Cerebral Cortex, May 2018;28: 1656–1672

doi: 10.1093/cercor/bhx061
Advance Access Publication Date: 10 March 2017
Original Article

ORIGINAL ARTICLE

Individualized Prediction of Reading Comprehension Ability Using Gray Matter Volume

Zaixu Cui[1], Mengmeng Su[1], Liangjie Li[1], Hua Shu[1] and Gaolang Gong[1,2]

[1]State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China and [2]Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing 100875, China

Address correspondence to Gaolang Gong. Email: gaolang.gong@bnu.edu.cn; Hua Shu. Email: shuhua@bnu.edu.cn

Zaixu Cui and Mengmeng Su contributed to the work equally.

Abstract

Reading comprehension is a crucial reading skill for learning and putatively contains 2 key components: reading decoding and linguistic comprehension. Current understanding of the neural mechanism underlying these reading comprehension components is lacking, and whether and how neuroanatomical features can be used to predict these 2 skills remain largely unexplored. In the present study, we analyzed a large sample from the Human Connectome Project (HCP) dataset and successfully built multivariate predictive models for these 2 skills using whole-brain gray matter volume features. The results showed that these models effectively captured individual differences in these 2 skills and were able to significantly predict these components of reading comprehension for unseen individuals. The strict cross-validation using the HCP cohort and another independent cohort of children demonstrated the model generalizability. The identified gray matter regions contributing to the skill prediction consisted of a wide range of regions covering the putative reading, cerebellum, and subcortical systems. Interestingly, there were gender differences in the predictive models, with the female-specific model overestimating the males' abilities. Moreover, the identified contributing gray matter regions for the female-specific and male-specific models exhibited considerable differences, supporting a gender-dependent neuroanatomical substrate for reading comprehension.

Key words: cross-validation, individual difference, linguistic comprehension, reading comprehension, reading decoding

Introduction

Reading is a unique skill to humans and is crucial for learning and career achievement in modern societies. Individual reading abilities differ greatly, and a number of neuroimaging studies have been devoted to investigate the neural basis for these individual differences (Shaywitz 1998; Hoeft et al. 2007; Richlan et al. 2013). The results, however, are far from conclusive.

To explore individual differences in brain and cognition, the majority of neuroimaging studies have chosen to correlate neuroimaging measures with cognitive scores across all recruited individuals. However, this strategy is limited by the uncertainty of whether the observed correlational result can be generalized to unseen individuals (Gabrieli et al. 2015; Dubois and Adolphs 2016). To overcome this limitation, the cross-validation (CV) approach can been applied, which inherently evaluates the model's ability to predict the outcome for a previously unseen individual (Dosenbach et al. 2010; Ullman et al. 2014; Finn et al. 2015; Rosenberg et al. 2015). In such an approach, a neuroimaging-based predictive model for specific cognitive scores is initially estimated using training samples and is subsequently validated by independent testing samples. Once the prediction performs well on the testing dataset, individual differences in relevant cognition can be efficiently captured with the estimated model. Furthermore, the predictive features adopted by the model can be used as effective neuroimaging markers for the corresponding cognition.

© The Author 2017. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

To date, only one neuroimaging study has explored the CV prediction for reading-related abilities, however, this study targets locating the reading-related brain regions by applying a searchlight algorithm, rather than evaluating the individualized prediction for the reading-related abilities per se (He et al. 2013). In particular, this study did not perform an analysis of reading comprehension, an important reading ability necessary to effectively comprehend information from the text input (Gough and Tunmer 1986). Regarding this ability, the influential model termed "Simple View of Reading" suggests that reading comprehension is mainly the product of 2 components: decoding ability and linguistic comprehension (Gough and Tunmer 1986; Hoover and Gough 1990). Whether and how neuroanatomical features can be used to predict these 2 crucial reading comprehension components remain largely unexplored, particularly using the neuroimaging-based CV method.

Notably, there are well-observed gender differences in reading skills, with females typically performing better than males (Chiu and McBride-Chang 2006). In line with this, numerous brain imaging studies have demonstrated significant gender effects on the functional organization of the reading brain in both adults (Shaywitz et al. 1995; Rossell et al. 2002) and children (Burman et al. 2008). For example, greater bilateral brain activation was observed among females, while relatively lateralized brain activation was observed in males (Shaywitz et al. 1995; Rossell et al. 2002; Burman et al. 2008). Therefore, it is likely that there exists a significant gender effect on the neuroimaging-based prediction of reading-related abilities, including the 2 main reading comprehension abilities: reading decoding and linguistic comprehension. However, this hypothesis has never been empirically evaluated.

The recently released Human Connectome Project (HCP) dataset provides an opportunity to explore these issues or hypotheses, and a few recent studies have applied this dataset to investigate the individualized prediction of fluid intelligence (Finn et al. 2015) and impulsivity (Marquand et al. 2016). Specifically, the HCP dataset includes high-quality magnetic resonance imaging (MRI) scans and a battery of cognitive tests for a large number of healthy adults (Van Essen et al. 2012, 2013; Barch et al. 2013). Importantly, 2 specific cognitive tests for measuring the reading decoding and linguistic comprehension abilities have been applied to each individual in this dataset, making it possible to thoroughly study neuroimaging-based predictive models for the 2 reading comprehension components.

In the present study, we sought to ascertain whether the whole-brain gray matter (GM) volume (GMV) pattern could effectively predict the 2 reading comprehension abilities in previously unseen individuals. Specifically, the HCP dataset was applied, and structural MRI data were used to extract GMV features for each individual. An elastic-net penalized linear regression was adopted to achieve a sparse model, and a strict 3-fold CV was employed to evaluate the generalizability of the estimated model. Particularly, our local center dataset, which consisted of 67 Chinese children, was further included to evaluate the model's generalizability. Finally, we explored gender effects on the predictive models by comparing gender-specific models for the 2 reading comprehension skills.

Materials and Methods

Participants

Three datasets were included: the HCP S500, HCP NEW400, and Beijing Normal University (BNU) datasets.

HCP S500 Dataset

There are 520 subjects in the HCP S500 release. For all inclusion/exclusion criteria, please see the study by Van Essen et al. (2013). Thirteen subjects were excluded because 10 of them lacked T1 structural images and the other 3 subjects had a giant posterior cranial fossa arachnoid cyst. Finally, 507 subjects (205 males; 22–35 years) were included in our current study, and their HCP IDs were listed in Supplementary Table 1.

HCP NEW400 Dataset

The newly released HCP S900 dataset include all individuals of the S500. There are 435 new individuals that were not included in the S500, and 372 of them (180 males; 22–35 years) have T_1 images. The set of these 372 individuals was referred to as the NEW400 dataset, which was used to evaluate the generalizability of the model derived from the HCP S500 dataset (see below). The HCP IDs for the included individuals of the NEW400 datasets were listed in Supplementary Table 2.

BNU Dataset

A dataset of Chinese children that was collected in Beijing Normal University, referred to as the BNU dataset, was used to validate the generalizability of the model constructed using HCP S500 dataset. Seventy-two Chinese primary school children were included. Five children were excluded because of severe head motion during MRI scanning (visually checking the motion artifacts on the T_1-weighted images). Finally, 67 subjects (39 males; 8–13 years, mean age = 10.97 years) were analyzed, in which there were 25 dyslexics and 42 typically developing children. All participants were right-handed (Oldfield 1971) native Mandarin speakers who attended school regularly. Normal or corrected-to-normal vision and hearing were confirmed in each subject. The participants' parents reported no evidence of current or past major neurological or psychiatric disorders for any individual. All children had normal intelligence quotients, with scores above 85 on the Chinese version of the Wechsler Intelligence Scale for Children (C-WISC) (Gong and Cai 1993). Written informed consent was obtained from the children and their parents after the details of the study were comprehensively explained. The Institutional Review Board of the Beijing Normal University Imaging Center for Brain Research approved the protocol.

Behavioral Scores

HCP S500 and NEW400 Datasets

Two reading comprehension-related tests, that is, the Oral Reading Recognition Test (ORRT, measuring the reading decoding ability) and the Picture Vocabulary Test (PVT, measuring the linguistic comprehension ability), were chosen from the HCP cognitive battery. Specifically, the ORRT and PVT were applied using the NIH Toolbox Cognition Battery (Gershon et al. 2013). In the ORRT, letters or words are visually presented on the screen, and participants are required to pronounce them accurately. In the PVT, participants hear a spoken word while viewing 4 pictures, and they are asked to choose the picture that best represents the meaning of the word. According to the NIH Toolbox national norms, the raw scores of the 2 tests were transferred into the age-adjusted score, with mean of 100 and standard deviation of 15. In addition, the scores for fluid intelligence and emotion processing were included as control measurements. Specifically, fluid intelligence was tested with Raven's Progressive Matrices (Bilker et al. 2012), and emotion processing was evaluated using the Penn Emotion Recognition

Test (Gur et al. 2001). In S500 dataset, one male and one female lack the fluid intelligence and emotion processing scores. In NEW400 dataset, 4 females lack the fluid intelligence and emotion processing scores.

BNU Dataset

This dataset did not receive the same reading tests as the HCP dataset (i.e., ORRT and PVT). However, 2 related tests out of the behavioral battery were used, that is, the Character Recognition Test (CRT) and the Vocabulary Definition Test (VDT), which corresponded well to the ORRT and PVT, respectively.

CRT: This task consists of 150 single characters, which are required to be learned by the end of primary school (Shu et al. 2003). All characters were arranged in increasing difficulty level and decreasing frequency. Children were required to name the characters with no time limit and were stopped when they failed to recognize 15 consecutive items. The standard *Z* score was used here. The character recognition ability is a widely used indicator representing Chinese children's reading decoding skill (McBride-Chang and Kail 2002; Pan et al. 2011; Li et al. 2012).

VDT: This test was adopted from the C-WISC (Gong and Cai 1993). This subtest consists of 32 words, which are orally presented to the children. The task for the children is to provide the definition for each word. Scoring was based on the scoring scheme in the test manual. The full score for each item was 2. The children's answers were rated by 2 well-trained experimenters with high inter-rater reliability during pilot tests. Again, the standard *Z* score was used. This task has been suggested to be a reasonable proxy for linguistic knowledge in previous studies (McBride-Chang et al. 2005; Lervag and Aukrust 2010; Zhang et al. 2013).

MRI Acquisition

HCP S500 and NEW400 Datasets

High-resolution (0.7-mm isotropic voxels) structural T_1-weighted images were acquired using a customized Siemens Skyra 3-T scanner with a 32-channel head coil. The preprocessed images produced by the PreFreeSurfer pipeline were used. For details on data acquisition and preprocessing, see the study by Glasser et al. (2013).

