Behavioral Neuroscience
2007, Vol. 121, No. 3, 602–613

0735-7044/07/$12.00 DOI: 10.1037/0735-7044.121.3.602

Prediction of Children's Reading Skills Using Behavioral, Functional, and Structural Neuroimaging Measures

Fumiko Hoeft
Stanford University School of Medicine and Stanford University

Takefumi Ueno and Allan L. Reiss
Stanford University School of Medicine

Ann Meyler
Carnegie Mellon University

Susan Whitfield-Gabrieli
Stanford University

Gary H. Glover
Stanford University School of Medicine

Timothy A. Keller
Carnegie Mellon University

Nobuhisa Kobayashi, Paul Mazaika, and Booil Jo
Stanford University School of Medicine

Marcel Adam Just
Carnegie Mellon University

John D. E. Gabrieli
Stanford University

The ability to decode letters into language sounds is essential for reading success, and accurate identification of children at high risk for decoding impairment is critical for reducing the frequency and severity of reading impairment. We examined the utility of behavioral (standardized tests), and functional and structural neuroimaging measures taken with children at the beginning of a school year for predicting their decoding ability at the end of that school year. Specific patterns of brain activation during phonological processing and morphology, as revealed by voxel-based morphometry (VBM) of gray and white matter densities, predicted later decoding ability. Further, a model combining behavioral and neuroimaging measures predicted decoding outcome significantly better than either behavioral or neuroimaging models alone. Results were validated using cross-validation methods. These findings suggest that neuroimaging methods may be useful in enhancing the early identification of children at risk for poor decoding and reading skills.

Keywords: functional magnetic resonance imaging, voxel-based morphometry, reading skill, outcome, prediction

Supplemental data: http://dx.doi.org/10.1037/0735-7044.121.3.602.supp

Fumiko Hoeft, Center for Interdisciplinary Brain Sciences Research, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, and Department of Psychology, Stanford University; Takefumi Ueno, Department of Anesthesia, Stanford University School of Medicine; Allan L. Reiss, Center for Interdisciplinary Brain Sciences Research, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine; Ann Meyler, Timothy A. Keller, and Marcel Adam Just, Center for Cognitive Brain Imaging, Department of Psychology, Carnegie Mellon University; Susan Whitfield-Gabrieli and John D. E. Gabrieli, Department of Psychology, Stanford University; Gary H. Glover, Department of Radiology, Stanford University School of Medicine; Nobuhisa Kobayashi, Paul Mazaika, and Booil Jo, Center for Interdisciplinary Brain Sciences Research, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine.

Susan Whitfield-Gabrieli and John D. E. Gabrieli are now at the Harvard-MIT Division of Health Sciences and Technology and the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

This study was supported by grants from the William and Flora Hewlett Foundation and the Richard King Mellon Foundation. Participants were recruited and characterized by the Power4Kids program, which is funded through a public–private partnership that includes the Haan Foundation for Children, Heinz Endowments, Grable Foundation, and the U.S.Department of Education. For a full description of this project, see http://www.haan4kids.org/power4kids/. We declare no competing financial interests.

We thank Jun Sese, Tokyo University, for his support on statistical analyses and Jennifer Martindale, Heather Taylor-Hill, Glenn McMillon, Wai Ting Siok, and Gayle Deutsch for assistance in data collection and analyses.

Correspondence concerning this article should be addressed to Fumiko Hoeft, Center for Interdisciplinary Brain Sciences Research, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, 401 Quarry Road, M/C 5795, Stanford, CA 94305-5795. E-mail: fumiko@stanford.edu

The relation between education and cognitive neuroscience is exciting but controversial. It is exciting because noninvasive brain imaging methods are providing unprecedented views of the structural and functional development of the child's brain, and such new views of the maturing brain may provide novel information relevant for enhancing educational practices (Goswami, 2006). The relation between education and neuroscience is controversial because many links represent speculative and potentially flawed interpretations associating animal experimentation with human education (Bruer, 2002). For this reason, the relation between education and neuroscience has been called "a bridge too far" (Bruer, 1997, p. 4).

A critical issue in relating education and cognitive neuroscience is that education involves behavioral goals that are most directly evaluated by behavioral measures. For example, the most direct measures of the effectiveness of reading instruction are behavioral tests of reading comprehension or fluency or of the subskills of reading such as single-word decoding. Cognitive neuroscience studies of children aim to delineate the neural substrates of behaviors, such as reading. For example, a number of studies have found neural correlates of reading in typical-reading or dyslexic children (for reviews, see Dehaene, Cohen, Sigman, & Vinckier, 2005; Eden & Zeffiro, 1998; McCandliss, Cohen, & Dehaene, 2003; Price & Mechelli, 2005; S. E. Shaywitz & Shaywitz, 2005). These studies illuminate the neurobiological substrates of reading, but it is unknown whether such studies provide information that goes beyond behavioral measures with regard to reading itself. Perhaps brain imaging studies, at the optimal limit, can only be as informative about behavior as behavioral measures themselves. By this view, the imaging measures are redundant with behavioral measures, providing neurobiological correlates of behavior (i.e., providing a different level of description of the same phenomenon). Further, some behavioral measures, such as age-standardized language and reading tests, have been optimized for measurement reliability and validity, and measurement reliability and validity are seldom studied in brain imaging research. Finally, measures of a particular kind typically correlate best with measures of the same kind, so that behavioral measures of reading would be expected to be most closely associated with the most important outcome of reading education, that is, the behavior of reading. The preceding points suggest that current brain imaging measures are unlikely to provide insights into reading performance that go beyond behavioral measures. Alternatively, it may be the case that even now, neuroscience measures of brain structure and function contribute novel, nonredundant information about reading ability.

The goal of this study was to examine directly whether current brain imaging measures can provide novel information for predicting future reading skills in healthy children. We considered prediction to be an important goal because improved prediction of reading skill can facilitate identification of children who may benefit most from intensified or alternative reading instruction so that reading failure is minimized. We focused on one reading skill thought to be essential for effective reading, namely, word decoding skill. Decoding refers to the ability to determine the sound of a word from letters and syllables. Decoding ability is fundamental to reading because learning to read involves learning to relate the sounds of known auditory language (phonology) to letters (orthography). Early and systematic emphasis on decoding leads to superior achievement of reading skills (Adams, 1990; Snow, Burns, & Griffin, 1998). Therefore, improved methods for early identification of young children at risk for impaired decoding abilities hold promise for improving the specificity and effectiveness of early intervention and later achievement of reading skills.

A relatively pure test of word decoding involves reading aloud pronounceable nonsense words, because their proper pronunciation can only be derived from decoding skills (as opposed to words memorized by sight). Such a test also measures phonemic awareness, that is, awareness that words are composed of separable sounds (i.e., phonemes) that are blended to produce words. Phonemic awareness is one of the best predictors of reading success (e.g., Juel, 1988). We therefore used decoding skill as an outcome measure by measuring performance on a widely used test of decoding: the Woodcock Reading Mastery Tests (WRMT) Word Attack subtest. In this test, children attempt to read aloud pronounceable nonwords of successive difficulty.

With advances in neuroimaging, it is possible to examine brain activation and morphometric patterns that are associated with later reading achievement and decoding skills. To date, however, there are only minimal data pertaining to the use of brain measures to predict later reading achievement. All studies in this area have utilized event-related potentials (ERPs) to examine the development of language and reading skills (Espy, Molfese, Molfese, & Modglin, 2004; Molfese, Molfese, & Modgline, 2001). We used data from functional magnetic resonance imaging (fMRI) and structural imaging (VBM) and examined the relations of those measures to future reading skills. One imaging study has suggested that variation in brain morphology, as elucidated by VBM, can be linked to phonetic learning of novel speech sounds in normal-reading adults (Golestani, Paus, & Zatorre, 2002). Although not focused on reading per se, these results are of interest as they suggest that tissue-specific features of particular brain regions (parietal gray and white matter) can, in part, predict the speed or facility of normal, healthy adults in learning novel speech sounds.

