

## SUPPLEMENTARY INFORMATION

### Consistency Maps of fMRI and VBM Permutation Analyses

We performed leave-one-out cross-validation analyses on fMRI and VBM whole brain analyses to examine the consistency of brain activation and morphometric patterns. We performed regression analyses as in **Definition of Prediction Models** in the **Main Text Methods**, but including 63 participants ( $N-1$ ) for each whole-brain group analysis and permuted 64 times. The thresholded statistical maps (fMRI:  $p = .001$  uncorrected,  $et = 10$ ; VBM:  $p = .00001$  FWE corrected,  $et = 10$ ) were transformed into binary maps and were summed for fMRI and VBM separately. Hence, for example, if significant correlation between Time2 WA-ss and brain activation was observed consistently in all 64 permutations in a certain voxel, then the spectral color map would indicate a value of 64. These analyses validated highly consistent activation and morphometric patterns (**Supplementary Figure 2** online).

### Direct Comparison of Models

We compared whether the simple combination of behavioral and neuroimaging predictors (*7-variable Combined Model*, **Main Text Figure 1a**, right-most flow-chart) explained the variance of Time2 WA-ss significantly better than the behavioral or neuroimaging models using the equation (Howell, 2002):

$$F_{(k_2 - k_1, N - k_2 - 1)} = \frac{(R_2^2 - R_1^2) / (k_2 - k_1)}{(1 - R_2^2) / (N - k_2 - 1)}$$

where  $R_1^2$  is the proportion of  $Y$ -variance explained by the first set of  $k_1$  predictors, and  $R_2^2$  is the proportion of  $Y$ -variance explained by the larger set of  $k_2$  predictors,  $N$  is the number of participants, hence having  $(k_2 - k_1, N - k_2 - 1)$ df (**Main Text Figure 1c**) (Howell, 2002). This allowed us to directly compare between models using multiple regression as 3-variable behavioral model and 4-variable neuroimaging models are sub-sets of the 7-variable combined model. Results are provided in the main text.

### Prediction Analyses Controlling for Initial Decoding Skills

There is a possibility that the results reported thus far may merely be a result of strong associations between Time1 WA-ss and the behavioral scores or prediction indices, rather than a unique contribution of Time2 WA-ss. Therefore, we performed partial correlation analyses for the three models and found that the results remained highly significant.

When partial correlation was performed between the behavioral predictors at Time1 and Time2 WA-ss controlling for Time1 WA-ss, most predictors remained significant: WJC-ss ( $r = .26$ ,  $p = .043$ ) and WJS-ss ( $r = .35$ ,  $p = .006$ ),  $ROI_{fMRI}$  RFG/MOG ( $r = .43$ ,  $p < .001$ ),  $ROI_{fMRI}$  RMFG ( $r = -.30$ ,  $p = .016$ ),  $ROI_{GM}$  RFGp ( $r = .39$ ,  $p = .002$ ),  $ROI_{WM}$  LIPL ( $r = .31$ ,  $p = .013$ ),  $ROI_{WM}$  LSTL ( $r = .27$ ,  $p = .031$ ). There was a trend for significance for  $ROI_{fMRI}$  LMTG ( $r = .23$ ,  $p = .075$ ), and  $ROI_{GM}$  RFGa ( $r = .19$ ,  $p$

= .14) was no longer significant. Note: one cannot calculate partial correlation with Time2 WA-ss and Time1 WA-ss partialing out Time1-WAss was therefore excluded from the analyses.

Finally, when the behavioral, neuroimaging and combined models were re-estimated only using those that remained significant after controlling for Time1 WA-ss, all models still explained variance of Time2 WA-ss strongly: behavioral model (multiple r-square = .21,  $p < .001$ ), neuroimaging model (multiple r-square = .40,  $p < .001$ ) and 8-variable combined model (multiple r-square = .59,  $p < .001$ ). Not surprisingly, now the neuroimaging model explained variance much more than the behavioral model.