BNU Dataset

All scans were performed using a 3-T Siemens Tim Trio MRI scanner in the Imaging Center for Brain Research, Beijing Normal University. Three-dimensional T_1-weighted images with high resolution were obtained using a 3D magnetization prepared rapid gradient echo (MPRAGE) sequence with the following parameters: slice thickness, 1.33 mm; no gap; 144 sagittal slices; repetition time, 2 530 ms; echo time, 3.39 ms; flip angle, 7°; acquisition matrix, 256 × 192; field of view, 256 × 192 mm^2; and resolution, 1 × 1 × 1.33 mm^3. An experienced radiologist reviewed all MR images to assess image quality and ensure the absence of visible neurological anomalies.

Image Processing

A GMV map in the Montreal Neurological Institute (MNI) space was generated for each individual using the SPM12 toolbox (http://www.fil.ion.ucl.ac.uk/spm/). This processing procedure included the following steps: 1) correcting for bias-field inhomogeneity; 2) segmenting into GM, white matter and cerebrospinal fluid density maps using the "new-segment" approach (Ashburner and Friston 2005); 3) applying Diffeomorphic Anatomical Registrations Through Exponentiated Lie Algebra (DARTEL) to generate a custom, study-specific template (Ashburner 2007); 4) warping each subject's GM density (GMD) image of the native space to the customized template; 5) affine registering the resultant image to the MNI space and standardizing the GMD map; 6) applying the modulation by multiplying the resulting GMD map with the nonlinear components of Jacobian determinant, which resulted in the GMV maps representing the local native-space GM volume after correcting for individual differences in whole-brain size; and 7) smoothing GMV maps using a 2-mm full-width at half-maximum Gaussian kernel. The moderate 2-mm smoothing kernel was chosen to make a balance between reserving anatomical details and compensating for registration errors (Gardumi et al., 2016). Finally, to create a GM mask, we smoothed GMD maps with 2-mm kernel size, averaged all subjects' resultant GMD maps, and applied a threshold of 0.2 to this average map (Krafnick et al. 2014; Xie et al. 2015). The GMV features were then restricted to this GM mask.

Notably, for a given prediction model, only subjects in the training set were used to estimate the DARTEL template and GM mask. The GMD image for each testing subject was then warped to this specific template, and the GMV features were extracted within this specific GM mask (see Supplementary Fig. 1). This ensures a complete isolation of the model training procedure from the testing individuals. The specific DARTEL template and GM mask based on all S500 subjects can be found at http://gonglab.bnu.edu.cn/wp-content/S500_All_DARTEL_Template_GMMask.rar.

Predictive Models

In the present study, we applied an elastic-net penalized linear regression model to predict the ORRT/PVT scores. We will briefly introduce the elastic-net algorithm and prediction framework as described below.

Elastic-Net Penalized Linear Regression

A linear regression model was adopted to predict individual's reading comprehension scores using the voxel-wise GMV features across the entire GM mask. The linear model can be formulized as follows:

$$y = \sum_{i=1}^{p} \beta_i x_i + \beta_0$$

where y is the reading comprehension score of the individual, p is the number of voxels in the GM mask, x_i is the GMV value at the voxel i and β_i is the regression coefficient.

To avoid overfitting and improve the prediction accuracy, penalization techniques have been frequently applied during model fitting (Zou and Hastie 2005). There are 2 common penalization techniques: 1) ridge regression and 2) least absolute shrinkage and selection operator (LASSO). Specifically, the ridge regression applies an L2-norm penalty, which minimizes the sum of the square of the regression coefficients and retains all features in the model (Hoerl and Kennard 1988). In contrast, LASSO applies an L1-norm penalty, which minimizes the sum of the absolute regression coefficients and retains only one representative predictor from the correlated predictors. Therefore, the LASSO will achieve a sparse model by excluding the majority of features from the model (Tibshirani 1996). The sparsity of the model is quite useful, as it can facilitate the optimization of the predictors and reduce the model complexity (Wright et al. 2010). Notably, the LASSO can only select N features at most in

the final model, where N is the sample size (Efron et al. 2004). This, however, can be problematic for a regression with few samples but large number of features, such as the present study (i.e., 510 samples but more than 180 000 features). To address this difficulty, elastic-net penalized linear regression was proposed, which uses a weighted combination of L1-norm and L2-norm penalties to allow for the number of the selected features to be larger than the sample size, while achieving a sparse model (Zou and Hastie 2005).

Specifically, the elastic-net penalty takes the following form:

$$\lambda \sum_{j=1}^{p} \left(\alpha \|\beta_j\|_{l_1} + \frac{1}{2}(1-\alpha)\|\beta_j\|_{l_2}^2 \right)$$

where β_j is the regression coefficient for the jth feature and α is a mixing parameter that controls the relative weighting of the L1-norm and L2-norm contributions. The elastic-net is equivalent to the ridge regression when $\alpha = 0$ and is equivalent to the LASSO when $\alpha = 1$. The regularization parameter λ controls the amount of shrinkage that was applied to β_j. If $\lambda = 0$, the effect of the elastic-net penalty is canceled. As λ increases from zero, the coefficients are progressively shrunk.

Prediction Framework Within S500 Dataset

The schematic overview for our prediction framework is shown in Figure 1 and Supplementary Figure 1. Specifically, for each subject, the GMV values of all GM voxels were extracted to generate a feature vector. Then, we applied a nested 3-fold CV (3 F-CV), with the outer 3 F-CV loop estimating the generalizability of the model and the inner 3 F-CV loop determining the optimal parameter set (α, λ) for the elastic-net model (Barretina et al. 2012). The outer 3 F-CV serves as the primary mechanism to prevent overfitting, together with the inner 3 F-CV for model selection. Here, we applied the scikit-learn library (version: 0.16.1) to implement the elastic-net penalized regression in the present study (http://scikit-learn.org/) (Pedregosa et al. 2011).

Outer 3 F-CV: In the outer 3 F-CV, all subjects were divided into 3 subsets. Particularly, we sorted the subjects according to their behavioral scores and then assigned the individuals with a rank of (1st, 4th, 7th, …, 508th) to the first, (2nd, 5th, 8th, …, 509th) to the second, and (3rd, 6th, 9th, …, 510th) to the third subset. This approach ensured the same distribution of behavioral scores for the 3 subsets and avoided random bias for the division. Of the 3 subsets, 2 were used as the training set, and the remaining one was used as the testing set. Each feature was linearly scaled to the range of zero to one across the

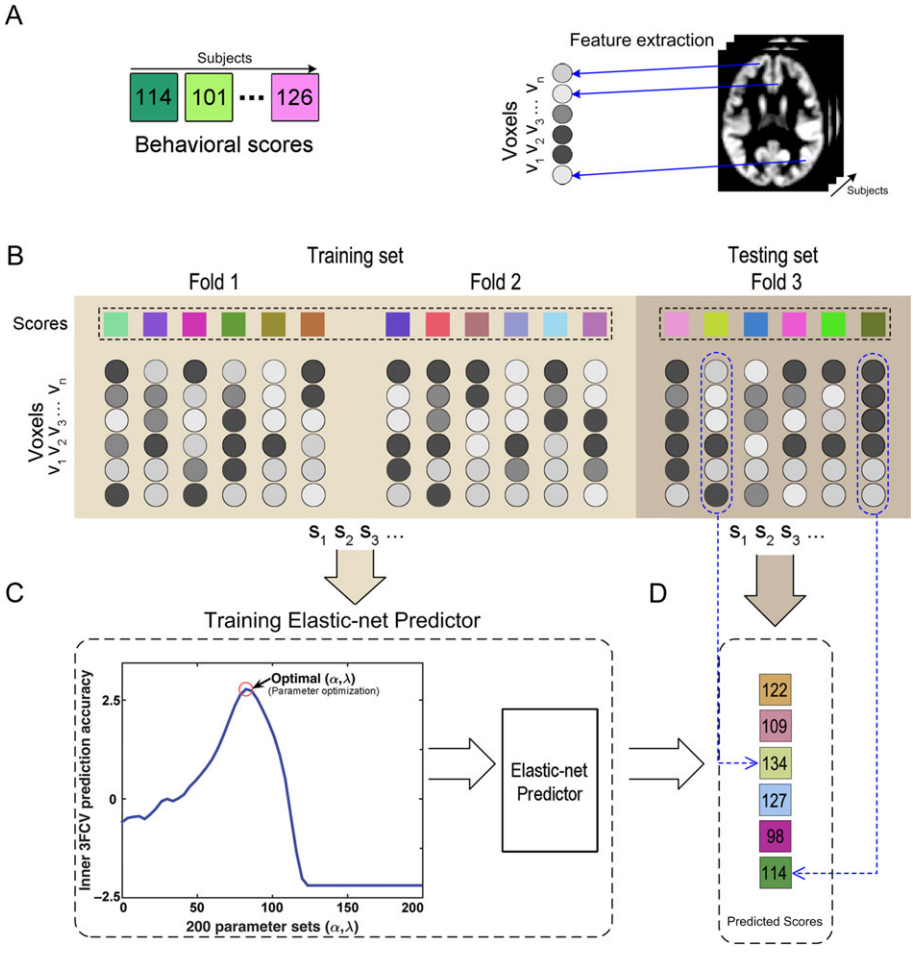

Figure 1. Schematic overview of the prediction framework. (*A*) Each subject provided a behavioral score and a whole-brain GMV map. All GMV values of voxels within the GM mask were extracted as the raw feature vector for each subject. (*B*) The whole dataset was partitioned into 3 subsets, 2 of which were used as the training sets and the remaining one was used as the testing set. (*C*) Inner 3-fold CVs were applied to determine the optimal parameter set (α, λ), and a prediction model was estimated by applying the elastic-net penalized regression to the training samples. (*D*) Applying the model to predict the behavioral scores for each testing subject. This figure was inspired by figure 1 reported in the study by Norman et al. (2006).

training dataset, and the scaling parameters were also applied to scale the testing dataset (Erus et al. 2015). The training and testing procedures were repeated 3 times such that each subset was used once as the testing set. Across the testing subjects for each fold, the Pearson correlation r and mean absolute error (MAE) between the actual reading comprehension scores and predicted scores were computed to quantify the accuracy of the prediction (Franke et al. 2010; Erus et al. 2015). The correlation r or MAE was averaged across the 3-folds to produce the final accuracy metrics.

Inner 3 F-CV: Within each loop of the outer 3 F-CV, we applied an inner 3 F-CV to determine the optimal α and λ. Specifically, the training set for each loop of the outer 3 F-CV was further partitioned into 3 subsets according to their rank of the behavioral scores, as like the outer loop. Two subsets were selected to train the model under a given parameter set of (α, λ), and the remaining subset was used to test the model. Regarding the (α, λ) choices, we applied a grid search: the α was chosen from 10 values in the range of [0.2, 1.0] and the λ was set as $\lambda = e^{\gamma}$, where γ was chosen from 20 values in the range of [−6, 5] (Barretina et al. 2012). This scheme resulted 200 (α, λ) parameter sets in total. For each set of (α, λ), one MAE and one correlation r were generated for each inner 3 F-CV loop, and a mean value across the 3 inner loops was then obtained for the MAE and correlation r, respectively. The sum of the mean correlation r and reciprocal of the mean MAE was defined as the inner prediction accuracy, and the (α, λ) set with the highest inner prediction accuracy across the 200 inner 3 F-CVs was chosen as the optimal parameter set. Notably, the mean correlation r and reciprocal of the mean MAE cannot be summed directly, because the scales of the raw values for these 2 measures are quite different. We therefore normalized the correlation r and MAE across the 200 samples, respectively, and then made the summation with the resultant normalized values.