In the present study, we investigated whether data obtained from fMRI and VBM can predict later decoding skills and whether fMRI and VBM data can be combined with behavioral data to produce a successful multimodal predictor of future reading skills. We studied 64 healthy children, identified by teachers as at risk for reading difficulty, who were between 8 and 12 years of age and varied in reading ability (see Method section for details on how the children were recruited and characterized). These children were identified as struggling readers by their teachers, but scores on standardized tests ranged widely from poor to average to above average.

We performed an fMRI study using a real-word rhyme judgment task interrogating phoneme awareness at the beginning of the school year (Time 1). At Time 1, optimized VBM analysis (C. D. Good et al., 2001) was also performed with high-resolution anatomical images. Further, a full battery of behavioral measures before (Time 1) and after one school year (Time 2) were obtained. We examined (a) how well decoding skills after one school year were predicted by initial fMRI and VBM results; (b) whether the combination of behavioral and neuroimaging results were more predictive than behavioral or neuroimaging results alone using multiple regression, and (c) the validity of the regression models. The model validity check is critical because the residual (or prediction) error of a multiple regression analysis may underestimate the errors found in practice when there are outliers in the data

or an excessive number of regressors in the model. We used leave-one-out validation analysis to demonstrate prediction ability and split-half reliability to demonstrate the stability of model estimation (see Supplemental Data online).

Method

Recruitment

All participants were children attending public schools surrounding Pittsburgh in Allegheny County, Pennsylvania and were recruited from a larger behavioral study of children in the Pittsburgh area. Although most children were within the normal range of reading, all students were initially identified as struggling readers by their teachers. These children were participants in the Power4Kids Reading Initiative, a randomized trial, field study of remedial instruction for children with a wide range of reading difficulties.[1] Parents received explanatory materials about the Power4Kids reading project in the mail, including the fMRI study, and those expressing interest in the fMRI study were recruited. The children gave verbal informed assent in the presence of a parent or guardian, who gave signed informed consent. The children were paid for their participation. A parent questionnaire was used to verify that all participants met inclusion criteria (e.g. right-handed, native English speakers, normal vision and hearing, no brain injury, sensory disorders, psychiatric disorders, attention deficit disorder, medication, claustrophobia, or metal in their bodies). Following recruitment and screening, the children were scanned and baseline measures were administered. All protocols were approved by the University of Pittsburgh and Carnegie Mellon University Institutional Review Boards, and informed assent and consent was obtained for participation from each child and guardian, respectively.

Participants

Children were healthy, right-handed, native English speakers between the ages of 8.2 and 12.4 years old. Out of 95 children tested at Time 1, 73 children returned for Time 2 behavioral assessment and 64 (37 females, 27 males) had complete and usable behavioral and neuroimaging measures. Many of the children underwent one of four different types of reading intervention, but there was no significant effect of intervention on their Time 1 or Time 2 standard scores of decoding (Time 1: $p = .92$, Time 2: $p = .44$).

Behavioral Evaluation

Reading ability was assessed witha standard battery of behavioral measures. Behavioral evaluations of reading and reading-related skills were obtained by Mathematica Policy Research (Princeton, NJ). Tests at Time 1 included WRMT Word Attack subtest; WRMT Word Identification subtest; WRMT Passage Comprehension subtest; AIMSweb (a test measuring reading fluency) Oral Reading Passage subtest; Clinical Evaluation of Language Fundamentals 3 (CELF) Formulated Sentences subtest; Comprehensive Test of Phonological Processing (CTOPP) Elision subtest; CTOPP Blending Word subtest; CTOPP Rapid Digit Naming subtest; CTOPP Rapid Letter Naming subtest; Group Reading Assessment and Diagnostic Evaluation (GRADE) Passage Comprehension subtest; Peabody Picture Vocabulary Test (PPVT); Rapid Autonomic Naming (RAN) Colors, Letters, Numbers, and Objects subtests; Test of Word Reading Efficiency (TOWRE) Phonemic Decoding Efficiency subtest; TOWRE Sight Word Efficiency subtest; the Woodcock Johnson (WJ) Spelling subtest; and the WJ Calculation subtest. Tests at Time 2 included WRMT Word Attack, Word Identification, and Passage Comprehension subtests, AIMSweb, GRADE Passage Comprehension subtest, WJ Spelling subtest, and WJ Calculation subtest and alternate forms. We compared differences between Time 1 and Time 2 behavioral scores and age using paired *t* tests.

fMRI Task Design

A real-word rhyme judgment task was used in the scanner with two conditions: rhyme and rest. During the rhyme condition, participants judged whether two visually presented words rhymed (e.g., *bait/gate, price/miss*) and indicated each response with a button press using their right hand for 'rhyme' and their left hand for 'non-rhyme'. Word pairs were selected so that the visual appearance of the last letters of the two words could not be regularly used to determine whether they rhymed. Stimuli were balanced for frequency of occurrence, number of letters, and syllables between the rhyme and nonrhyme trials and across blocks (Zeno, Ivens, Millard, & Duvvuri, 1995; see Hoeft et al., 2006, for the list of stimuli). Each trial lasted a total of 6 s, consisting of a 4-s period where the two words were presented simultaneously, followed by a 2-s fixation cross. Each task block consisted of a 2-s cue period followed by five trials (32 s total). During the rest block, participants saw a 15-s fixation cross on the screen. The entire scan was 234 s long, including two practice trials at the beginning, and consisted of four rhyme blocks and five rest blocks.

Image Acquisition

The fMRI imaging and imaging-related procedures were performed at the Brain Imaging Research Center (Carnegie Mellon University and University of Pittsburgh). A 3.0 tesla (T) Allegra scanner was used (Siemens Medical, Malvern, PA). A T2*-weighted gradient echo, resonant echo planar pulse sequence sensitive to blood oxygen level-dependent contrast was used with the following acquisition parameters: TR (repetition time) = 1,000 ms, TE (time to echo) = 30 ms, flip-angle = 60°, field of view (FOV) = 20 × 20 cm, matrix size = 64 × 64, axial-oblique plane with 16 slices, and a voxel size of 3.12 × 3.12 × 6 mm with a 1-mm gap. In addition, a T1-weighted 3D-MPRAGE with the following parameters was acquired: TR = 2,000, TE = 3.34, flip angle = 7°, dimensions = 256 × 256 × 160, axial plane, voxel size = 1 × 1 × 1 mm.

fMRI Data Analysis

Statistical analysis was performed with statistical parametric mapping software (SPM99; Wellcome Department of Cognitive Neurology, London, United Kingdom). After image reconstruction, each participant's data was slice-time corrected (ascending,

[1] See www.haan4kids.org/power4kids/ for details on the Power4 Kids Reading Initiative.

reference slice 8) and realigned to the first functional volume. Sessions were then normalized with the mean functional volume resampled to 2 × 2 × 2 mm voxels in Montreal Neurological Institute (MNI) stereotaxic space (12 nonlinear iterations, 7 × 8 × 7 nonlinear basis functions, medium regularization, sinc interpolation). Spatial smoothing was done with a Gaussian filter (8-mm full-width half maximum). Each participant's data, which was high-pass filtered at 96 s and globally scaled, was analyzed with a fixed effects model incorporating their 6 motion parameters (x, y, z, pitch, roll, yaw) as regressors. Motion was minimal in these children (Table 1). There were no significant differences between younger and older children: Grades 3 and 5, $t(62) = .05$, $p = .95$; for Grade 3, $n = 26$, $M = 0.23$; for Grade 5, $n = 38$, $M = 0.24$). Further, there were no significant correlation with age ($r = .14$, $p = .28$).

Group analysis was performed with a random effects model with the rhyme versus rest contrast images (one per participant, per contrast identified by fixed effect analysis). One-sample t tests were conducted to identify regions involved in phonological processing ($p = .01$, false-discovery rate corrected; extent threshold (et) = 10 voxels).

Further, we performed simple regression analysis using Time 2 WRMT Word Attack standard scores for age as a covariate of interest. We identified regions that showed significant positive or negative correlation with contrast values and Time 2 Word Attack standard scores (which we defined as regions of interest, or ROI_{fMRI}; $p = .001$, et = 10) and extracted contrast estimates for each participant for further analyses.