### **Rationale of Analyses Used**

Correlation measures how well an overall data set matches a model (a linear regression model in our case). The residual error of a multiple regression analysis describes the accuracy of a model, but its accuracy can be limited by outliers in the data or an excessive number of regressors in the model. We use additional validation methods to check for these errors. Similar methods occur in existing neuroimaging literature (bootstrap: (Woodhouse et al., 2003), split-half reliability: (Bullmore et al., 1999; Golland & Fischl, 2003; Nichols & Holmes, 2002; Poulakis et al., 2004; Thompson et al., 2003), leave-one-out cross validation: (Vos et al., 2005)). Validation tells how stable a model is, by comparing a model built from one set of data to a model built from an entirely separate set of data (in our case, a split half validation). Validation protects against over-fitting a model in the correlation analysis. Finally, prediction refers to how well a participant's score can be predicted based on the input components to the model, where the model is built independently from that participant's data. The leave-one-out analysis shows the prediction performance of the model. Note that split-half reliability and leave-one-out cross-validation should use entirely independent data sets to be strictly valid. The analyses used here subtly share information between data sets because the best behavioral tests and the best neuroimaging regions were determined by the whole data set. However, these tests and regions are also well-established from earlier studies and could have been treated as a priori information for this study. Thus, we believe that this contamination effect on validation can safely be neglected.

### **Bootstrap Analyses: Stability of Models using Different Training Set Sizes**

This analysis was designed to test for the stability of the behavioral, neuroimaging and combined models (**Main Text Figure 1b**). A training set was defined as a random sub-sample of the data ranging from 10 to 63 participants with 2000 iterations per training-set size (Efron & Tibshirani, 1993). For each training-set, multiple regression analysis was performed, and the obtained F values were averaged and converted to P values. Multiple r-square and 95 % confidence intervals for each training-set size were calculated and plotted against training-set size, i.e., number of participants included in the analyses (**Supplementary Figure 3a** online). Number of participants required to obtain significance ( $p = .001$ ) were obtained for each model to examine the stability of the three models.

When correlations between Time2 WA-ss and prediction indices were examined, the combined model explained the variance of Time2 WA-ss better than the behavioral model, which in turn was better than the neuroimaging model at all training-set sizes (**Supplementary Figure 3a** online). In order to reach a significance level of  $p = .001$ , for example, a minimum of 22 participants were required using the combined model, 26 using the behavioral model, and 33 using the neuroimaging model.

### **Split-Half Reliability Analyses: Predictability of Models**

A potential risk in multiple regression is that too many regressors may overfit the data (and underestimate prediction error). A split-half reliability check shows that the model estimates are stable with respect to the choice of input data, and hence the model does not overfit the data. In this validation test, we investigated whether data from half the participants (training set) can predict the group characteristics of the remaining participants (validation set) in the behavioral, neuroimaging or 8-variable combined model (**Main Text Figure 1b**). Training set size was fixed to 32 participants, randomly sampled from the total sample, and the test set size was the remaining 'hold-out' set, fixed to 32 participants without overlap. We performed multiple regression analyses with the training set which gave  $b_i$ 's. In the training set, we calculated the correlation with predicted values derived from either the behavioral, neuroimaging or combined predictors in the 32 participants, and Time2 WA-ss. We then applied these coefficients  $b_i$ 's to predict Time2 WA-ss for each participant in the validation set. Simple correlation was performed between the actual and the predicted Time2 WA-ss in all 32 participants of the validation set. 2000 iterations were performed for each training-set (Efron & Tibshirani, 1993), and the 2000 correlation coefficients (also known as the cross-validated square multiple correlation) from the validation set were plotted for each model with their mean and 95 % confidence intervals (**Supplementary Figure 3b** online). This analysis also showed that the combined model explained variance better than the behavioral model, which, in turn, performed better than the neuroimaging model