Accordingly, each loop of the outer 3 F-CV ended up with a specific optimal parameter set (α, λ). The 2D images showing the normalized correlation r and MAE as a function of (α, λ) for each outer loop were included as Supplementary Figures 2 and 3. The optimal (α, λ) was then used to estimate the final elastic-net predictive model with the training set of that outer 3 F-CV loop.

The Python/Matlab scripts for our elastic-net analyses and the resultant PVT/ORRT models estimated by using the HCP S500 dataset have been made available online: https://github.com/ZaixuCui/CC_Reading_Prediction.

Significance of Prediction Performance

The permutation test was applied to determine whether the obtained final accuracy metrics (i.e., the mean correlation r or mean MAE of the 3 outer 3 F-CV loops) were significantly better than expected by chance (Dosenbach et al. 2010). Specifically, the above prediction procedure was re-applied 1 000 times. For each time, we permuted the behavioral scores across the training samples without replacement. The P value of the mean correlation r was calculated by dividing the number of permutations that showed a higher value than the actual value for the real sample by the total number of permutations (i.e., 1 000). Similarly, the P value of the mean MAE was the portion of permutations that showed a lower value than the actual value for the real sample.

Specificity of the Predictive Model

To assess the specificity of the predictive models for reading decoding and linguistic comprehension, we examined the correlation between the predicted ORRT or PVT scores and their actual scores, after adjusting for the fluid intelligence score. Putatively, the predictive model will capture specific GMV features that underlie individual differences for the 2 reading comprehension abilities rather than the general cognitive ability, if the resultant correlation remains significant; this strategy has been previously applied (Rosenberg et al. 2015). In addition, to further evaluate the model's specificity, we adopted the emotional processing scores as a control measurement, and we tested whether the predicted ORRT or PVT scores were significantly correlated with this control measurement.

Gender-Specific Predictive Models

For either the male or female group of S500 dataset, we applied the same nested 3 F-CV prediction framework to each group, separately. Specifically, we trained and evaluated a female-specific model using female subjects and a male-specific model using male subjects for the 2 reading comprehension scores: ORRT female-specific model, ORRT male-specific model, PVT female-specific model, and PVT male-specific model. To evaluate the gender effect on reading comprehension prediction, the female-specific and male-specific models using all subjects of each gender group were further applied to predict the scores for the subjects in the other gender group, respectively. For each gender group, the predicted scores between the 2 gender-specific models were compared using a paired t-test.

Independent Validation Using HCP NEW400 Dataset

To evaluate the generalizability of predictive models, we applied the acquired S500 models to predict the NEW400 subjects. Specifically, for both ORRT and PVT scores, we trained a predictive model using all S500 subjects (e.g., ORRT model and PVT model). The resultant ORRT model and PVT model were then applied to predict relevant scores for each of the NEW400 subjects (372 in total).

Like above, to validate the specificity of the acquired models, we correlated the predicted PVT/ORRT scores and their actual scores, after controlling for the fluid intelligence score. Also, we further tested whether the predicted PVT/ORRT scores were significantly correlated with the emotion processing scores.

Finally, for both ORRT and PVT, the acquired S500 male- or female-specific models were applied to predict the scores for NEW400 males or females, respectively.

Independent Validation Using BNU Dataset

To further evaluate the generalizability of the acquired S500 models to children and across centers, we applied the acquired S500 models to predict relevant reading comprehension scores in BNU dataset.

The ORRT model was applied to predict the CRT scores and the PVT model was applied to predict the VDT scores of the BNU children. The gender-specific models for PVT and ORRT of S500 dataset were applied to predict the CRT and VDT scores for the BNU boys and girls separately.

Contributing GM Voxels

The GM voxels with a nonzero regression coefficient/weight in the models trained using all S500 subjects (i.e., ORRT model and PVT model) can be deemed as the contributing voxels for the ORRT and PVT prediction, as reported previously

(Toiviainen et al. 2013; Khundrakpam et al. 2015). Similarly, gender-specific models for PVT and ORRT of S500 dataset were applied to locate the contributing GM voxels for gender-specific prediction.

Notably, the L1-norm penalization of elastic-net algorithm tends to select only a representative voxel from the correlated relevant voxels (Zou and Hastie 2005; Carroll et al. 2009; Grosenick et al. 2013), however, the correlated relevant voxels could be also informative and includable when mapping the behavior-related regions. To address this, for nonzero weighted voxels identified by the regularized elastic-net model, we searched out all possible voxels (with a zero weight according to the regularized elastic-net model) that correlate tightly ($r > 0.95$) with these voxels, and further assigned them the regression coefficient/weight of the correlated nonzero weighted voxel. Both these nonzero voxels and their tightly correlated voxels can be considered as important voxels that relate to reading comprehension abilities. The absolute regression coefficient/weight of a voxel represents the importance of the GMV feature to ORRT or PVT (Dosenbach et al. 2010; Erus et al. 2015; Cui et al. 2016).

Results

Overall Prediction Accuracy Within HCP S500

For each CV fold, there were 338 training subjects and 169 testing subjects. As shown in Figure 2*A* and *B*, the predicted ORRT/PVT scores were highly correlated with the actual scores for all the 3 folds. The mean correlations of the 3 folds were 0.40 and 0.43 for the ORRT and PVT, respectively. According to the permutation tests, these correlations were significantly higher than those expected by chance ($P < 0.001$). The mean MAEs for ORRT and PVT were 11.73 and 11.16, which were significantly lower than those expected by chance (permutation tests, $P < 0.001$).

Importantly, the predicted ORRT/PVT scores for all 3 folds remained significantly correlated with the actual scores (ORRT, mean $r = 0.30$; PVT, mean $r = 0.32$; permutation tests, both $P < 0.001$), even after controlling for the fluid intelligence score (Fig. 2*C*,*D*). Furthermore, the predicted ORRT/PVT scores showed no significant correlation with the emotion processing scores in most folds (ORRT: first fold, $r = -0.09$, $P = 0.27$; second fold, $r = 0.29$, $P < 0.001$; third fold, $r = 0.15$, $P = 0.06$; PVT: first fold, $r = 0.09$, $P = 0.26$; second fold $r = 0.18$, $P = 0.02$; third fold, $r = 0.04$, $P = 0.60$) (Fig. 2*E*,*F*). These results supported the cognitive specificity of the acquired predictive model to the ORRT/PVT.

Gender-Specific Prediction

Using male or female subjects, gender-specific predictive models were generated, and both gender-specific models could significantly predict the ORRT/PVT scores for subjects in the same gender. In females, the mean r and MAE of the female-specific model were 0.34 and 11.79 for the ORRT (Fig. 3*A*) and 0.43 and 10.37 for the PVT (Fig. 3*C*), respectively. In males, the mean r and MAE of the male-specific model were 0.34 and 11.84 for the ORRT (Fig. 3*B*) and 0.42 and 11.83 for the PVT (Fig. 3*D*), respectively. According to the permutation tests, all these mean correlation r were significantly higher than by chance (all $P < 0.001$), and all these MAEs were significantly lower than by chance (all $P < 0.001$).

Intriguingly, the males' predicted scores using the female-specific model were also significantly correlated with the actual scores of males (ORRT: $r = 0.40$; PVT: $r = 0.42$; permutation tests, both $P < 0.001$). (Fig. 3*F*,*H*). Notably, for both the ORRT and PVT, the males' predicted scores using the female-specific model were significantly higher than the predicted scores using the male-specific model (pair t-test, both $P < 0.001$) (Fig. 3*J*,*L*).

Likewise, the females' predicted scores using the male-specific model were significantly correlated with the actual scores of females (ORRT: $r = 0.32$; PVT: $r = 0.36$; permutation test, both $P < 0.001$) (Fig. 3*E*,*G*). For both the ORRT and PVT, the females' predicted scores using the male-specific model were significantly lower than the predicted scores using the female-specific model (pair t-test, both $P < 0.001$) (Fig. 3*I*,*K*).

Independent Prediction in NEW400 Dataset

The S500 models were applied to predict the ORRT and PVT scores for subjects in NEW400 dataset. The predicted ORRT and PVT scores were significantly correlated with the actual scores (ORRT: $r = 0.28$; PVT: $r = 0.34$; permutation tests, both $P < 0.001$) (Fig. 4*A*,*B*). The MAEs for ORRT and PVT were 11.58 and 11.07 (permutation tests, both $P < 0.001$). After controlling the fluid intelligence, the prediction accuracy remained significant for both ORRT ($r = 0.16$, permutation tests, $P = 0.001$) and PVT ($r = 0.21$, permutation tests, $P < 0.001$) (Fig. 4*C*,*D*). The predicted scores for both ORRT and PVT did not correlate with the emotion processing scores (ORRT: $r = 0.04$, $P = 0.419$; PVT: $r = 0.01$, $P = 0.853$) (Fig. 4*E*,*F*). These results further support the model generalizability and specificity to predict the ORRT/PVT.

The S500 male-specific model was used to predict the scores of NEW400 males. The correlation r between the predicted male's scores and the actual scores was 0.22 (permutation test, $P = 0.003$) for ORRT and 0.20 (permutation test, $P = 0.007$) for PVT (Fig. 5*C*,*D*). The corresponding MAEs for ORRT and PVT were 12.23 (permutation test, $P = 0.002$) and 12.16 (permutation test, $P = 0.76$). Also, the S500 female-specific model can significantly predict the scores of NEW400 females for both ORRT ($r = 0.28$, permutation test, $P = 0.003$; MAE = 11.69, permutation test, $P < 0.001$) and PVT ($r = 0.34$, MAE = 11.30, permutation tests, both $P < 0.001$) (Fig. 5*A*,*B*).

Independent Prediction in BNU Children

Given the ORRT-CRT and PVT-VDT correspondence, the S500 model was used to predict the CRT/VDT scores for the BNU children. The resultant predicted ORRT and PVT scores for BNU children were correlated with the actual CRT and VDT scores, respectively. Strikingly, the predicted ORRT scores were significantly correlated with the actual CRT scores across all children ($r = 0.24$, permutation test, $P = 0.024$) (Fig. 6*A*), and there is a trend for the correlation between the predicted PVT scores and actual VDT scores ($r = 0.20$, permutation test, $P = 0.062$) (Fig. 6*B*). Considering the significant differences between the HCP and BNU cohort (e.g., subjects' age range, MRI acquisition, and language type), these results provide further support for the generalizability of the acquired ORRT/PVT models. Notably, the scale of the predicted scores is based on ORRT/PVT scores in HCP dataset, therefore is quite different from the scale of actual CRT/VDT scores in BNU dataset (Fig. 6*A*,*B*).