Voxel-Based Morphometry Data Analysis

Statistical analysis was performed with SPM2 (Wellcome Department of Cognitive Neurology, London, United Kingdom). After image reconstruction and coregistration with functional images, we used an optimized voxel-based statistical analysis (C. D. Good et al., 2001) with tools modified by Christian Gaser (http://dbm.neuro.uni-jena.de/vbm.html). Images were segmented into gray matter, white matter, and cerebrospinal fluid and normalized to a segmented template using the following parameters for nonlinear normalization: 25-mm cutoff, medium regularization, 16 iterations. Normalization parameters were applied to the initial anatomic volume, and the normalized anatomic images were partitioned into gray matter and white matter. Spatial smoothing was performed at full-width half maximum 12 mm. We performed analyses using both the standard adult template as well as the customized template including all participants and found similar results in terms of location and statistical significance. We also found similar results for modulated and nonmodulated VBM results. Here, we report the results using a customized template without modulation.

We performed multiple regression analysis of gray matter and white matter densities using Time 2 WRMT Word Attack –standard scores as a covariate of interest and total gray matter or white matter volume as a nuisance variable. We identified regions that showed significant positive or negative correlation with gray matter or white matter density and Time 2 WRMT Word Attack standard scores (ROI_{GM}, ROI_{WM}, respectively, where GM = gray matter and WM = white matter; $p = .000001$, family-wise error corrected, et = 0) and extracted the average density values for each ROI and for each participant for further analyses.

For both fMRI and VBM, statistical images were overlaid onto the SPM or medical image viewing software MRIcro (http://www.sph.sc.edu/comd/rorden/mricro.html) template image for three-dimensional viewing. Peak coordinates of brain regions with significant effects were converted from MNI to Talairach space with the mni2tal function (http://www.mrc-cbu.cam.ac.uk/Imaging/Common/mnispace.shtml). Brain regions were identified from these x, y, and z coordinates with Talairach Daemon (Research Imaging Center, University of Texas Health Science Center, San Antonio, TX) and confirmed with the Talairach atlas (Talairach & Tournoux, 1988).

Definition of Prediction Models

We performed prediction analyses using a method similar to other studies predicting outcome where there were a number of predicting variables (Poulakis et al., 2004; Woodhouse et al., 2003) (Figure 1a). All analyses were performed with Matlab (MathWorks, Natick, MA). First, simple regression analyses were performed between Time 2 WRMT Word Attack standard scores

Table 1
Demographics

	Time 1		Time 2	
Variable	Raw scores	Standard scores	Raw scores	Standard scores
Age (years)	10.00* (1.09)		10.55* (1.09)	
WMRT				
WA	19.50* (9.06)	93.73 (10.57)	25.28* (7.97)	98.01 (10.68)
ID	53.59* (14.28)	90.48 (9.92)	61.06* (11.51)	92.88 (8.84)
PC	31.39* (7.88)	93.95 (11.6)	34.17* (6.06)	94.67 (8.88)
Rhyme (%)	78.4 (15.9)		NA	
fMRI motion (mm)	0.24 (0.13)		NA	

Note. Values presented are means (and standard deviations); WRMT = Woodcock Reading Mastery Test; WA = Word Attack (pseudo-word reading); ID = Word Identification (real-word reading); PC = Passage Comprehension; Rhyme = in-scanner real-word rhyme judgment task performance; fMRI = functional magnetic resonance imaging; NA = not applicable.
* $p < .001$, comparing Time 1 and Time 2 performance.

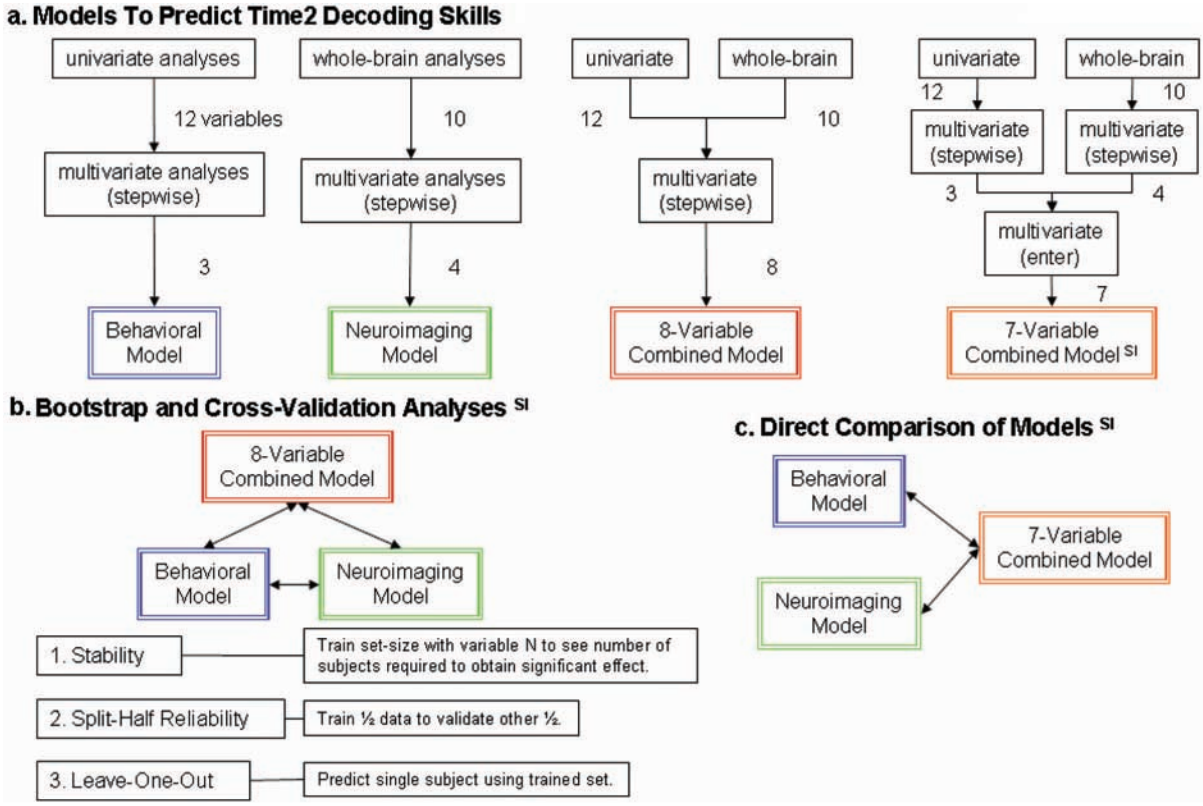

Figure 1. Flow charts and schematic diagrams of model creation and statistical comparisons of models. SI indicates that the corresponding methods and results are provided in the Supplemental information online.

and each of the Time 1 behavioral measures. Behavioral variables that correlated significantly ($p < .05$) were then subjected to a multiple regression analysis with Time 2 Word Attack standard scores.

It is possible that we missed some important predictors (suppressor variables) that on their own do not correlate with outcome but may reduce error variance by explaining additional variance. However, we first performed simple regression analyses to select variables that were entered into multiple regression analyses, in order to match the methods used to derive neuroimaging predictors and also to reduce the number of variables objectively.

In multiple regression analyses, y_i' are the Time 2 WRMT Word Attack standard scores ($i = 1, \ldots, N$, where N = total number of participants), and x_{ki} ($k = 1, \ldots, K$, where K = total number of behavioral variables) are the behavioral scores. The predicted Time2 WA-ss Y_i's were then denoted with weights b_k's for each participant i. Using the least square method (Minotani, 2004), we determined $b_1, \ldots b_k$, and constant term b_0, to minimize the sum of squared deviations and provide the best fit of the multiple regression model, the best correlation coefficient r^2 of the model, and the best contribution R^2 for each variable. The regression residual is represented by ε_i.

$$\sum_{i=1}^{N} (Y_i - Y_i')^2$$

$$Y_i' = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i$$

First, we performed multiple regression with all behavioral variables defined in the simple regression analyses using the enter procedure. Next, using the stepwise procedure (criteria: probability-of-F-to-enter $\leq .05$, probability-of-F-to-remove $\geq .1$, which are the default settings in Matlab), we obtained the behavioral model. We also performed forward and backward procedures and obtained similar results not only for the behavioral model but also for the neuroimaging and combined models. In the multiple regression model obtained from the stepwise procedure, behavioral measures that contributed significantly according to the above criteria were defined as *behavioral predictors*. Y_i' was defined as the prediction index PI_i for a given participant i.