### **Leave-One-Out Cross Validation Method Comparing Three Models With Matching Number Of Variables**

In order to exclude the possibility that the combined model showed significantly better predictability due to having more variables (eight for the combined model vs. three or four for the behavioral and neuroimaging models, respectively), we performed secondary analyses creating behavioral and neuroimaging models with eight variable each matching the number of variables included in the Combined Model, and with three variables each matching the number of variables included in the Behavioral Model. For the 8-variable behavioral and neuroimaging models, we stopped multiple regression analysis (backward stepwise procedure) when the model reached eight variables. For the 3-variable neuroimaging and the combined models, we stopped multiple regression analysis (forward stepwise procedure) when the model reached three variables. We performed identical leave-one-out cross validation analyses and one-way ANOVA to compare between models.

For the 8-variable comparison, the deviation was 5.44 (sd = 4.13) WA-ss for the behavioral

model, 5.25 (sd = 4.70) for the neuroimaging model, and 4.17 (sd = 3.52) for the combined model (**Supplementary Figure 5a** online). There was a significant effect of models ( $F_{(1,63)} = 6.85$ ,  $p = .011$ ) which was driven by the significantly greater accuracy (i.e., less deviation) of the combined model compared to the behavioral ( $t_{(63)} = 2.62$ ,  $p = .011$ ) and the neuroimaging models ( $t_{(63)} = 2.32$ ,  $p = .024$ ). There was no significant difference between the neuroimaging and behavioral models ( $t_{(63)} = .29$ ,  $p = .78$ ).

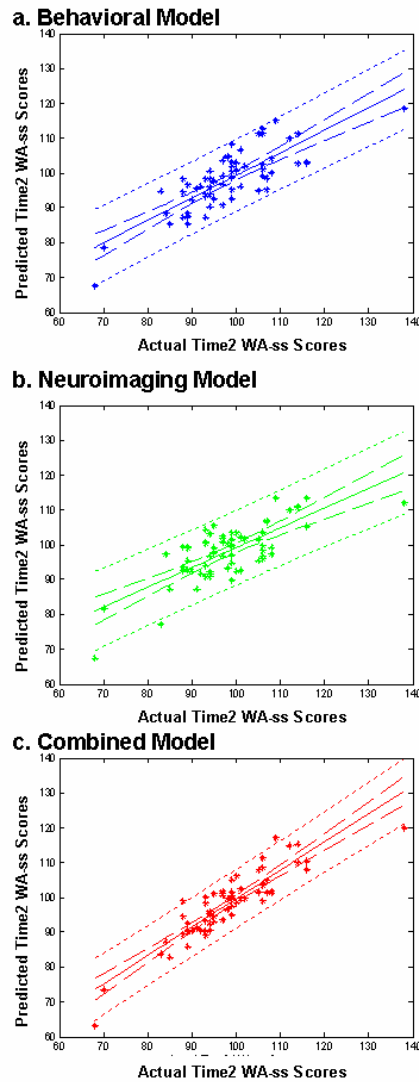
For the 3-variable comparison, the deviation was 5.33 (sd = 4.24) WA-ss for the behavioral model, 5.96 (sd = 4.83) for the neuroimaging model, and 4.84 (sd = 3.95) for the combined model (**Supplementary Figure 5b** online). While the 3-variable comparison showed effects in the same direction as the 8-variable comparison, there was only a trend for a significant effect of models ( $F_{(1,63)} = 3.18$ ,  $p = .079$ ) which was driven by the significantly greater accuracy (i.e., less deviation) of the combined model compared to the neuroimaging ( $t_{(63)} = 2.29$ ,  $p = .027$ ) but not compared to the behavioral model ( $t_{(63)} = 1.04$ ,  $p = .30$ ). There was also a non-significant difference between the neuroimaging and behavioral models ( $t_{(63)} = 1.07$ ,  $p = .29$ ).