The male-specific and female-specific models also exhibited decent generalizability. Specifically, the ORRT and PVT female-specific models of the S500 dataset could significantly predict the CRT ($r = 0.55$, permutation test, $P = 0.002$) and VDT ($r = 0.41$, permutation test, $P = 0.029$) scores for the BNU girls (Fig. 6*C*,*D*). The ORRT/PVT male-specific models also showed a trend for predicting the CRT ($r = 0.16$, permutation test, $P = 0.114$) and VDT ($r = 0.17$, permutation test, $P = 0.084$) scores for the BNU boys (Fig. 6*E*,*F*).

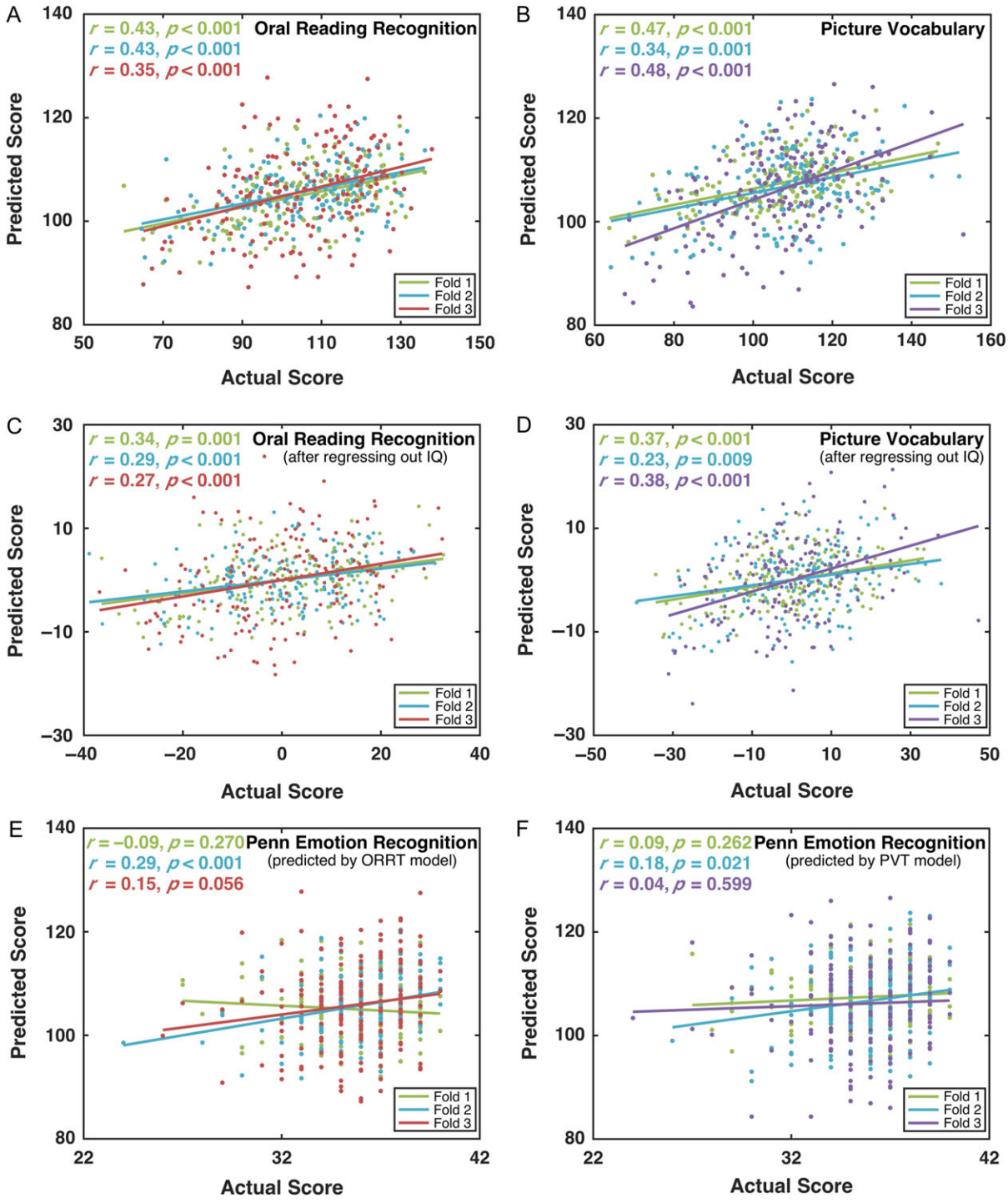

Figure 2. Prediction performance of the estimated ORRT and PVT models using HCP S500 dataset. The predicted scores and actual scores for both the ORRT (*A*) and PVT (*B*) models were significantly correlated across the testing samples in each fold of the CV. After regressing out the fluid intelligence scores, the correlations remained significant for both the ORRT (*C*) and PVT (*D*). In contrast, the correlation between the emotion processing scores and the predicted scores of the ORRT (*E*) or PVT (*F*) was not significant in most folds. The *P* values in (*A*–*D*) were calculated using permutation tests (i.e., 1 000 times).

Contributing GM Voxels

Both nonzero weighted voxels of the elastic-net model and their tightly correlated voxels were considered as important voxels/features that relate to the predicted reading comprehension abilities. Given the very scattered distribution of these voxels, we set a cluster-size threshold of 5 voxels, and sorted clusters according to the regression coefficient value of the cluster peak voxel. The most important 10 clusters for the prediction (i.e., the top 10 clusters) were illustrated for both ORRT and PVT (Fig. 7). The entire maps showing the spatial

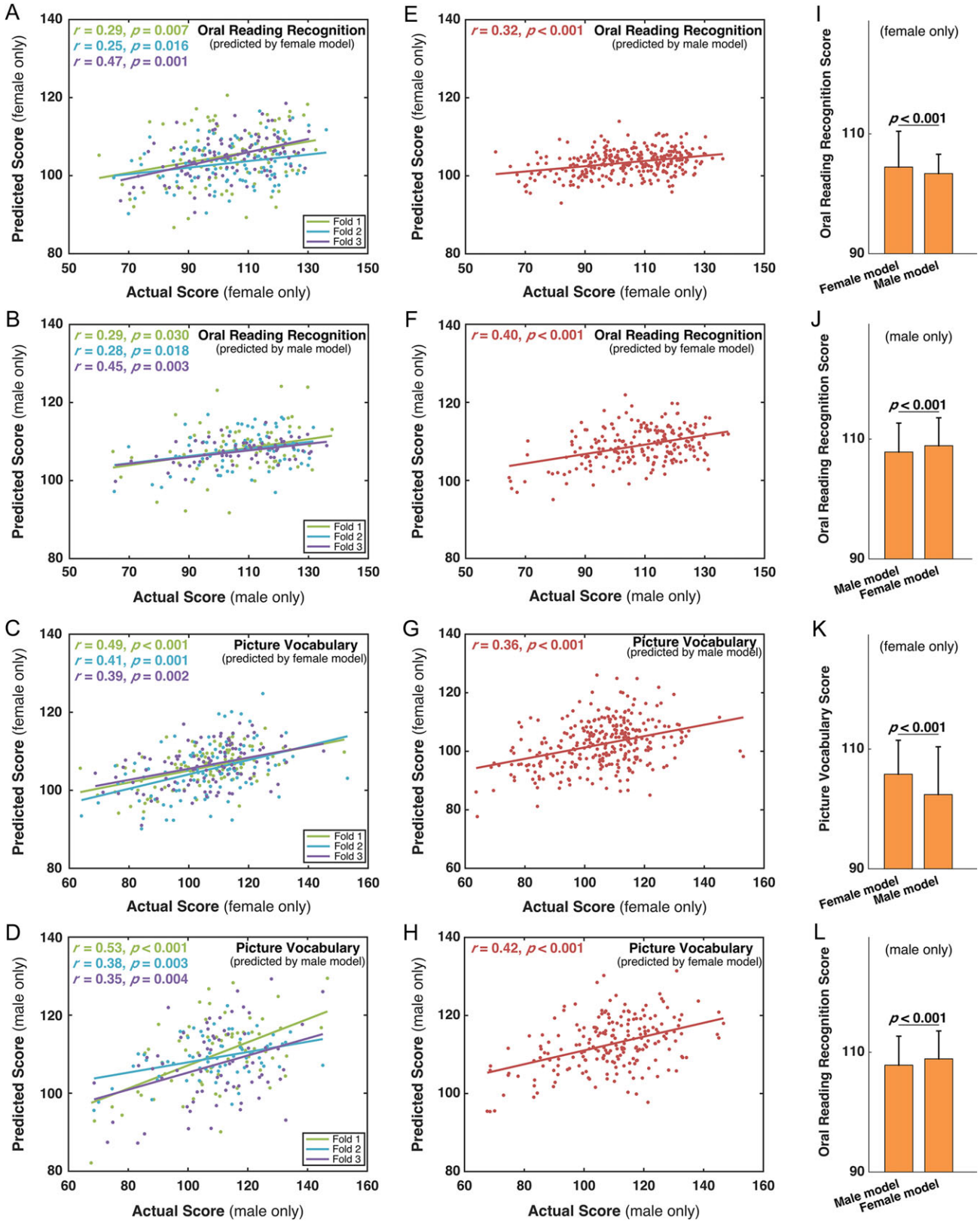

Figure 3. Prediction performance of gender-specific models for the ORRT and PVT using HCP S500 dataset. For each gender group, the predicted scores using the same gender-specific model and actual scores for both ORRT (*A*, *B*) and PVT (*C*, *D*) were significantly correlated across the testing samples in each fold of the CV. For each gender group, the correlations between the predicted scores using the other gender-specific model and the actual scores for both the ORRT (*E*, *F*) and PVT (*G*, *H*) were also significant. For females, the predicted PVT/ORRT scores using the male-specific model were significantly lower than those predicted by the female-specific model (*K*, *I*). For males, the predicted ORRT/PVT scores using the female-specific model were significantly higher than those predicted by the male-specific model (*J*, *L*). All the *P* values of correlations were calculated using the permutation tests (i.e., 1 000 times).