Prediction indices were plotted against Time2 WRMT Word Attack standard scores, and r^2 and p values were computed. Linear regression lines and two types of 95% prediction intervals (Minotani, 2004) were drawn. Prediction intervals were calculated as follows. The predicted value from the regression line $\hat{Y}_i$ can be defined as:

$$\hat{Y}_i = b_0 + b_1x_1 + b_2x_2 + \ldots$$

where b_0, b_1, b_2, and so forth are the results of the regression model fit. The residue ε_i is therefore:

$$\varepsilon_i = Y_i - \hat{Y}.$$

Assuming Gaussian data, the prediction interval with 95% confidence of the mean Time 2 WRMT Word Attack standard scores was calculated as:

$$\hat{Y}_0 \pm 1.96 \times \sqrt{\frac{\sum \varepsilon_i^2}{N-2}\left(\frac{1}{N}+\frac{(PI_o-\overline{PI})^2}{\sum_i (PI_i-PI)^2}\right)}$$

for the ranges of the prediction indices where there are 95% probabilities that the next experimental group line regression will occur (95% prediction interval, group), and

$$\hat{Y}_0 \pm 1.96 \times \sqrt{\frac{\sum \varepsilon i^2}{N-2}\left(1+\frac{1}{N}+\frac{(PI_o-\overline{PI})^2}{\sum_i (PI_i-PI)^2}\right)}$$

for the ranges of the prediction indices where there are 95% probabilities that the next experiment's individual's Time 2 WRMT Word Attack standard score value will occur (95% prediction interval, individual; Minotani, 2004), where PI = prediction index, $\hat{Y}_i = \alpha_0 + \alpha_1 PI$, N = number of participants, PI_0 = a new given PI with which we predicted the confidence interval, PI_i's = the original PIs (data itself), and $\overline{PI}$ = mean value of PI_i's. We defined the neuroimaging and combined (behavioral and neuroimaging) models similarly and subsequently compared different models.

For neuroimaging data, extracted contrast values from the fMRI analysis (ROI_{fmri}), and gray matter and white matter density values from VBM analyses (ROI_{GM}, ROI_{WM}, respectively) were submitted to similar multiple regression analyses used to identify behavioral predictors and were defined as neuroimaging predictors. Prediction indices were calculated similarly and correlation with prediction indices and Time 2 WRMT Word Attack standard scores were examined (Figure 1a). This was defined as the neuroimaging model.

Additionally, the 12 behavioral variables and 10 neuroimaging predictors that showed correlation with Time 2 WRMT Word Attack standard scores were combined and submitted to similar analyses to identify combined predictors, which yielded eight predictors. Prediction indices were calculated similarly, and correlation with prediction indices and Time 2 Word Attack standard scores were examined (Figure 1a). This was defined as the eight-variable combined model

There was some variation between the number of days between Time 1 and Time 2 testing sessions, but this interval did not correlate with Time 2 WRMT Word Attack standard scores ($r = -.06$, $p = .66$). Therefore, time between testing sessions was not considered in further analyses.

Prediction Analyses Controlling for Initial Decoding Skills, Age, or PPVT Standard Scores

We performed separate partial correlation analyses for each model using Time1 WRMT Word Attack –standard scores as covariates of no interest to examine whether results reflected only strong associations between Time 1 Word Attack –standard scores and the behavioral or neuroimaging measures, rather than a unique contribution of Time 2 Word Attack standard scores. We also partialed out Time 1 age and PPVT standard scores to avoid bias from these variables. We chose PPVT in place of an IQ measure because IQ was not obtained in this study; PPVT highly correlates with full-scale IQ (.90) in children on the *Wechsler Intelligence Scale for Children, Third Edition* (WISC—III; Dunn & Dunn, 1997).

Leave-One-Out Cross-Validation Method

In the leave-one-out validation analysis, we tested whether single participant Time 2 WRMT Word Attack standard scores were predicted from the remaining 63 participants in the behavioral, neuroimaging, or eight-variable combined models (Figure 1b). We first performed a multiple regression analysis with 63 participants, leaving out the single participant to be tested. The 63 participants were resampled 64 times, giving the best fitted b_i's for each sample. The b_i's were then applied to the omitted participant, yielding a prediction index. The 64 predicted values were plotted against Time 2 WRMT Word Attack –standard scores. Mean prediction indices of the 64 trained sets (i.e.,predicted value) were correlated with WRMT Word Attack Time 2 standard scores, and a linear regression line and 95% prediction intervals of individual expected Time 2 Word Attack standard scores (see above *Definition of Prediction Models* for definition) were drawn for each model. Absolute differences between the predicted values and the actual Time 2 Word Attack standard score values of the omitted participant were calculated. One-way repeated measures analysis of variance (ANOVA) and post hoc comparisons were performed between models.

Results

Demographic and Behavioral Measures

Demographic information is presented in Table 1. WRMT Word Attack standard scores were the critical outcome measure, and scores ranged from well above (138) to well below (66) the expected mean of 100. Time 1 and Time 2 Word Attack standard scores correlated highly ($r = .68$, $p < .001$).

Behavioral Model: Predicting Later Decoding Skills Using Behavioral Measures

We first performed simple correlations between each test administered at Time 1 and Time 2 WRMT Word Attack standard scores (Figure 1a) and found measures most closely related to phonological processing to correlate with Time 2 Word Attack standard scores: standard scores on Time 1 WRMT Word Attack ($r = .74$, $p < .001$), WRMT Word Identification ($r = .64$, $p < .001$), TOWRE Phonemic Decoding Efficiency ($r = .57$, $p < .001$), TOWRE Sight Word Efficiency ($r = .46$, $p < .001$), CTOPP Blending Word ($r = .35$, $p = .005$), CTOPP Elision ($r = .34$, $p = .006$), AIMSweb ($r = .50$, $p < .001$), WRMT Passage Comprehension ($r = .50$, $p < .001$), GRADE Passage Comprehension ($r = .40$, $p = .001$), Woodcock Johnson Calculation ($r = .41$, $p = .001$), Woodcock Johnson Spelling ($r = .60$, $p < .001$), and the in-scanner rhyme judgment task (Rhyme; $r = .48$, $p < .001$).

Among these 12 Time 1 behavioral measures that correlated significantly with Time 2 WRMT Word Attack standard scores, 3 variables remained as significant predictors when multiple regression analysis was performed, multiple $r^2 = .65$, $F(3, 60) = 36.59$, $p < .001$. These 3 remaining variables were defined as behavioral predictors, WRMT Word Attack, $t = 5.70$, $p < .001$; Woodcock

Johnson Spelling, $t = 3.05$, $p = .003$; Woodcock Johnson Calculation, $t = 2.53$, $p = .014$, and combined into prediction indices calculated by summing the constant and multiplying the 3 variables with their respective coefficients (see Supplemental Figure A1 online). Thus, this behavioral model was predictive of later decoding skills.