## SUPPLEMENTARY REFERENCES

- Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., & Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans Med Imaging*, *18*(1), 32-42.
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.
- Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. *Inf Process Med Imaging*, *18*, 330-341.
- Howell, D. C. (2002). *Statistical Methods fo Psychology* (5 ed.). Florence, KY: Wadsworth
- Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*, *15*(1), 1-25.
- 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. *Eur Urol*, *46*(5), 571-578.
- Thompson, P. M., Hayashi, K. M., de Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., et al. (2003). Dynamics of gray matter loss in Alzheimer's disease. *J Neurosci*, *23*(3), 994-1005.
- Vos, M. J., Berkhof, J., Postma, T. J., Hoekstra, O. S., Barkhof, F., & Heimans, J. J. (2005). Thallium-201 SPECT: the optimal prediction of response in glioma therapy. *Eur J Nucl Med Mol Imaging*.
- 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. *Am J Physiol Endocrinol Metab*, *284*(5), E1009-1017.

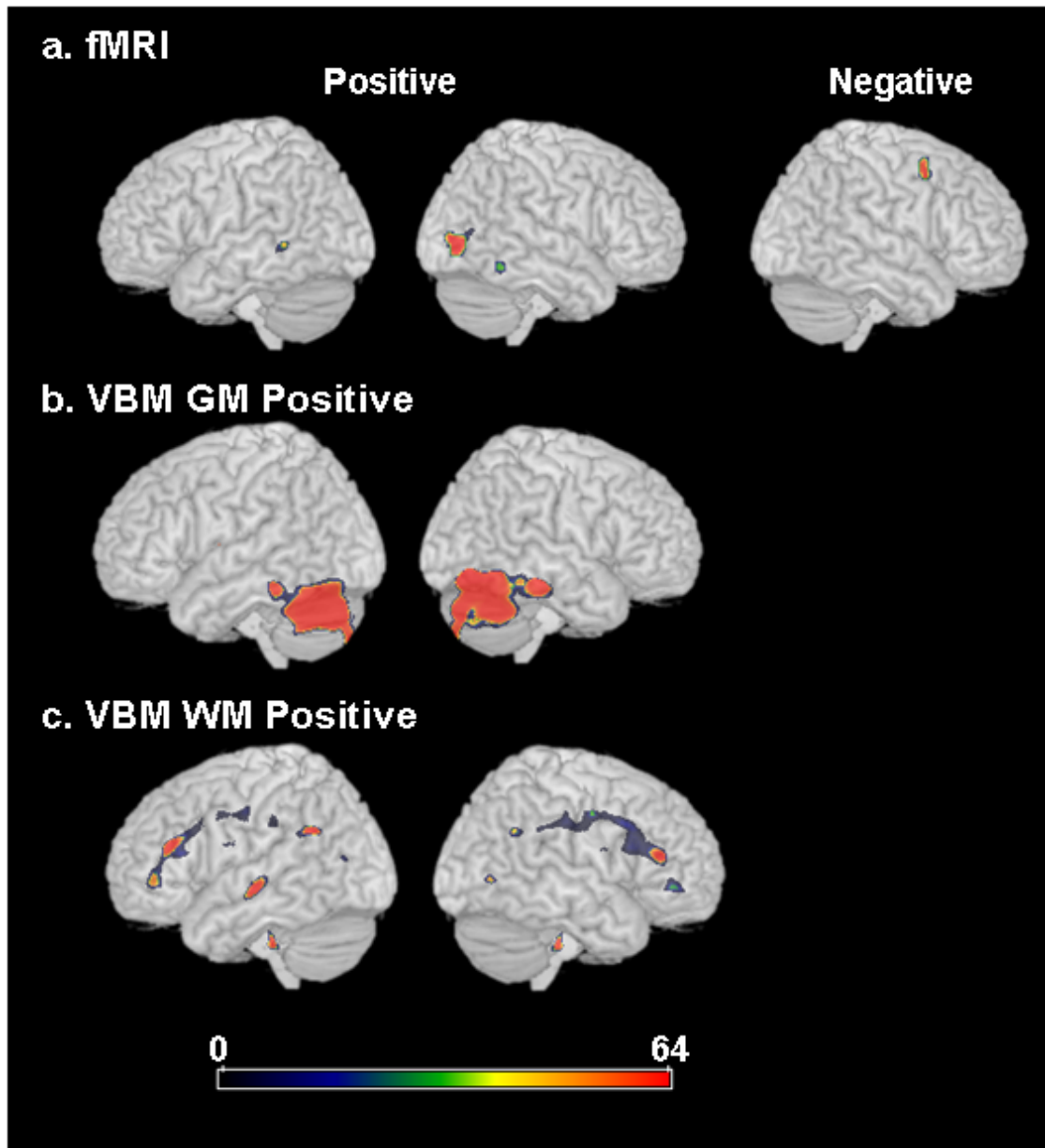
**SUPPLEMENTARY FIGURES**

**Supplementary Figure 1.**



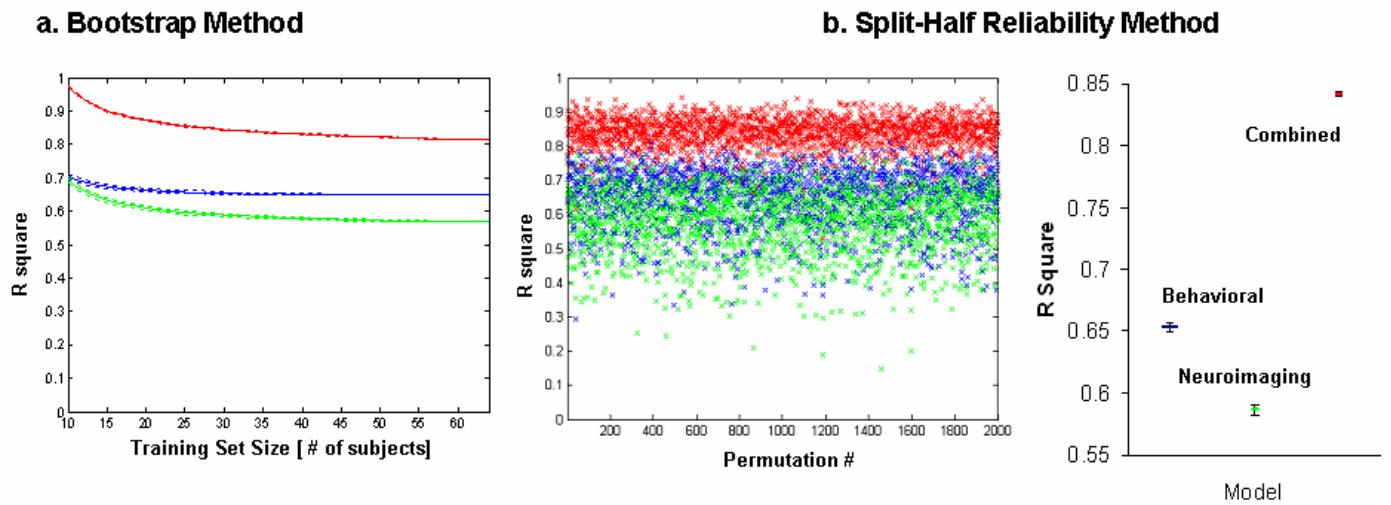
**Prediction Of Time2 Decoding Skills Using Behavioral, Neuroimaging Or Combined Models.** Each coefficient from multiple regression analyses between Time2 WA-ss and behavioral, neuroimaging or combined predictors were multiplied with their respective Time1 variables and summed with the constant to yield one Prediction Index for each participant. Time2 WA-ss was correlated with Behavioral Prediction Indices (a), Neuroimaging Prediction Indices (b), or Combined Prediction Indices (c). Linear regression lines (solid line), 95 % prediction intervals of the expected mean Time2 WA-ss (dashed lines close to linear regression line) and 95 % prediction intervals of the expected individual Time2 WA-ss (dotted/dashed lines further from linear regression line) are drawn for each model.

