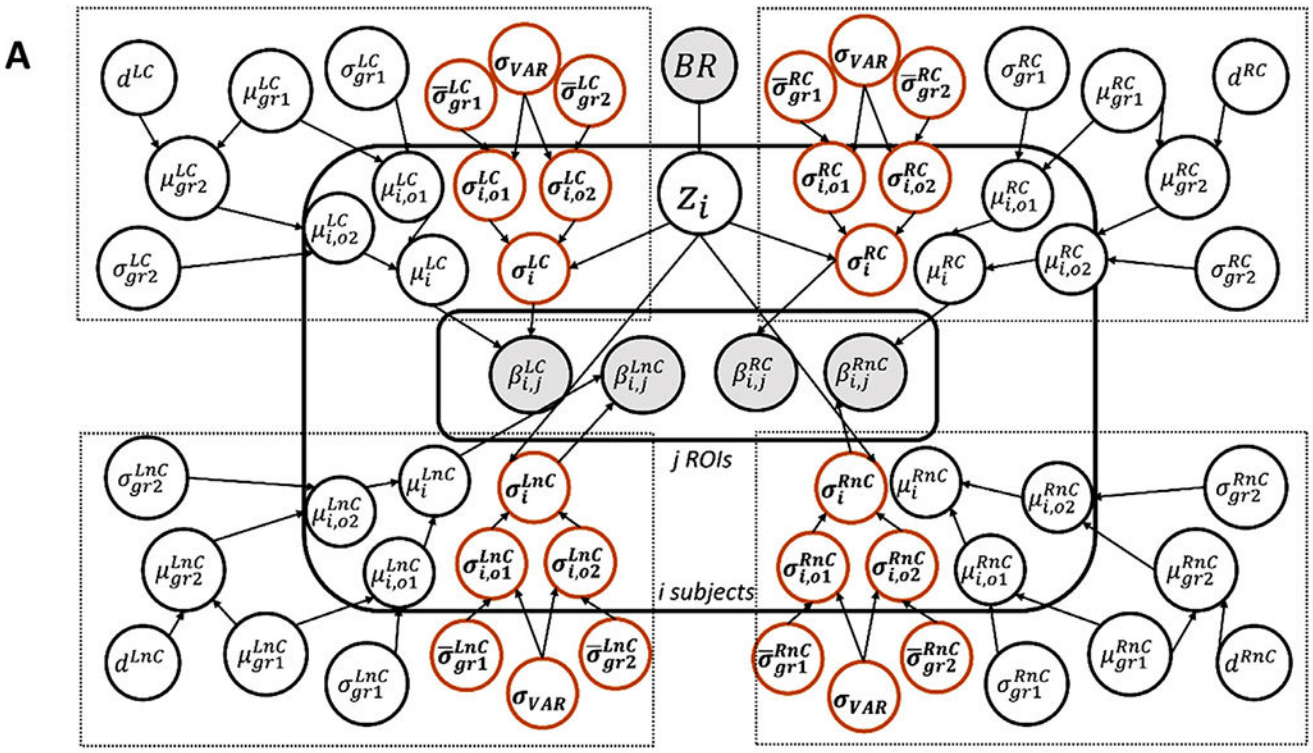

Fig. 5.

Visual depiction of the (intra-subject intra-region) variability-based model. Panel A: the model's architecture (see caption of Fig. 2 for notation legend). Parameters circled in orange and bold were not part of the parameters in the mean- and heterogeneity-based models above (shown in Fig. 2). Panel B: central priors in the variability-based model. Mean and inter-subject variability were constrained to be equal in the two groups. Importantly, the variability mean in Group 1 was constrained to be a priori larger than that of Group 2: the mean intra-subject variability in Group 1 could a priori receive any value between 0 and 2 standard deviations (e.g., in the left-canonical network: $\bar{\sigma}_{gr1}^{LC} \sim U(0,2)$), while the mean intra-subject variability in Group 2 was constrained to be smaller than this value (yet still larger than zero; $\bar{\sigma}_{gr2}^{LC} \sim U(0, \bar{\sigma}_{gr1}^{LC})$. Priors regarding mean activation that are not listed here are identical to those in the mean-activation models above. Priors that are not listed for the left non-canonical, right-canonical and right non-canonical sub-networks are similar to the left-canonical network. See Supplementary Materials S4 for details, and the project's OSF