Neuroimaging Model: Predicting Later Decoding Skills Using Functional Magnetic Resonance Imaging and Voxel-Based Morphometry Measures

We compared fMRI activation for real-word rhyme judgments versus rest state (Figure 2, Table 2). We then performed whole-brain regression analyses correlating Time 1 fMRI activation and gray matter or white matter VBM densities with Time 2 WRMT Word Attack standard scores and found 10 brain regions that showed significant positive or negative correlations (ROI_{fMRI}s, ROI_{GM}s, and ROI_{WM}s, respectively; Figure 3 and Table 3). Mean contrasts (effect size calculated as the linear combination of beta parameters) or density information were extracted from these ROIs for each participant. Consistency maps from permutation analyses of fMRI and VBM multiple regression analyses showed consistent activation and morphometric patterns for these ROIs (see Supplemental Figure B and text online). Using multiple regression, four ROIs were found to contribute significantly, which were defined as neuroimaging predictors: multiple $r^2 = .57$, $F(4, 59) = 19.41$, $p < .001$; ROI_{fMRI} right fusiform ~ middle occipital gyri (RFG/MOG), $t = 4.30$, $p < .001$; ROI_{fMRI} left middle temporal gyrus (LMTG): $t = 3.21$, $p = .002$, ROI_{fMRI} right middle frontal gyrus (RMFG): $t = -2.36$, $p = .021$; ROI_{GM} right posterior fusiform gyrus (RFGp): $t = 3.90$, $p < .001$; Figure 3 and Table 3). Prediction indices were calculated from these four predictors as described above (Supplemental Figure A2 online). Thus, the neuroimaging model was also predictive of later decoding skills.

Combined Model: Predicting Later Decoding Skills Combining Behavioral and Neuroimaging Measures

We repeated multiple regression and prediction analyses using the 12 behavioral and 10 neuroimaging variables that showed significant correlation with Time2 WRMT Word Attack standard scores (Figure 1a). The goal was to achieve an index that would best predict Time 2 Word Attack standard scores while minimizing the total number of predictors. Using multiple regression, we found that 8 variables contributed significantly, which were then defined as combined predictors: multiple $r^2 = .81$, $F(8, 55) = 30.18$, $p < .001$; standard scores on Word Attack, $t = 4.16$, $p < .001$; Woodcock Johnson Calculation, $t = 2.94$, $p = .005$; Woodcock Johnson Spelling, $t = 2.79$, $p = .007$; ROI_{fMRI} RFG/MOG, $t = 3.95$, $p < .001$; ROI_{GM} RFGp, $t = 2.74$, $p = .008$; ROI_{GM} right anterior frontal gyrus (RFGa), $t = 2.94$, $p = .005$; ROI_{WM} left inferior parietal lobule (LIPL), $t = 2.26$, $p = .028$; ROI_{WM} left superior temporal lobe (LSTL), $t = 2.14$, $p = .037$; Figure 3 and Table 3) Prediction indices were calculated as described above (see Supplemental Figure A3 online). Thus, the combined model was also predictive of Time 2 Word Attack standard scores.

When behavioral predictors were entered first, neuroimaging predictors explained 23% more variance in addition to the variance explained by behavioral predictors, $F(4, 56) = 6.42$, $p < .001$. When neuroimaging predictors were entered first, behavioral predictors explained 15% of the variance in addition to the variance explained by neuroimaging predictors, $F(3, 56) = 14.78$, $p < .001$. Thus, the combined model was significantly better than the behavioral model or the neuroimaging model (see Supplemental Figure A3 and text online).

Prediction Analyses Controlling for Initial Decoding Skills, Age, or PPVT Scores

There is a possibility that the results reported thus far may merely be a result of strong associations between Time 1 WRMT Word Attack standard scores and the behavioral scores or prediction indices, rather than a unique contribution of Time 2 Word Attack standard scores. Therefore, we performed partial correlation analyses for the three models, partialing out Time 1 Word Attack standard scores, and found that the results remained significant (behavioral $r^2 = .21$, neuroimaging $r^2 = .40$, combined $r^2 = .59$; all ps $< .001$; details are provided in the Supplemental text online).

We also partialed out Time 1 age and PPVT standard scores. The results remained significant: for age, behavioral $r^2 = .63$, neuroimaging $r^2 = .56$, and combined $r^2 = .81$; for PPVT, behavioral $r^2 = .63$, neuroimaging $r^2 = .57$, and combined $r^2 = .81$; all p's $< .001$). Hierarchical regression analyses of the three models entering Time 1 WRMT Word Attack standard scores, age, or PPVT standard scores first and examining the remaining variance showed similarly significant results.

Validation of Predictability of Models

The following tests use permutation approaches by effectively reformulating the question on the differences between the predictability of the models as measured by classifier performance in the traditionally used framework of hypothesis testing (P. Good, 1994). The model validity check is critical because the residual (or prediction) error of a multiple regression analysis may underestimate the errors found in practice when there are outliers in the data or an excessive numbers of regressors in the model. We used leave-one-out cross-validation analysis to suppress possible effects of outliers. We further performed bootstrap and split-half reliability to demonstrate the stability and predictability of model estimation (Supplemental Figure C and text online).

Using leave-one-out cross-validation analyses, we tested whether single participant Time 2 WRMT Word Attack standard scores (validation participant) can be predicted from the remaining 63 participants (training set) in the behavioral, neuroimaging, or combined models, which allows for testing of generalization error

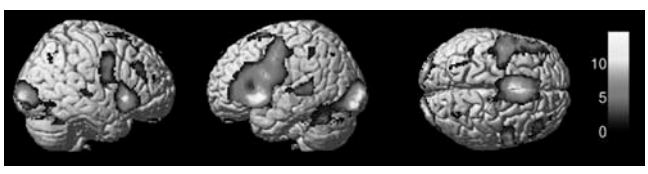

Figure 2. Brain activation for rhyme judgment condition compared with-resting baseline. Greater activation is evident in the rhyme judgment condition. Scaling bar represents t values.

Table 2
Brain Activation for Rhyme > Rest Contrast

Region	Brodmann area	Talairach coordinates			*t*	*p* (FDR)	Voxels
		x	y	z			
Bilateral superior, inferior frontal gyri, insula, middle, inferior occipital gyri, cuneus	6, 18, 47, 13	4	9	55	14.46	<.001	37,815
Right precuneus, superior parietal lobule	7	24	−52	45	6.17	<.001	297
Left middle, inferior frontal, precentral gyri	6, 9, 4	−55	2	40	5.56	<.001	947
Left inferior, superior parietal lobules, angular gyrus, precuneus	7, 39	−30	−56	47	4.25	<.001	180
Left middle frontal gyrus	46, 9	−40	36	20	3.98	.001	233
Right inferior parietal lobule	40	44	−41	44	3.97	.001	68
Left superior temporal gyrus	22	−53	−23	3	3.93	.001	73
Left transverse, superior temporal gyri	41	−42	−32	11	3.57	.002	35
Left superior, middle frontal gyrus	10	−32	49	14	3.09	.006	18

Note. FDR = false-discovery rate.

(Figure 1b and Method section). All models showed significant correlation between prediction indices and Time 2 Word Attack standard scores (p's $< .001$), indicating the validity of the models (see Supplemental Figure D1–3 online). For each of these 64 predicted values, we measured the deviation from the actual Time 2 Word Attack standard scores to measure how accurately the validation participant's Time 2 Word Attack standard scores could be predicted. The deviation was on average 5.33 (SD = 4.24) Word Attack standard score points for the behavioral model, 5.81 (SD = 4.81) points for the neuroimaging model, and 4.17 (SD = 0.52) points for the combined model (see Supplemental Figure D4 online). There was a significant effect of models, $F(1, 63) = 5.33$, $p = .024$, which was driven by the significantly greater accuracy of predicting the validation participant's Time 2 Word Attack standard scores (i.e., less deviation) of the combined model compared with the behavioral, $t(63) = 2.31$, $p = .024$, and the neuroimaging models, $t(63) = 3.02$, $p = .004$. There was no significant difference between the neuroimaging and behavioral models, $t(63) = 0.78$, $p = .44$.