## Supplementary Figure 2.



**Consistency maps of brain activation and morphometry patterns from permutation analyses.** **a.** Positive (left and middle) and negative (right) correlations with brain activation during the rhyme judgment condition compared to resting baseline and Time2 WA-ss. **b.c.** Positive correlation with gray (**b**) or white (**c**) matter volume and Time2 WA-ss. For VBM, no voxels were significant for negative correlations. Maximum possible value is 64.

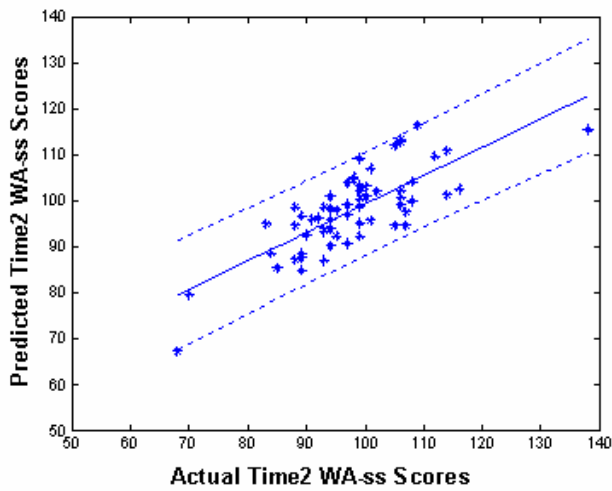
**Supplementary Figure 3.**



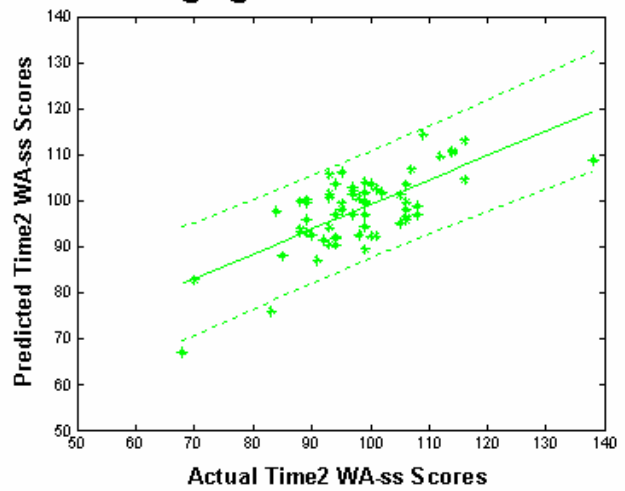
**Stability and Predictability of Models Using Bootstrap and Split-Half Reliability Methods** **a.** Number of participants included in the training-set is plotted against the estimated r-square of the prediction. The behavioral model is indicated in blue (middle solid line), the neuroimaging model in green (lower solid line) and the 8-variable combined model in red (upper solid line). 95 % confidence intervals derived from 2000 permutations are in dotted lines. **b.** Parameters obtained from multiple regression analyses in the randomly sampled training set (N = 32) were applied to the validation set constituting the 'hold-out' remaining participants (N = 32). All 2000 iterations of the correlation coefficients of the validation set for the three models (**b** left panel) and mean average values (**b** right panel) are plotted (behavioral model: blue, neuroimaging model: green, 7-variable combined model: red). Error bars overlaid on the mean averages represent 95 % confidence intervals.