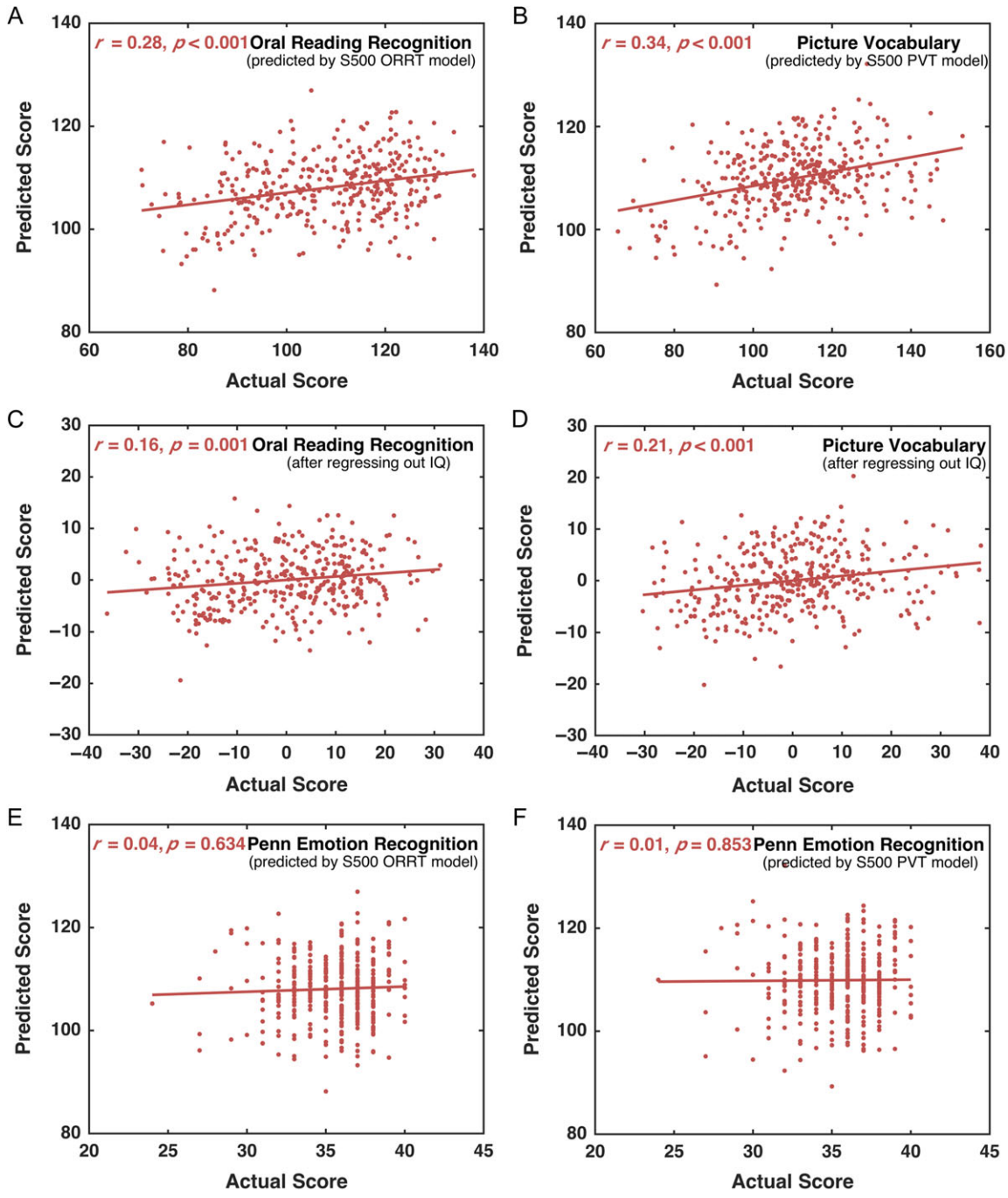

Figure 4. Prediction performance of the HCP S500-based ORRT and PVT models in the HCP NEW400 dataset. The scores predicted by the S500-based ORRT/PVT model and the actual scores were significantly correlated across the NEW400 samples (*A*, *B*). After regressing out the fluid intelligence scores, the correlations remained significant for both the ORRT (*C*) and PVT (*D*). In contrast, the correlation between the emotion processing scores and the predicted scores of the ORRT (*E*) or PVT (*F*) was not significant. The *P* values in (*A*–*D*) were calculated using permutation tests (i.e., 1 000 times).

pattern of all prediction-relevant GM voxels are illustrated in Supplementary Figure 4.

As shown in Figure 7 and Table 1, the most important GM voxels for predicting the ORRT involved widespread regions that mainly covered the putative reading system (e.g., lingual gyrus, inferior temporal gyrus, inferior frontal gyrus, and middle frontal gyrus), cerebellum (cerebellar tonsil), uncus, and postcentral gyrus. The most important GM voxels contributing to PVT prediction also involved multiple regions within the putative reading system (e.g., medial/middle frontal gyrus, inferior frontal gyrus, and superior temporal gyrus), the cerebellum (cerebellar tonsil), uncus, and parahippocampus (Fig. 7 and Table 2).

For the ORRT female-specific model, the identified most important GM voxels were mainly located around the inferior frontal gyrus, postcentral gyrus, superior temporal gyrus, cuneus, and posterior cingulate (Fig. 7 and Table 1). In contrast, the male-specific model mainly involved the uncus, medial

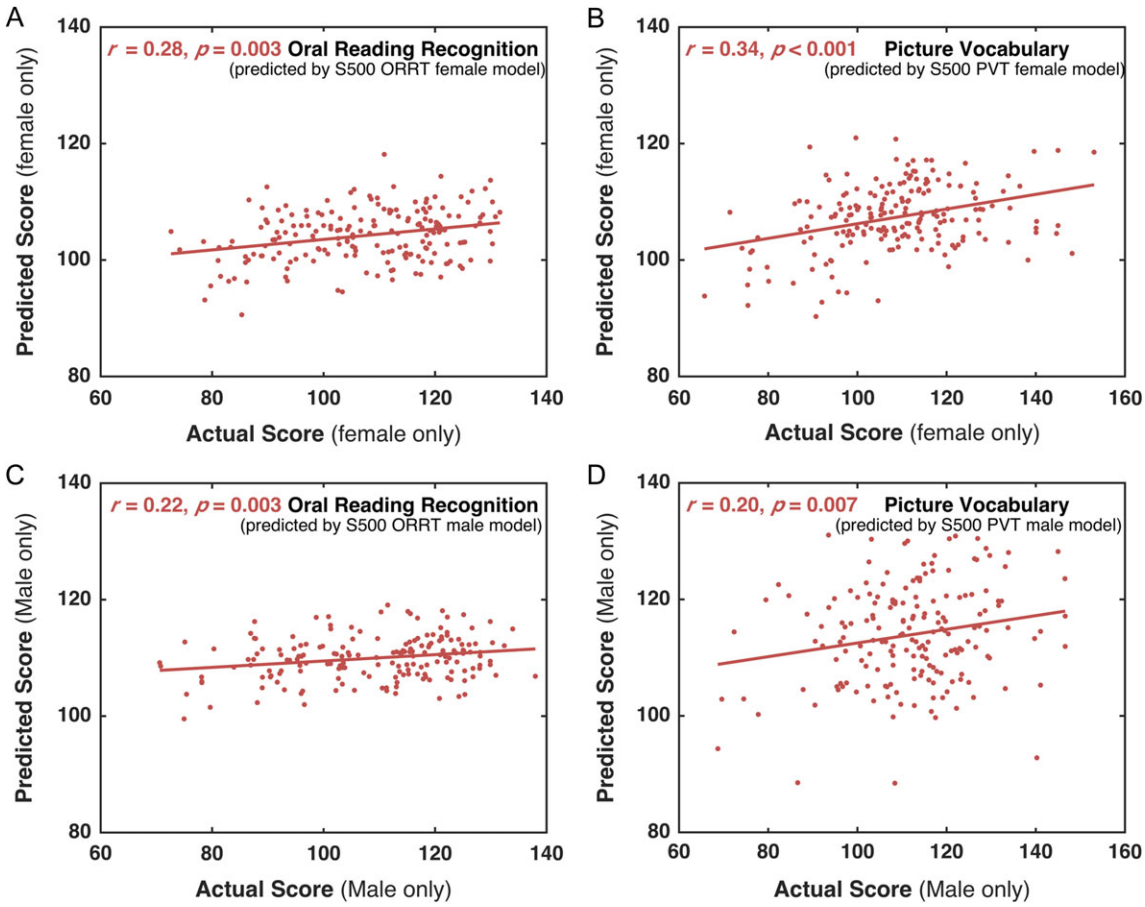

Figure 5. Prediction performance of the HCP S500-based gender-specific models in the NEW400 dataset. The scores of NEW400 females predicted by S500 female model were significantly correlated with their actual scores for both ORRT (*A*) and PVT (*B*). Similarly, S500 male model can significantly predict NEW400 males' ORRT (*C*) and PVT scores (*D*). All the *P* values were calculated using the permutation tests (i.e., 1 000 times).

frontal gyrus, thalamus, inferior parietal lobule, fusiform gyrus, and inferior temporal gyrus.

Regarding the PVT female-specific model, the most important GM contributing voxels were distributed around the middle frontal gyrus, lingual gyrus, middle temporal gyrus, and inferior frontal gyrus (Fig. 7 and Table 2). The male-specific model also covered widely distributed regions, including the middle frontal gyrus, superior frontal gyrus, cingulate gyrus, cerebellar tonsil, and lingual gyrus.

Discussion

Using a large cohort of healthy adults in the HCP dataset, the present study successfully built GMV feature-based multivariate models that efficiently captured individual differences in reading comprehension abilities (i.e., reading decoding and linguistic comprehension) and could significantly predict these abilities for unseen individuals. The CV using the HCP S500 cohort and another 2 independent testing datasets (i.e., NEW400 dataset and BNU cohort of children) demonstrated decent generalizability of these models. Particularly, there was a gender effect on the predictive models, with the female-specific model overestimating the males' reading comprehension abilities, while the male-specific model underestimated the females' abilities. Intriguingly, the GM regions contributing to the prediction exhibited considerable differences between the male-specific and female-specific models, suggesting distinguished neuroanatomical substrates for reading comprehension between males and females.