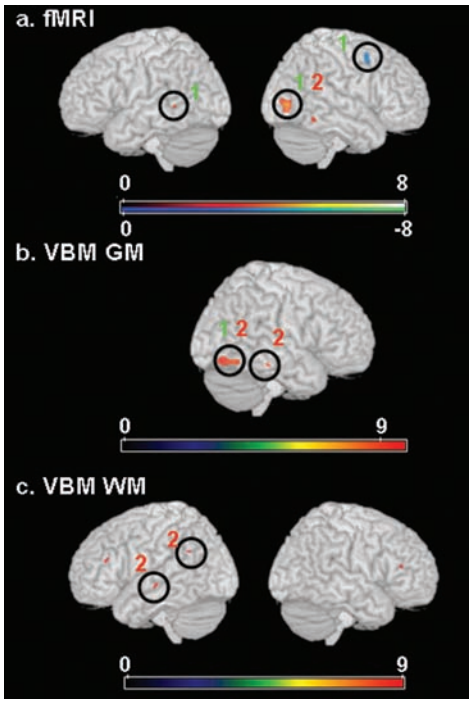

Figure 3. Neuroimaging rredictors of future decoding skill. (a) Regions showing a relation between fMRI (functional magnetic resonance imaging) brain activation (rhyme > rest) and Time 2 Woodcock Reading Master Test Word Attack standard scores. (b) Regions showing a relation between gray matter (GM) density and and Time 2 Word Attack standard scores. (c) Regions showing a relation between white matter (WM) density and Time 2 Word Attack standard scores. The numeral 1 (in green) indicates predictors included in the neuroimaging model, and the numeral 2 (in red) indicates predictors included in the combined model (note: these models were derived independently; hence, the slight differences in predictors that were included in the models). Scaling bars indicate Tesla values. VBM = voxel-based morphometry.

One might expect that the combined model performs better simply as a result of the increased number of predictors in the model. Hence, we repeated the analyses including the same number of predictors (eight or three) for each model (see Supplemental text online). Eight variables per model were chosen as the number of variables to match with the combined model, which had the most number of variables. To avoid bias toward the combined model, we also tested with three variables per model, which were chosen as the number of variables to match with the behavioral model,which contains the least number of variables and which biases toward the behavioral model. With the number of predictors held constant, the combined model performed better than the behavioral or the neuroimaging model with either eight or three variables (see Supplemental Figure E and text online).

Discussion

We examined how well behavioral and brain measures taken at the beginning of the school year predicted a critical ability for reading, that is decoding skills, at the end of the school year for children 8–12 years of age. Standardized behavioral measures of reading and language yielded a behavioral model that accounted for 65% of the variance in end-of-the-year performance on the WRMT Word Attack subtest, a standardized test of decoding. Brain imaging measures, comprised of both functional (fMRI) and gray and white matter morphological (VBM) scores, yielded a neuroimaging model that accounted for 57% of later variance in

Table 3
Whole-Brain Correlations With Time 2 Woodcock Reading Mastery Test Word Attack Standard Scores

Region	Brodmann area	Talairach coordinates x	y	z	*t*	*p*[a]	Voxels
	fMRI						
Positive correlation with Time 2 WRMT WA ss							
Right fusiform gyrus; middle occipital gyri[b,c]	18,19	28	−75	6	4.61	<.001	133
Right fusiform gyrus	37	24	−49	−9	3.49	<.001	12
Left middle temporal gyrus[b]	22	−51	−47	2	3.47	<.001	12
Negative correlation with Time 2 WRMT WA ss							
Right middle frontal gyrus[b]	6	42	8	46	3.83	<.001	52
	Voxel-based morphometry						
Gray matter: Positive correlation with Time 2 WRMT WA ss		30	−72	−6	9.82	<.001	508
Right fusiform gyrus; posterior[b,c]	19	30	−64	−8	9.03	<.001	
Right fusiform gyrus, anterior[c]	20	42	−30	−13	9.07	<.001	18
White matter: Positive correlation with Time 2 WRMT WA ss							
Right frontal lobe		29	34	13	8.96	<.001	33
Left medial frontal lobe		−19	32	19	8.94	<.001	48
Left superior temporal lobe[c]		−32	−19	−3	8.94	<.001	45
Left inferior parietal lobule[c]		−31	−51	31	8.93	<.001	24

Note. WRMT WA ss = Woodcock Reading Mastery Test Word Attack standard scores; fMRI = functional magnetic resonance imaging.
[a] For fMRI, *p* values are uncorrected; for VBM, *p* values are corrected for family-wise error. [b] Brain regions that remained as *neuroimaging predictors.* [c] Brain regions that remained as *combined predictors.*

decoding ability (which was nonsignificantly less than the behavioral model). Most importantly, the combined model of behavioral and neuroimaging measures was most predictive of later decoding skills, and explained 81% of the variance. The combined model was significantly better than either the behavioral or the neuroimaging model, as indicated by direct comparisons with multivariate analyses and validation tests. Thus, neuroimaging provided a unique kind of predictive information that was not merely redundant with behavioral measures. The combination of behavioral and neuroimaging most accurately predicted how much a year of education would influence a fundamental reading skill.

Subsidiary analyses supported the reliability and validity of all three models. The leave-one-out cross-validation, bootstrap, and split-half reliability analyses indicated that the findings were not due to either outlier values or too many regressors in the models, and that the models were stable. The findings held when the number of regressors was equated across the models or when initial WRMT Word Attack standard scores were used as a covariate (although this reduced the explained variance of all models). Thus, the behavioral and neuroimaging measures were similar in their validity and robustness as predictors.

In agreement with previous studies that have repeatedly shown phonological awareness to be one of the best predictors of reading success (e.g., Juel, 1988), many behavioral tests that showed significant correlation with Time 2 WRMT Word Attack standard scores were related to decoding and phoneme awareness. Time 1 Word Attack scores, which should be most predictive of Time 2 Word Attack scores, that is, the outcome measure in our study, accounted for 49% of the variance on their own. Further, grapheme–phoneme knowledge (spelling, Time 1 Woodcock Johnson Spelling standard scores) was another strong predictor. One rather surprising predictor, however, was the children's ability to calculate (Time 1 Woodcock Johnson Calculation standard scores). This is, however, in agreement with what has been found previously (Nairoo, 1972). The prediction of later decoding skills from the ability to calculate may also be related to a higher prevalence of dyscalculia (a condition with a specific disturbance of arithmetic ability) in dyslexia (a developmental disorder characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction (Lyon, Shaywitz, & Shaywitz, 2003) ranging from 17 to 64% (e.g., Badian, 1999; Gross-Tsur, Manor, & Shalev, 1996). These studies suggest that there may be a relationship between calculation and reading skills. In addition, neuroimaging studies have shown relationships between language or phonological processing and calculation in language-related brain regions, including the left parietal region (Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; Simon, Mangin, Cohen, Le Bihan, & Dehaene, 2002). However, children with reading difficulties and children with comorbid reading and mathematics difficulties progress at about the same rate in reading achievement (Jordan, Kaplan, & Hanich, 2002). The exact role of calculation abilities in predicting later decoding skills needs further investigation. Other reading measures tested in these children did not remain as predictors in the behavioral model, which may partly be due to collinearity effects of Time 1 Word Attack standard scores and other reading tests.

Some have questioned the utility of behavioral tests for accurately predicting risk for poor decoding skills (Hammill, Mather, Allen, & Roberts, 2002). In a behavioral study examining 200 children between 1st and 6th grades with multiple regression analyses, phonology composites accounted for 40% of the variance in younger children and 42% of the variance when all children were combined in predicting word identification skills. The authors of that study concluded, however, that none of the composites studied met criteria that are considered to be practically useful. More recent multivariate studies, however, show better predictive

values (e.g., Bowey, 2005), in which the results are more consistent with our results.

Neuroimaging predictors of later superior decoding skills included greater brain activation in the right fusiform ~ middle occipital and left middle temporal gyri, lesser activation in the right middle frontal gyrus, greater gray matter density in right fusiform gyrus, and greater white matter density in the left superior temporal and inferior parietal regions. Findings that greater left temporal-lobe activation and white matter are associated with superior decoding skill are consistent with prior studies in normal and dyslexic readers (e.g., structural studies: Deutsch et al., 2005; Klingberg et al., 2000; Silani et al., 2005; functional studies: Turkeltaub, Gareau, Flowers, Zeffiro, & Eden, 2003). The relation between lesser right frontal activation and later superior decoding skill may be related to findings indicating that the development of reading ability involves a reduction of right-hemisphere activation and a growth of left-hemisphere activation. Less expected were the relations of greater right fusiform activation and gray matter density with later superior decoding. Some studies have reported reduced VBM gray matter (W. E. Brown et al., 2001; Eckert et al., 2005) and activation (Aylward et al., 2003) in the right occipitotemporal region in dyslexia, which may imply that greater activation leads to better reading outcome. Other studies, however, have reported a developmental decrease in right fusiform activation associated with increasing age or gains in reading and language tasks in a cross-sectional study of healthy children and adults (T. T. Brown et al., 2005; Turkeltaub et al., 2003); these studies may indicate that less activation (more like that of adults) leads to better outcome and, hence, may be inconsistent with our results.