page for full codes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

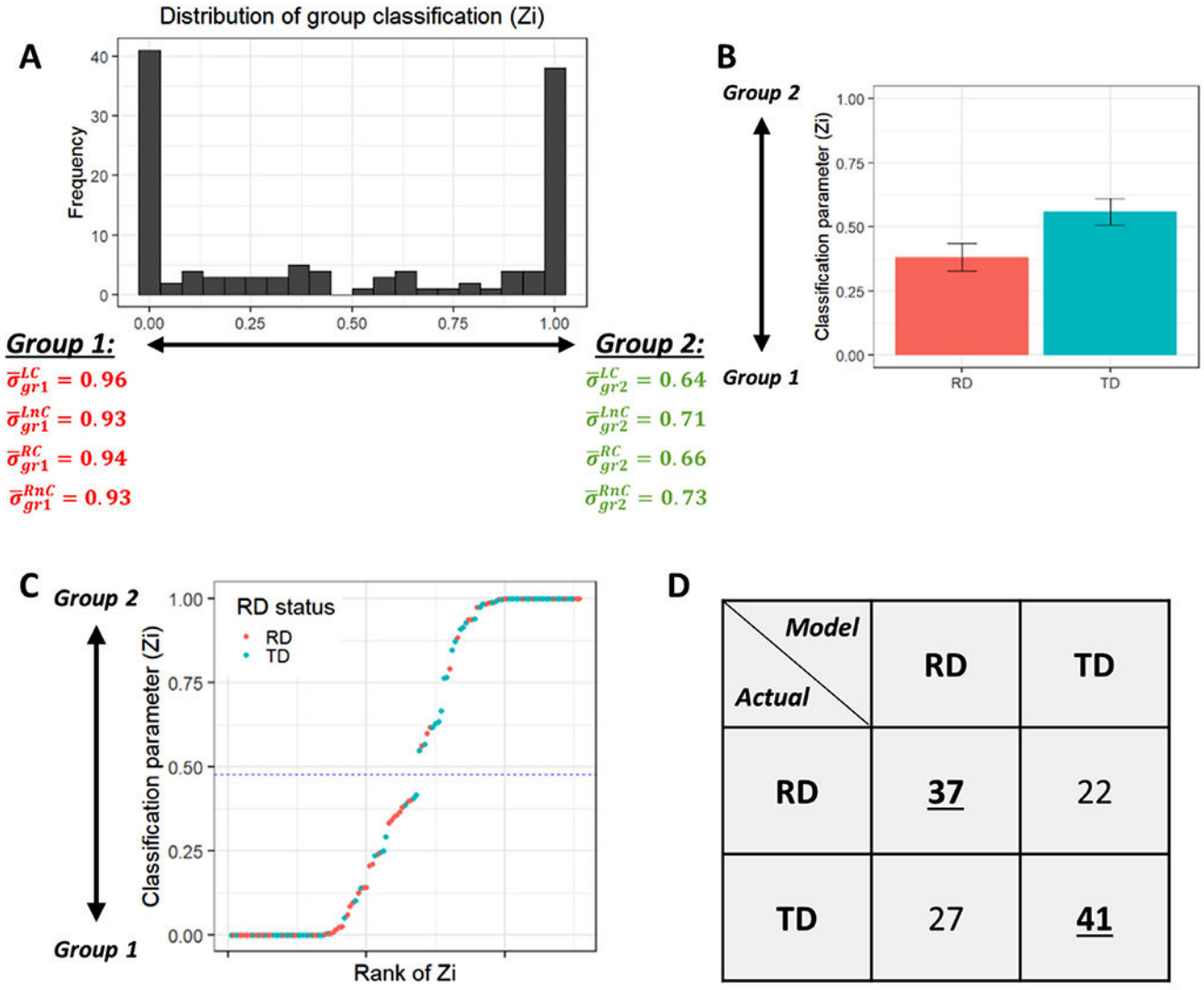

Fig. 6.

Results of the *variability-based model (model 3)*. Panel A: histogram of the group classification parameter (z_i) and estimated mean inter-region standard deviations (means of posterior distributions). Standard deviation parameters in green were constrained to be smaller (i.e., less variable) than those in red. Panel B: mean group classification parameter in the two subject groups (RD and TD individuals). Panel C: individual-level distribution of the group classification parameter (y-axis: estimated z_i values; x-axis: ranks of z_i) and behavioral RD/TD status (in color). The horizontal line presents the median estimated z_i across individuals, used as a threshold for the dichotomous classification. Panel D: Cross-tabulation of dichotomous classification success. Counts of successful classification (on the diagonal) are in bold/underline font. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