Individualized Prediction of the Reading Decoding and Linguistic Comprehension Abilities

To identify neuroimaging markers for capturing cognitive individual differences, it has been recently advocated to push the traditional correlational analysis across all samples to the individualized prediction that naturally evaluates whether the identified neuroimaging markers can be generalized and used in practice (Gabrieli et al. 2015; Dubois and Adolphs 2016). In line with this, the present study applied a strict 3-fold CV to assess the GMV-based prediction for unseen individuals, and significant prediction accuracies for the independent testing subjects in NEW400 and BNU datasets were achieved. Particularly, the estimated HCP S500 models were able to significantly predict the corresponding scores for completely independent individuals in the BNU dataset (collected in Beijing, China), further supporting the models' robustness and generalizability across sites. Notably, the subjects in the BNU dataset are Chinese children, and 25 of them were recognized as dyslexics. The HCP S500 model predicted that the dyslexic children would show relatively lower values for the score of character recognition, which is the main discriminative behavioral test for identifying Chinese dyslexics (Shu et al. 2003; Zhang et al. 2012; Cui et al. 2016). Therefore, the models possess potential

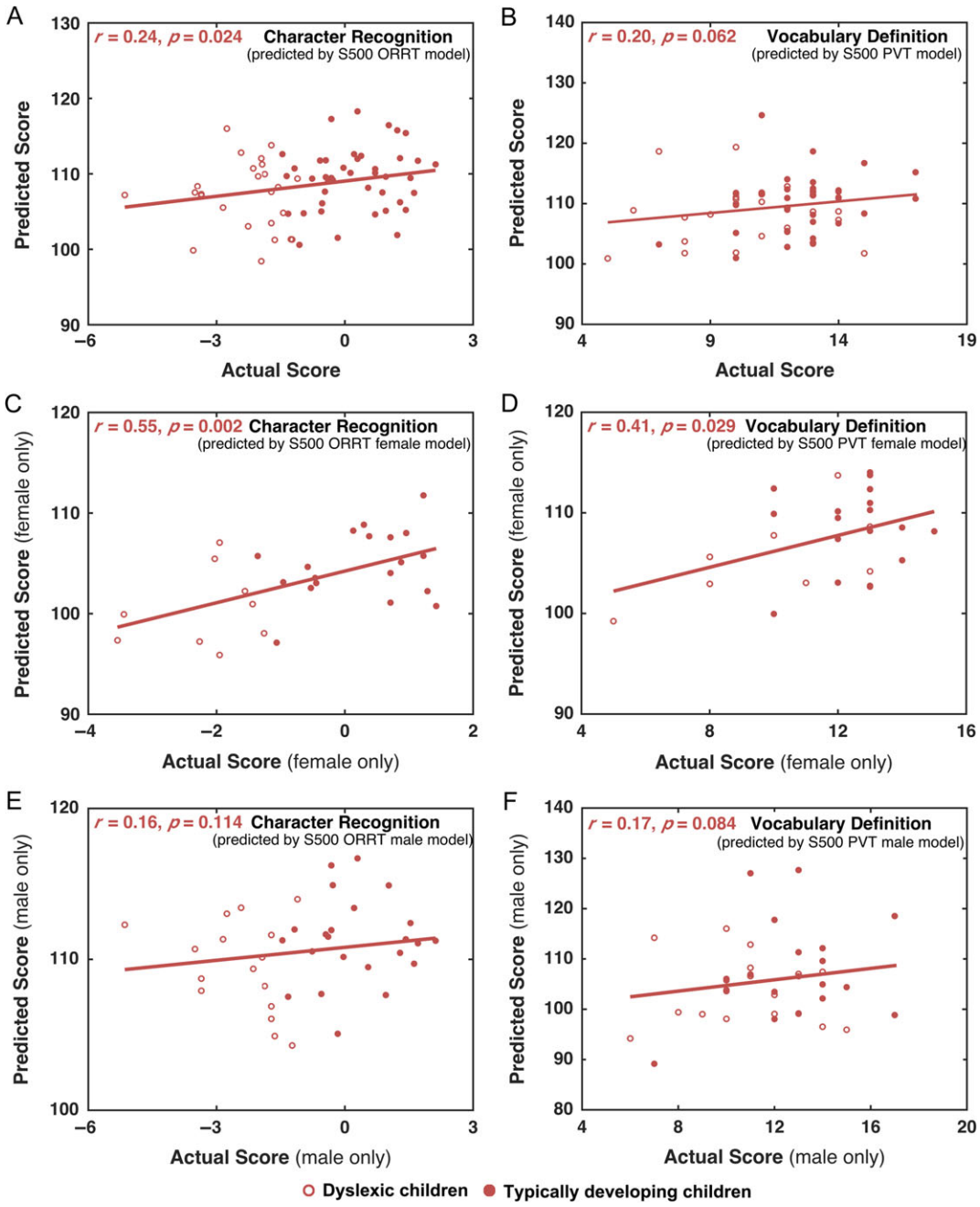

Figure 6. Prediction performance of the HCP S500-based ORRT and PVT models in the BNU dataset. The scores predicted by the HCP-based ORRT/PVT model and the actual CRT/VDT scores using the whole dataset (*A*, *B*), female group (*C*, *D*), and male group (*E*, *F*) were mostly significantly correlated. All the *P* values were calculated using the permutation tests (i.e., 1 000 times).

clinical significance to identify dyslexics. Finally, the HCP models were estimated using English-speaking adults but could be generalized to predict the reading comprehension abilities in Chinese-speaking children, strongly suggesting some common neuroanatomical substrate underlying these abilities among alphabetic and logographic languages.

It should be noted that a significant difference from chance indicates the meaningfulness of a prediction model, and an effect size (e.g., *r*) tells more about what extent the model can predict scores for unseen individuals and therefore is more suitable to assess generalizability. However, to determine a "good" generalizability, there is no definitive standard for a cutoff value in relevant effect size. Generally in statistics, a correlation *r* greater than 0.3 would be considered as a well-accepted effect size (Cohen 1992). While many of our predictions reached such a well-accepted effect size, some of them had an *r* value smaller than 0.3 though showing a significant difference from chance ($P < 0.05$). Moreover, in terms of whether the models can be solely applied in practice and replace relevant behavioral tests when evaluating individual's reading comprehension abilities or identifying dyslexic patients, the prediction results (even for the highest *r* value of 0.55 in Fig. 6*C*) are not good enough. The predicted data from our models however might join the data from behavioral tests,

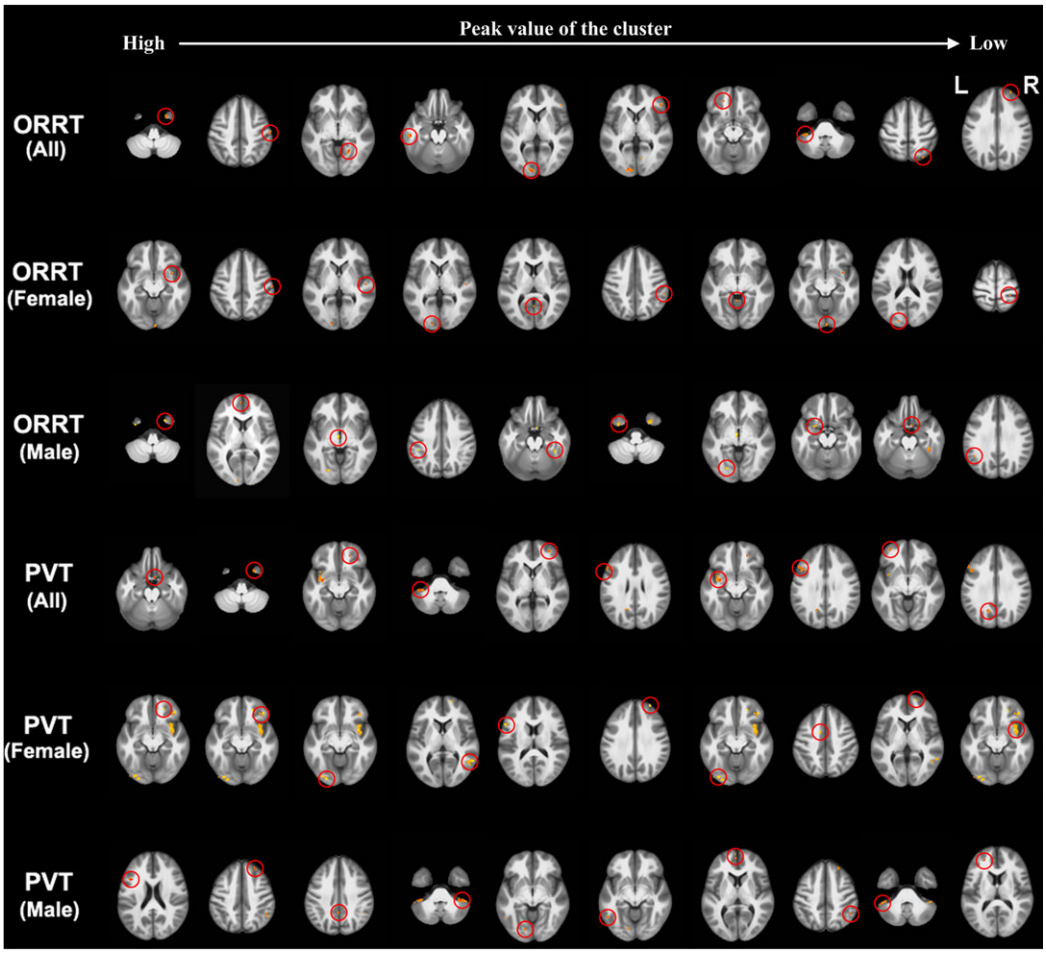

Figure 7. The most important contributing 10 clusters for the S500 PVT/ORRT prediction models using all subjects, female group, or male group. The clusters are arranged from the left to the right, according to a descending order of the peak value of the cluster. Each cluster was marked out by a red circle. The entire maps showing the spatial pattern of all prediction-relevant GM voxels are illustrated in Supplementary Figure 4.

resulting a more accurate assessment for individuals. Further investigation is highly desired to keep improving the accuracy and generalizability of our currently proposed predictive models.

Regarding neuroimaging features, the present study adopted the GMV, a neuroimaging measure characterizing the gross GM morphology. Notably, the GMV was extracted from the structural MRI data; this feature is very easy to acquire and requires no task in the scanner. Structural MRI-based features, such as the GMV, are advantageous for neuroimaging-based prediction or discrimination in practice. In fact, the GMV has been applied to predict individual age (Erus et al. 2015), working memory (Ullman et al. 2014), and clinical scores (Stonnington et al. 2010). Here, we further demonstrate the possibility of predicting reading-related abilities using the GMV feature, indicating a diagnostic role for the GMV in various reading-related disorders.

Computationally, our current results demonstrate the effectiveness of elastic-net penalized linear regression for prediction analysis with a large number of features. Compared with traditional linear regression, this algorithm applied the elastic-net regularization technique, which could prevent overfitting and improve the generalization ability by adding additional constraints or penalty to the model (Carroll et al. 2009; Teipel et al. 2015, 2016). This technique performs automatic feature selection while training the model, leading to a sparse predictive model, which is particularly attractive in cases of a large number of features and a small sample size. A few studies have applied the elastic-net regression to decode neural activity (Carroll et al. 2009; Grosenick et al. 2013) or predict behavioral phenotype (Fagerholm et al. 2015) and age (Khundrakpam et al. 2015). The present study further indicated the promising role of elastic-net regression in reading prediction and relevant biomarker discovery.

Neuroanatomical Substrates Underlying Reading Decoding and Linguistic Comprehension

Previous studies have consistently proposed that the human reading system consists of frontal, temporoparietal, and occipito-temporal cortical regions, responsible for mapping visual information onto auditory and semantic representations (Fiez and Petersen 1998; McCandliss and Noble 2003; Vandermosten et al. 2012). This putative reading system could be further divided into 2 distinct neural routes: dorsal phonological route (including superior temporal gyrus, inferior parietal lobule, and inferior frontal gyrus) and ventral orthographic route (including occipito-temporal regions) (Schlaggar and McCandliss 2007).