It is difficult to directly relate our findings, which are derived from a sample of children with a broad range of reading ability, to prior imaging studies mentioned above that examined severely dyslexic, highly skilled typically reading participants, or both. Speculatively, it may also be that the development of reading ability in the age range of our study depends transiently on mechanisms supported by the right fusiform gyrus before becoming dependent on left-hemisphere mechanisms in the adult reader. This possibility is supported not only by findings of right fusiform activation in children performing reading tasks (Aylward et al., 2003; T. T. Brown et al., 2005; Turkeltaub et al., 2003) but also by evidence that effective remediation for dyslexia involves not only increased activation in left-hemisphere language areas but also increased activation in many right-hemisphere areas (Aylward et al., 2003; B. A. Shaywitz et al., 2004; Temple et al., 2003).

The combined model predicted later decoding skills significantly better than either the behavioral or neuroimaging models, but our study has several limitations. First, the sample is not epidemiologically representative and, therefore, generalizability of the findings is unknown. Second, participants were followed for only for one school year, and evaluation of longer term predictive models will be important. Third, we focused on one measure of word decoding, a pseudoword reading task that measures phonological processing as the outcome measure. We chose this measure because decoding accounts for most of the variance in reading comprehension, the development of language specific phonology is essential for reading success, and early and systematic emphasis on decoding leads to better achievement of reading skills (Adams, 1990; Hulme & Snowling, 2005; E. Richardson, DiBenedetto, & Adler, 1982; Shankweiler et al., 1999; Snow et al., 1998; Snowling, 1987). There may, however, be better measures or combinations of measures that will be more suitable as an outcome measure (Leonard, Eckert, Given, Virginia, & Eden, 2006), because reading comprehensions and reading fluency involve many processes beyond single-word decoding. In addition, one might argue that gains in reading ability may be a more appropriate outcome measure than Time 2 scores. In our preliminary analyses (results not shown), activation of brain regions that predicted gains in WRMT Word Attack scores were, however, similar to those of this study.

Fourth, behavioral and neuroimaging predictors used here were selected from univariate and multivariate analyses with the total sample and were applied to the validation tests, that is, for each permutation, ROIs were not re-identified, and the same contrast estimates and gray and white matter volume were used throughout. In addition, there may be important predictors that may not show significant correlation in a univariate analysis but may contribute significantly when included in multivariate analyses. Fifth, we deliberately chose a purely empirical approach so that we could optimize both behavioral and brain predictions of Time 2 scores. Such a purely empirical approach may highlight brain regions that are not yet well understood in reading, and that may merit a hypothesis-based approach in the future, but reduces our ability to interpret why certain brain measures predicted future decoding skill. Future studies with a predefined set of brain regions (a more theoretical approach) or utilizing multi-voxel pattern analysis (MVPA) will be of interest (Norman, Polyn, Detre, & Haxby, 2006). Sixth, the models created examined linear relationships, and an increasing number of studies show nonlinear effects of development (e.g., Shaw et al., 2006). Whether this approach has true clinical utility is unknown. Although the leave-one-out cross-validation analyses showed significantly greater prediction for the combined as compared with the behavioral model, the gain was only 1.17 Word Attack standard score points on average (4.16 vs. 5.33). It is thought that the sensitivity index, specificity index, and positive predictive value should all reach at least 75% in order for a measure to be considered acceptable for practical use and suitable for screening purposes (Gredler, 2000). In our sample, and using the behavioral, neuroimaging, and the combined models, we found that high sensitivity, specificity, and positive predictive value greater than 75% were achieved in classifying children with reading disability at Time 2 (data not shown); reading disability was defined by Time 2 Word Attack standard scores of 85 or below, which is a common threshold. Ultimately, identification of a predefined set of predictors independent of the sample in a prospective study followed up for a longer period of time that passes the above threshold (a minimum of 75%) will be necessary.

More generally, our findings relate to one potential practical use of neuroimaging, namely, the prediction of future health or behavior. Neuroimaging has been used to predict the outcome of treatment for depression (e.g., Canli et al., 2005; Siegle, Carter, & Thase, 2006) and the conversion from healthy aging to Alzheimer's disease (e.g., Apostolova et al., 2006; Bookheimer et al., 2000; de Leon et al., 2001). These studies have often used smaller samples and a single imaging modality, and only one study has examined the validity of a model with a permutation test (Apostolova et al., 2006). Another study, however, examined preoperative behavior, brain volume, and fMRI in 10 temporal-lobe epilepsy patients to predict post-operative memory and compared sensitivity, specificity, and positive predictive value in a small

sample of 10 participants (M. P. Richardson et al., 2004). They found that left–right hippocampal encoding activity difference showed reasonable sensitivity, specificity, and positive predictive value (20–100 %) for predicting the amount of pre- to post-operative memory decline. Over time, perhaps in combination with an individual's genetic information, neuroimaging may contribute to increasingly accurate predictions of future behaviors. In the case of reading difficulties, identification of children at risk as early as possible seems desirable so that interventions may be implemented prior to reading failure and perhaps prior to the development of disadvantageous reading habits that may slow the effectiveness of interventions. Judicious use of predictive measures would require consideration of predictive accuracy at the individual level, such as sensitivity, specificity, and cost–benefit balances. Ethical considerations will also be important to avoid abuses of neuropredictive measures (although these ethical considerations are not fundamentally different from those of other kinds of predictive measures; Illes & Raffin, 2005). In this vein, it will also be important to recognize that brain dysfunction in dyslexia can be altered by remediation (e.g., Temple et al., 2003), indicating that effective education can guide beneficial plasticity.

Taken together, these findings indicate that neuroimaging measures predict decoding skill after a year of school almost as well as do current standardized tests and that behavioral tests and neuroimaging measures in combination predict decoding skill significantly better than either kind of measure alone. The significantly greater predictive accuracy of the combined behavioral–neuroimaging model than either model alone shows that neuroimaging is measuring brain functions and structures relevant to reading that are not fully measured by their behavioral correlates in standardized testing. There are still many steps to be taken to show that neuroimaging measures have sufficient value before such measures ought to be considered for practical prediction of the need for educational intervention. Combined behavioral and neuroimaging measures, however, hold promise for improving the specificity and effectiveness of early intervention and later achievement of reading skills. The present findings, therefore, suggest a point where a useful bridge can be built between cognitive neuroscience and education.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Apostolova, L. G., Dutton, R. A., Dinov, I. D., Hayashi, K. M., Toga, A. W., Cummings, J. L., et al. (2006). Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. *Archives of Neurology, 63,* 693–699.

Aylward, E. H., Richards, T. L., Berninger, V. W., Nagy, W. E., Field, K. M., Grimme, A. C., et al. (2003). Instructional treatment associated with changes in brain activation in children with dyslexia. *Neurology, 61,* 212–219.

Badian, N. A. (1999). Persistent arithmetic, reading, or arithmetic and reading disability. *Annals of Dyslexia: An Interdisciplinary Journal of the Orton Dyslexia Society, 49,* 45–70.

Bookheimer, S. Y., Strojwas, M. H., Cohen, M. S., Saunders, A. M., Pericak-Vance, M. A., Mazziotta, J. C., et al. (2000). Patterns of brain activation in people at risk for Alzheimer's disease. *New England Journal of Medicine, 343,* 450–456.

Bowey, J. A. (2005). Predicting individual differences in learning to read. In C. Hulme & M. Snowling (Eds.), *Science of reading: A handbook* (pp. 155–172). Malden, MA: Blackwell.

Brown, T. T., Lugar, H. M., Coalson, R. S., Miezin, F. M., Petersen, S. E., & Schlaggar, B. L. (2005). Developmental changes in human cerebral functional organization for word generation. *Cerebral Cortex, 15,* 275–290.