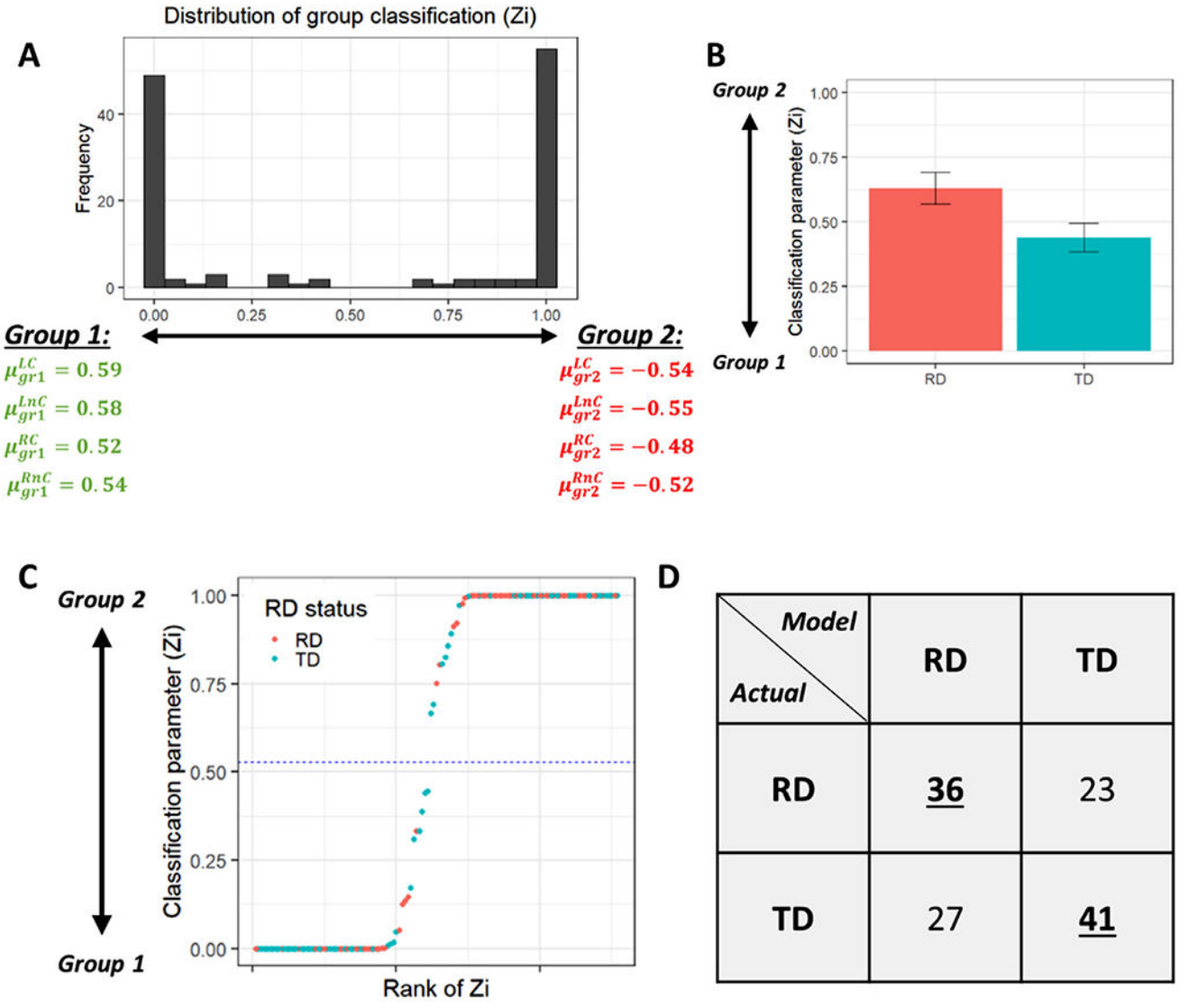

Fig. 7.

Results of the *connectivity-based, model (model 4)*. Panel A: histogram of the group classification parameter (z_i) and estimated mean connectivity values (means of posterior distributions). Note that in this model these means reflect connectivity values (*r*-to-*z*-transformed values), not mean activation. Parameters in green were constrained to be larger (i.e. more connectivity) than those in red. Panel B: mean group classification parameter in the two subject groups (RD and TD individuals). Panel C: individual-level distribution of the group classification parameter (y-axis: estimated z_i values; x-axis: ranks of z_i) and behavioral RD/TD status (in color). The horizontal line presents the median estimated z_i across individuals, used as a threshold for the dichotomous classification. Panel D: Cross-tabulation of dichotomous classification success. Counts of successful classification (on the diagonal) are in bold/underline font. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 1

Basic characteristics of the two participant groups.