Table 1 The identified most important 10 clusters contributing the ORRT prediction in the S500 model using all, female, or male subjects, respectively

ID	Region	Hemisphere	Cluster size	Talairach			Weight
			Voxels	X	Y	Z	
Model based on all subjects of S500 dataset							
1	Uncus	R	34	−24	2	−52	0.330
2	Postcentral gyrus	R	21	−60	18	48	0.275
3	Lingual gyrus	R	45	−20	58	−4	0.239
4	Inferior temporal gyrus	L	21	52	24	−20	0.226
5	Cuneus	L	62	10	94	4	0.219
6	Inferior frontal gyrus	R	15	−46	−34	2	0.213
7	Middle frontal gyrus	L	10	18	−40	−14	0.201
8	Cerebellar tonsil	L	16	40	42	−42	0.199
9	Superior parietal lobule	R	7	−30	70	56	0.196
10	Superior frontal gyrus	R	9	−28	−56	30	0.193
Model based on female subjects of S500 dataset							
1	Inferior frontal gyrus	R	5	−34	−6	−12	1.700
2	Postcentral gyrus	R	6	−60	18	48	1.491
3	Superior temporal gyrus	R	6	−50	16	2	1.330
4	Cuneus	L	12	14	90	2	1.059
5	Posterior cingulate	R	5	−2	58	8	0.804
6	Inferior parietal lobule	R	8	−54	34	44	0.670
7	Culmen	L	10	2	48	−6	0.651
8	Lingual gyrus	R	14	−2	96	−12	0.645
9	Cuneus	L	5	16	88	18	0.623
10	Postcentral gyrus	R	5	−28	36	66	0.572
Model based on male subjects of S500 dataset							
1	Uncus	R	38	−24	2	−50	0.289
2	Medial frontal gyrus	L	5	2	−54	6	0.270
3	Thalamus	L	65	4	12	−4	0.241
4	Inferior parietal lobule	L	9	44	36	38	0.231
5	Fusiform gyrus	R	31	−40	42	−20	0.230
6	Inferior temporal gyrus	L	20	34	8	−48	0.198
7	Lingual gyrus	L	14	22	78	−4	0.196
8	Subcallosal gyrus	L	45	24	−4	−16	0.187
9	Medial frontal gyrus	R	6	−6	−6	−22	0.183
10	Supramarginal gyrus	L	7	46	52	30	0.180

In the present study, the prediction models identified a wide range of important GM regions that significantly contributed to the prediction of the 2 reading comprehension components; these regions mainly covered the putative reading system, cerebellum system, and subcortical system. This widespread distribution across various systems is compatible with the fact that reading comprehension is a complex skill consisting of multiple cognitive components, such as phonological processes, orthographic processes, and lexical-semantic processes (Pugh et al. 2000). As expected to some degree, the putative reading system was found to be prominent in predicting both reading decoding and linguistic comprehension. For example, some classical regions in reading comprehension exhibited a great deal of contribution for the prediction, including the superior temporal system and inferior frontal gyrus. This contribution implied an important role of phonological processing in the 2 reading comprehension skills (Pugh et al. 2000; Hoeft et al. 2011; Price 2012; Richlan et al. 2013).

On the other hand, the results showed that the left cerebellar tonsil was very important in predicting both the reading decoding and linguistic comprehension skills. This finding is compatible with previous functional MRI (fMRI) studies, in which cerebellar activation was repeatedly observed during reading-related tasks (Turkeltaub et al. 2002; Jobard et al. 2003). Our observed cerebellar contribution to the prediction provides further support for the cerebellum theory, in which impairments in cerebellar are believed to play an essential role in reading disability via affecting procedural learning, fluent processing and the acquisition of automatic processes (Nicolson et al. 2001; Nicolson and Fawcett 2007; Stoodley and Stein 2011).

Gender Differences in Predicting Reading Comprehension Abilities

The estimated female-specific model could be used to predict the scores of males, and the predicted scores were significantly correlated with the actual scores of males, and vice versa. This finding indicates some common mechanism between the female and male models. Despite the significant correlation, the female-specific model overestimated the males' scores, and the male-specific models underestimated the females' scores. This result is not that surprising, given the previously observed age-related female advantages in reading abilities (Wallentin 2009). Possibly, the female-specific model captures specific brain features on a more mature developmental stage and thus overestimated the males' performance. Genetic differences and different response styles to the environment of the 2 genders might also play a role in this gender difference.

Table 2 The identified most important 10 clusters contributing the PVT prediction in the S500 model using all, female, or male subjects, respectively

ID	Region	Hemisphere	Cluster size	Talairach			Weight
			Voxels	X	Y	Z	
Model based on all subjects of S500 dataset							
1	Medial frontal gyrus	R	26	−6	−6	−22	0.297
2	Uncus	R	16	−26	2	−52	0.258
3	Middle frontal gyrus	R	8	−20	−42	−12	0.256
4	Cerebellar tonsil	L	18	40	42	−42	0.253
5	Middle frontal gyrus	R	16	−32	−50	2	0.243
6	Inferior frontal gyrus	L	15	54	−20	26	0.234
7	Parahippocampus extend to superior temporal gyrus	L	114	32	4	−12	0.233
8	Middle frontal gyrus	L	80	52	−24	28	0.221
9	Middle frontal gyrus	L	10	34	−52	−6	0.200
10	Precuneus	L	22	16	60	30	0.197
Model based on all subjects of female S500 dataset							
1	Middle frontal gyrus	R	9	−20	−42	−12	0.313
2	Middle frontal gyrus	R	28	−42	−38	−12	0.241
3	Lingual gyrus	L	35	26	90	−10	0.238
4	Middle temporal gyrus	R	42	−58	52	6	0.236
5	Inferior frontal gyrus	L	68	54	−18	14	0.236
6	Superior frontal gyrus	R	13	−30	−54	32	0.229
7	Inferior occipital gyrus	L	28	34	86	−12	0.228
8	Cingulate gyrus	L	19	10	−2	46	0.220
9	Medial frontal gyrus	R	6	−16	−62	4	0.220
10	Inferior frontal gyrus	R	195	−34	−6	−12	0.215
Model based on all subjects of male S500 dataset							
1	Middle frontal gyrus	L	20	42	−16	20	0.645
2	Superior frontal gyrus	R	17	−26	−38	46	0.640
3	Cingulate gyrus	L	12	0	48	38	0.636
4	Cerebellar tonsil	R	39	−38	40	−42	0.592
5	Lingual gyrus	L	19	14	80	−6	0.578
6	Middle occipital gyrus	L	11	52	58	−8	0.559
7	Medial frontal gyrus	L	21	2	−54	6	0.558
8	Inferior parietal lobule	R	32	−48	50	48	0.556
9	Culmen	L	25	50	46	−40	0.551
10	Medial frontal gyrus	L	22	24	−44	12	0.548

In concordance, there were differences in the GM contributing spatial pattern between the female-specific and male-specific models. Although the GM spatial maps for both female-specific and male-specific models involved the putative reading system, cerebellum system, and subcortical system, the specific constitution within each system was distinct between genders, particularly within the putatively reading system. Specifically, the male-specific ORRT model for reading decoding involved the left inferior temporal gyrus, but it did not contribute to the female-specific model. According to the well-known dual route model of word reading (Coltheart and Rastle 1994), the left inferior temporal gyrus belongs to the ventral lexicosemantic route for reading (Jobard et al. 2003). The involvement of the ventral lexicosemantic route in males may reflect that males depend more on the semantic processing in the ORRT test, compared with females. However, the ORRT is a simple phonological decoding task for measuring reading decoding; thus, the involvement of semantic processing may be redundant for efficient reading decoding. Therefore, recruitment of the lexicosemantic route in males may impede this type of processing.

Regarding the linguistic comprehension ability, both male-specific and female-specific PVT models revealed an important role of the left lingual gyrus, suggesting the involvement of word processing in this particular cognitive processing (Howard et al. 1992; Price et al. 1994; Hoeft et al. 2007). However, the female-specific PVT model involved the left inferior frontal gyrus, but the male-specific model did not identify this region. Notably, the inferior frontal gyrus identified in the female-specific model previously revealed an association with articulation (Fiez and Petersen 1998) and was more recently implicated in the extraction of phonological elements (Gandour et al. 2002). Given the PVT performance requires the ability to extract useful phonological components and transform them into efficient articulation form (Fiez and Petersen 1998). It is possible that the involvement of the left inferior frontal gyrus in females is to enhance their performance via advantageous phonological processing skills, but males cannot benefit from this process.

Finally, a few issues relating to the current study should be addressed. First, the large HCP sample used to estimate the models included only adults with an age range from 22 to 35 years. While the model could be generalized to another independent sample of children, it would be intriguing to exclusively obtain models specific to different age ranges and further evaluate the predictive differences between them. In addition, the currently used HCP sample involves twins and siblings, which may limit the generalizability of our results to some degree. Second, the present study performed cross-sectional prediction, although longitudinally predictive models that

predict future reading decoding and linguistic comprehension outcomes using early brain imaging data are of great importance and should be thoroughly investigated in the future. Finally, further investigations are encouraged to achieve a better and more reliable prediction performance by combing the GMV and other neuroimaging features, for example, those derived from diffusion-weighted and fMRI data.

Supplementary Material

Supplementary material can be found at *Cerebral Cortex* online.

Funding

This work was supported by the National Science Foundation of China (81671772), the 973 program (2014CB846103, 2013CB837300), the National Science Foundation of China (81322021, 31271082, 31671126, 31611130107), the 863 program (2015AA020912), and the Fundamental Research Funds for the Central Universities. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 National Institutes of Health (NIH) Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Notes

Conflict of Interest: None declared.

References

Ashburner J. 2007. A fast diffeomorphic image registration algorithm. NeuroImage. 38:95–113.

Ashburner J, Friston KJ. 2005. Unified segmentation. NeuroImage. 26:839–851.

Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, et al. 2013. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage. 80:169–189.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 483: 603–607.

Bilker WB, Hansen JA, Brensinger CM, Richard J, Gur RE, Gur RC. 2012. Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. Assessment. 19: 354–369.

Burman DD, Bitan T, Booth JR. 2008. Sex differences in neural processing of language among children. Neuropsychologia. 46:1349–1362.

Carroll MK, Cecchi GA, Rish I, Garg R, Rao AR. 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage. 44:112–122.

Chiu MM, McBride-Chang C. 2006. Gender, context, and reading: a comparison of students in 43 countries. Sci Stud Read. 10: 331–362.

Cohen J. 1992. A power primer. Psychol Bull. 112:155–159.

Coltheart M, Rastle K. 1994. Serial processing in reading aloud: evidence for dual-route models of reading. J Exp Psych. 20: 1197.