Brown, W. E., Eliez, S., Menon, V., Rumsey, J. M., White, C. D., & Reiss, A. L. (2001). Preliminary evidence of widespread morphological variations of the brain in dyslexia. *Neurology, 56,* 781–783.

Bruer, J. T. (1997). Education and the brain: A bridge too far. *Educational Researcher, 26,* 1–13.

Bruer, J. T. (2002). Avoiding the pediatrician's error: How neuroscientists can help educators. *Nature Neuroscience, 5,* 1031–1033.

Canli, T., Cooney, R. E., Goldin, P., Shah, M., Sivers, H., Thomason, M. E., et al. (2005). Amygdala reactivity to emotional faces predicts improvement in major depression. *Neuroreport, 16,* 1267–1270.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Science, 9,* 335–341.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284,* 970–974.

de Leon, M. J., Convit, A., Wolf, O. T., Tarshish, C. Y., DeSanti, S., Rusinek, H., et al. (2001). Prediction of cognitive decline in normal elderly subjects with 2-[(18)F]fluoro-2-deoxy-D-glucose/positron-emission tomography (FDG/PET). *Proceedings of the National Academy of Sciences of the United States of America, 98,* 10966–10971.

Deutsch, G. K., Dougherty, R. F., Bammer, R., Siok, W. T., Gabrieli, J. D., & Wandell, B. (2005). Children's reading performance is correlated with white matter structure measured by diffusion tensor imaging. *Cortex, 41,* 354–363.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Eckert, M. A., Leonard, C. M., Wilke, M., Eckert, M., Richards, T., Richards, A., et al. (2005). Anatomical signatures of dyslexia in children: Unique information from manual and voxel based morphometry brain measures. *Cortex, 41,* 304–315.

Eden, G. F., & Zeffiro, T. A. (1998). Neural systems affected in developmental dyslexia revealed by functional neuroimaging. *Neuron, 21,* 279–282.

Espy, K. A., Molfese, D. L., Molfese, V. J., & Modglin, A. (2004). Development of auditory event-related potentials in young children and relations to word-level reading abilities at age 8 years. *Annals of Dyslexia, 54,* 9–38.

Golestani, N., Paus, T., & Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron, 35,* 997–1010.

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage, 14,* 21–36.

Good, P. (1994). *Permutation tests: A practical guide to resampling methods for testing hypothesis*: Springer-Verlag.

Goswami, U. (2006). Neuroscience and education: From research to practice? *Nature Reviews Neuroscience, 7,* 406–411.

Gredler, G. R. (2000). Early childhood screening for developmental and educational problems. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (pp. 399–411). Boston: Allyn & Bacon.

Gross-Tsur, V., Manor, O., & Shalev, R. S. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine and Child Neurology, 38,* 25–33.

Hammill, D. D., Mather, N., Allen, E. A., & Roberts, R. (2002). Using semantics, grammar, phonology, and rapid naming tasks to predict word identification. *Journal of Learning Disabilities, 35,* 121–136.

Hoeft, F., Hernandez, A., McMillon, G., Taylor-Hill, H., Martindale, J. L., Meyler, A., et al. (2006). Neural basis of dyslexia: A comparison between dyslexic and nondyslexic children equated for reading ability. *Journal of Neuroscience, 26,* 10700–10708.

Hulme, C., & Snowling, M. (Eds.). (2005). *Science of reading: A handbook*. Malden, MA: Blackwell.

Illes, J., & Raffin, T. A. (2005). No child left without a brain scan? Toward a pediatric neuroethics. *Cerebrum, 7,* 33–46.

Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology, 94,* 586–597.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80,* 437–447.

Klingberg, T., Hedehus, M., Temple, E., Salz, T., Gabrieli, J. D., Moseley, M. E., et al. (2000). Microstructure of temporo-parietal white matter as a basis for reading ability: Evidence from diffusion tensor magnetic resonance imaging. *Neuron, 25,* 493–500.

Leonard, C., Eckert, M., Given, B., Virginia, B., & Eden, G. (2006). Individual differences in anatomy predict reading and oral language impairments in children. *Brain, 129,* 3329–3342.

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia, 53,* 1–14.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends Cogn Sci, 7,* 293–299.

Minotani, C. (2004). *Toukeigaku-nyuumon* [*Introduction to statistics*]. Tokyo: Tokyo Tosho.

Molfese, V. J., Molfese, D. L., & Modgline, A. A. (2001). Newborn and preschool predictors of second-grade reading scores: An evaluation of categorical and continuous scores. *Journal of Learning Disabilities, 34,* 545–554.

Nairoo, S. (1972). *Specific dyslexia.* London: Pitman.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Science, 10,* 424–430.

Poulakis, V., Witzsch, U., de Vries, R., Emmerlich, V., Meves, M., Altmannsberger, H. M., et al. (2004). Preoperative neural network using combined magnetic resonance imaging variables, prostate specific antigen, and Gleason score to predict prostate cancer recurrence after radical prostatectomy. *European Urology, 46,* 571–578.

Price, C. J., & Mechelli, A. (2005). Reading and reading disturbance. *Current Opinion in Neurobiology, 15,* 231–238.

Richardson, E., DiBenedetto, B., & Adler, A. (1982). Use of the decoding skills test to study differences between good and poor readers. *Advances in Learning and Behavioral Disabilities, 1,* 25–74.

Richardson, M. P., Strange, B. A., Thompson, P. J., Baxendale, S. A., Duncan, J. S., & Dolan, R. J. (2004). Pre-operative verbal memory fMRI predicts post-operative memory decline after left temporal lobe resection. *Brain, 127,* 2419–2426.

Shankweiler, D., Lundquist, E., Katz, L., Stuebing, K. K., Fletcher, J. M., Brady, S., et al. (1999). Comprehension and decoding: Patterns of association in children with reading difficulties. *Scientific Studies of Reading, 3,* 69–94.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature, 440,* 676–679.

Shaywitz, B. A., Shaywitz, S. E., Blachman, B. A., Pugh, K. R., Fulbright, R. K., Skudlarski, P., et al. (2004). Development of left occipitotemporal systems for skilled reading in children after a phonologically-based intervention. *Biological Psychiatry, 55,* 926–933.

Shaywitz, S. E., & Shaywitz, B. A. (2005). Dyslexia (specific reading disability). *Biological Psychiatry, 57,* 1301–1309.

Siegle, G. J., Carter, C. S., & Thase, M. E. (2006). Use of fMRI to predict recovery from unipolar depression with cognitive behavior therapy. *American Journal of Psychiatry, 163,* 735–738.

Silani, G., Frith, U., Demonet, J. F., Fazio, F., Perani, D., Price, C., et al. (2005). Brain abnormalities underlying altered activation in dyslexia: A voxel based morphometry study. *Brain, 128,* 2453–2461.

Simon, O., Mangin, J. F., Cohen, L., Le Bihan, D., & Dehaene, S. (2002). Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. *Neuron, 33,* 475–487.

Snow, C. E., Burns, S. M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Snowling, M. J. (1987). *Dyslexia: A cognitive developmental perspective.* Oxford, United Kingdom: Basil Blackwell.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain.* New York: Thieme.

Temple, E., Deutsch, G. K., Poldrack, R. A., Miller, S. L., Tallal, P., Merzenich, M. M., et al. (2003). Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America, 100,* 2860–2865.

Turkeltaub, P. E., Gareau, L., Flowers, D. L., Zeffiro, T. A., & Eden, G. F. (2003). Development of neural mechanisms for reading. *Nature Neuroscience, 6,* 767–773.

Woodhouse, L. J., Reisz-Porszasz, S., Javanbakht, M., Storer, T. W., Lee, M., Zerounian, H., et al. (2003). Development of models to predict anabolic response to testosterone administration in healthy young men. *American Journal of Physiology, Endocrinology, and Metabolims, 284,* E1009–E1017.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* New York: Touchstone Applied Science.

Received November 15, 2006
Revision received January 12, 2007
Accepted January 31, 2007 ■