	RD group	TD group	Comparison
TOWRE:Sight Word Efficiency[a]	M = 82.19, SD = 7.51	M = 103.04, SD = 8.87	$t(125) = 14.19$, $p < .001$
TOWRE:Phonemic Decoding Efficiency[a]	M = 88.29, SD = 14.02	M = 103.78, SD = 9.56	$t(125) = 7.35$, $p < .001$
Age	M = 18.95, SD = 2.64	M = 20.69, SD = 2.56	$t(125) = 3.77$, $p < .001$
Gender	38 Males; 21 Females	34 Males; 34 Females	$\chi^2_{(1)} = 2.67$, p = .10
Proportion motion-censored TRs[b]	M = 0.10, SD = 0.14	M = 0.07, SD = 0.10	$t(125) = 1.37$, $p = .17$

[a] Values are standard scores.

[b] defined as the proportion of volumes that exceeded a thresholds of 0.3 mm Euclidean movement and/or with more than 10% of voxels marked as outliers by the *regress_censor_outliers* flag.

Table 2

MNI coordinates of the canonical left (and right) reading ROIs, derived from Martins et al. (2015).

Region	MNI coordinates		
	x	y	z
Inferior Frontal Gyrus (BA45)	(−)52	20	18
Inferior Frontal Gyrus (BA44)	(−)52	18	14
Precentral Gyrus	(−)46	2	42
Middle Frontal Gyrus	(−)42	4	48
Fusiform Gyrus	(−)42	−68	−22
Inferior Occipital Gyrus	(−)44	−74	−4
Middle Occipital Gyrus	(−)42	−86	−2
Inferior Temporal Gyrus	(−)48	−62	−20
Supplementary Motor Area	(−)4	24	56
Intra-Parietal Sulcus	(−)42	−48	48
Temporal Pole[a]	(−)52	4	−10

[a]Group-level activation for printed words in this left-canonical ROI did not reach significance, and it was therefore not included in the corresponding input matrix.

Table 3

Results of a logistic regression model predicting behavioral group membership from the group classification parameter (z_i) in the intra-subject variability-based model while controlling for age and motion.

Predictor	Estimate (β)	SE	*Z*-value	*p*-value
Classification parameter (z_i)	1.37	0.52	2.65	.008
Age	0.27	0.08	3.42	<.001
Gender[a]	−0.82	0.42	−1.96	.050
Motion	0.44	1.83	0.24	.809

Notes: SE = Standard Error. Significant *p*-values are shown in bold.

[a]dummy-coded variable, reference level set to “male”.

Table 4

Results of a logistic regression model predicting behavioral group membership from the group classification parameter (z_i) in the connectivity-based model while controlling for age and motion.

Predictor	Estimate (β)	SE	*Z*-value	*p*-value
Classification parameter (z_i)	−1.02	0.43	−2.36	0.018
Age	0.23	0.07	3.03	0.002
Gender[a]	−0.54	0.40	−1.36	0.173
Motion	−2.45	1.86	−1.32	0.187

Notes: SE = Standard Error. Significant *p*-values are shown in bold.

[a]dummy-coded variable, reference level set to “male”.

Table 5

Results of a logistic regression model predicting behavioral group membership from the group classification parameters (z_i) of both the variability- and connectivity-based models, as well as control variables (age and motion).

Predictor	Estimate (β)	SE	*Z*-value	*p*-value
z_i: Variability-based model	1.45	0.54	2.68	0.007
z_i: Connectivity-based model	−1.09	0.45	−2.40	0.017
Age	0.25	0.08	3.24	0.001
Gender[a]	−0.90	0.44	−2.05	0.040
Motion	−0.67	1.96	−0.34	0.732

Notes: SE = Standard Error. z_i = Estimated classification parameter; Significant *p*-values are shown in bold.

[a]dummy-coded variable, reference level set to “male”.

Table 6

Cross-tabulation showing the relation between actual (i.e. behavioral) group membership and the conjunction of the dichotomous classification based on the variability-based and connectivity-based models. Values in bold/underline show counts of successful classification among individuals who had similar classification under both models.

		Models' classification			
		RD both models	**RD connectivity only**	**RD variability only**	**TD both models**
Actual status	RD	22	14	15	8
	TD	10	17	17	24