Cui Z, Xia Z, Su M, Shu H, Gong G. 2016. Disrupted white matter connectivity underlying developmental dyslexia: a machine learning approach. Hum Brain Mapp. 37:1443–1458.

Dosenbach NUF, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, Nelson SM, Wig GS, Vogel AC, Lessov-Schlaggar CN. 2010. Prediction of individual brain maturity using fMRI. Science. 329:1358–1361.

Dubois J, Adolphs R. 2016. Building a science of individual differences from fMRI. Trends Cogn Sci. 20:425–443.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. Ann Stat. 32:407–451.

Erus G, Battapady H, Satterthwaite TD, Hakonarson H, Gur RE, Davatzikos C, Gur RC. 2015. Imaging patterns of brain development and their relationship to cognition. Cereb Cortex. 25:1676–1684.

Fagerholm ED, Hellyer PJ, Scott G, Leech R, Sharp DJ. 2015. Disconnection of network hubs and cognitive impairment after traumatic brain injury. Brain. 138:1696–1709.

Fiez JA, Petersen SE. 1998. Neuroimaging studies of word reading. Proc Natl Acad Sci USA. 95:914–921.

Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT. 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci. 18:1664–1671.

Franke K, Ziegler G, Klöppel S, Gaser C. 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. NeuroImage. 50:883–892.

Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S. 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron. 85:11–26.

Gandour J, Wong D, Lowe M, Dzemidzic M, Satthamnuwong N, Tong Y, Li X. 2002. A cross-linguistic FMRI study of spectral and temporal cues underlying phonological processing. J Cogn Neurosci. 14:1076–1087.

Gardumi A, Ivanov D, Hausfeld L, Valente G, Formisano E, Uludag K. 2016. The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. NeuroImage. 132:32–42.

Gershon RC, Slotkin J, Manly JJ, Blitz DL, Beaumont JL, Schnipke D, Wallner-Allen K, Golinkoff RM, Gleason JB, Hirsh-Pasek K, et al. 2013. IV. NIH Toolbox Cognition Battery (CB): measuring language (vocabulary comprehension and reading decoding). Monogr Soc Res Child Dev. 78:49–69.

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, et al. 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 80:105–124.

Gong Y, Cai T. 1993. Wechsler intelligence scale for children, Chinese revision (C-WISC). China: Map Press Hunan.

Gough PB, Tunmer WE. 1986. Decoding, reading, and reading disability. Rem Spec Educ. 7:6–10.

Grosenick L, Klingenberg B, Katovich K, Knutson B, Taylor JE. 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage. 72:304–321.

Gur RC, Ragland JD, Moberg PJ, Bilker WB, Kohler C, Siegel SJ, Gur RE. 2001. Computerized neurocognitive scanning: II. The profile of schizophrenia. Neuropsychopharmacology. 25:777–788.

He Q, Xue G, Chen C, Chen C, Lu ZL, Dong Q. 2013. Decoding the neuroanatomical basis of reading ability: a multivoxel morphometric study. J Neurosci. 33:12835–12843.

Hoeft F, McCandliss BD, Black JM, Gantman A, Zakerani N, Hulme C, Lyytinen H, Whitfield-Gabrieli S, Glover GH, Reiss AL, et al. 2011. Neural systems predicting long-term outcome in dyslexia. Proc Natl Acad Sci USA. 108:361–366.

Hoeft F, Meyler A, Hernandez A, Juel C, Taylor-Hill H, Martindale JL, McMillon G, Kolchugina G, Black JM, Faizi A,

et al. 2007. Functional and morphometric brain dissociation between dyslexia and reading ability. Proc Natl Acad Sci USA. 104:4234–4239.

Hoerl A, Kennard R. 1988. Ridge regression. In: Kotz S, Balakrishnan N, editors. Encyclopedia of statistical sciences. New York: Wiley. p. 129–136.

Hoover WA, Gough PB. 1990. The simple view of reading. Read Writ. 2:127–160.

Howard D, Patterson K, Wise R, Brown WD, Friston K, Weiller C, Frackowiak R. 1992. The cortical localization of the lexicons. Positron emission tomography evidence. Brain. 115(Pt 6): 1769–1782.

Jobard G, Crivello F, Tzourio-Mazoyer N. 2003. Evaluation of the dual route theory of reading: a metanalysis of 35 neuroimaging studies. Neuroimage. 20:693–712.

Khundrakpam BS, Tohka J, Evans AC, Brain Development Cooperative Group. 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. Neuroimage. 111:350–359.

Krafnick AJ, Flowers DL, Luetje MM, Napoliello EM, Eden GF. 2014. An investigation into the origin of anatomical differences in dyslexia. J Neurosci. 34:901–908.

Lervag A, Aukrust VG. 2010. Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. J Child Psychol Psychiatry. 51:612–620.

Li H, Shu H, McBride-Chang C, Liu HY, Peng H. 2012. Chinese children's character recognition: Visuo-orthographic, phonological processing and morphological skills. J Res Read. 35: 287–307.

Marquand AF, Rezek I, Buitelaar J, Beckmann CF. 2016. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biol Psychiatry. 80:552–561.

McBride-Chang C, Cho JR, Liu H, Wagner RK, Shu H, Zhou A, Cheuk CS, Muse A. 2005. Changing models across cultures: associations of phonological awareness and morphological structure awareness with vocabulary and word recognition in second graders from Beijing, Hong Kong, Korea, and the United States. J Exp Child Psychol. 92:140–160.

McBride-Chang C, Kail RV. 2002. Cross-cultural similarities in the predictors of reading acquisition. Child Dev. 73: 1392–1407.

McCandliss BD, Noble KG. 2003. The development of reading impairment: a cognitive neuroscience model. Ment Retard Dev Disabil Res Rev. 9:196–204.

Nicolson RI, Fawcett AJ. 2007. Procedural learning difficulties: reuniting the developmental disorders? Trends Neurosci. 30: 135–141.

Nicolson RI, Fawcett AJ, Dean P. 2001. Developmental dyslexia: the cerebellar deficit hypothesis. Trends Neurosci. 24:508–511.

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 10:424–430.

Oldfield RC. 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 9:97–113.

Pan J, McBride-Chang C, Shu H, Liu H, Zhang Y, Li H. 2011. What is in the naming? A 5-year longitudinal study of early rapid naming and phonological sensitivity in relation to subsequent reading skills in both native Chinese and English as a second language. J Educ Psychol. 103:897.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res. 12:2825–2830.

Price CJ. 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage. 62:816–847.

Price CJ, Wise RJ, Watson JD, Patterson K, Howard D, Frackowiak RS. 1994. Brain activity during reading. The effects of exposure duration and task. Brain. 117(Pt 6): 1255–1269.

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). Ment Retard Dev Disabil Res Rev. 6:207–213.

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: 3055–3065.

Rosenberg MD, Finn ES, Scheinost D, Papademetris X, Shen X, Constable RT, Chun MM. 2015. A neuromarker of sustained attention from whole-brain functional connectivity. Nat Neurosci. 19:165–171.

Rossell SL, Bullmore ET, Williams SC, David AS. 2002. Sex differences in functional brain activation during a lexical visual field task. Brain Lang. 80:97–105.

Schlaggar BL, McCandliss BD. 2007. Development of neural systems for reading. Annu Rev Neurosci. 37:475–503. 30.

Shaywitz BA, Shaywitz SE, Pugh KR, Constable RT, Skudlarski P, Fulbright RK, Bronen RA, Fletcher JM, Shankweiler DP, Katz L, et al. 1995. Sex differences in the functional organization of the brain for language. Nature. 373:607–609.

Shaywitz SE. 1998. Dyslexia. N Engl J Med. 338:307–312.

Shu H, Chen X, Anderson RC, Wu N, Xuan Y. 2003. Properties of school Chinese: implications for learning to read. Child Dev. 74:27–47.

Stonnington CM, Chu C, Kloppel S, Jack CR Jr., Ashburner J, Frackowiak RS, Alzheimer Disease Neuroimaging Initiative. 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage. 51:1405–1413.

Stoodley CJ, Stein JF. 2011. The cerebellum and dyslexia. Cortex. 47:101–116.

Teipel SJ, Grothe MJ, Metzger CD, Grimmer T, Sorg C, Ewers M, Franzmeier N, Meisenzahl E, Kloppel S, Borchardt V, et al. 2016. Robust detection of impaired resting state functional connectivity networks in Alzheimer's disease using elastic net regularized regression. Front Aging Neurosci. 8:318.

Teipel SJ, Kurth J, Krause B, Grothe MJ, Alzheimer's Disease Neuroimaging Initiative. 2015. The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment – Beyond classical regression. Neuroimage Clin. 8:583–593.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 58:267–288.

Toiviainen P, Alluri V, Brattico E, Wallentin M, Vuust P. 2013. Capturing the musical brain with Lasso: dynamic decoding of musical features from fMRI data. NeuroImage. 88C:170–180.

Turkeltaub PE, Eden GF, Jones KM, Zeffiro TA. 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. NeuroImage. 16:765–780.

Ullman H, Almeida R, Klingberg T. 2014. Structural maturation and brain activity predict future working memory capacity during childhood development. J Neurosci. 34:1592–1598.

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, WU-Minn HCP Consortium. 2013. The WU-Minn

Human Connectome Project: an overview. Neuroimage. 80: 62–79.

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TE, Bucholz R, Chang A, Chen L, Corbetta M, Curtiss SW, et al. 2012. The Human Connectome Project: a data acquisition perspective. Neuroimage. 62:2222–2231.

Vandermosten M, Boets B, Wouters J, Ghesquiere P. 2012. A qualitative and quantitative review of diffusion tensor imaging studies in reading and dyslexia. Neurosci Biobehav Rev. 36:1532–1552.

Wallentin M. 2009. Putative sex differences in verbal abilities and language cortex: a critical review. Brain Lang. 108: 175–183.

Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S. 2010. Sparse representation for computer vision and pattern recognition. Proc IEEE. 98:1031–1044.

Xie Y, Cui Z, Zhang Z, Sun Y, Sheng C, Li K, Gong G, Han Y, Jia J. 2015. Identification of amnestic mild cognitive impairment using multi-modal brain features: a combined structural MRI and Diffusion Tensor Imaging Study. J Alzheimers Dis. 47:509–522.

Zhang Y, Tardif T, Shu H, Li H, Liu H, McBride-Chang C, Liang W, Zhang Z. 2013. Phonological skills and vocabulary knowledge mediate socioeconomic status effects in predicting reading outcomes for Chinese children. Dev Psychol. 49:665–671.

Zhang Y, Zhang L, Shu H, Xi J, Wu H, Zhang Y, Li P. 2012. Universality of categorical perception deficit in developmental dyslexia: an investigation of Mandarin Chinese tones. J Child Psychol Psychiatry. 53:874–882.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 67: 301–320.