**Supplementary Figure 4.**

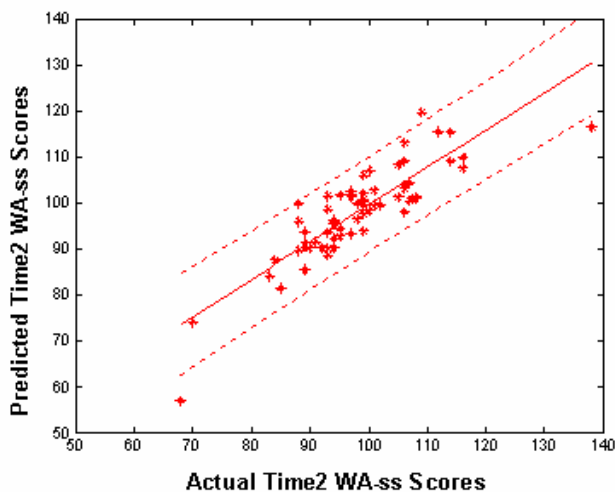
**a. Behavioral Model**



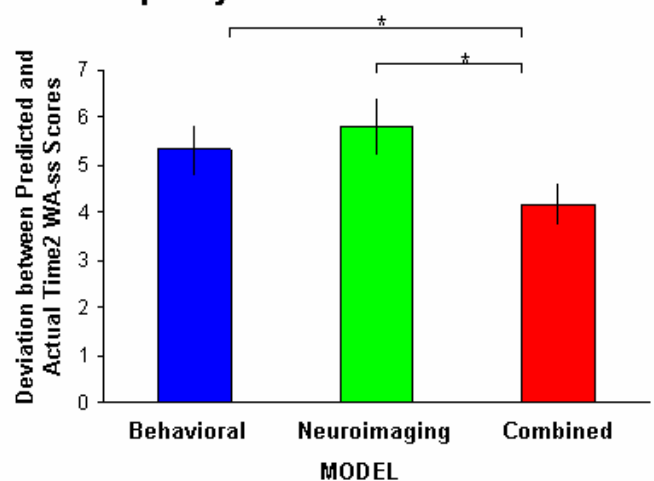
**b. Neuroimaging Model**



**c. Combined Model**



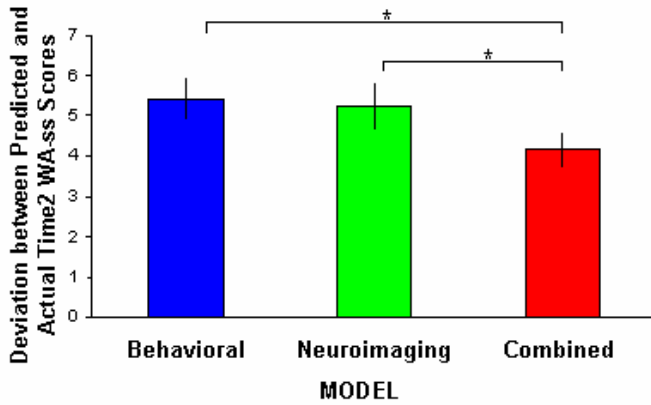
**d. Discrepancy scores**



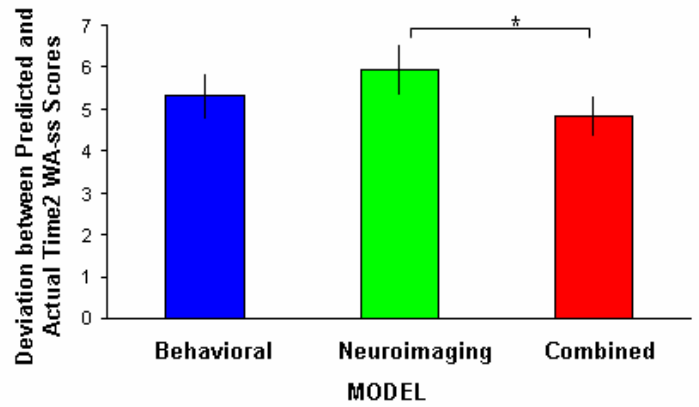
**Leave-One-Out Cross Validation Results And Discrepancy Between Predicted and Actual Time2 Reading Ability (WA-ss).** Parameters obtained from prediction analyses in the training set (N = 63) were applied to the omitted participant, which yielded one predicted value, and was repeated 64 times for all possible permutations. Predicted values obtained from the omitted single test participants are plotted for the behavioral model (a), neuroimaging model (b) and the combined model (c). Linear regression lines are obtained from the training sets and drawn as solid lines. 95 % prediction intervals of the expected individual Time2 WA-ss are in dotted curves. d. Deviation of predicted values from the actual Time2 WA-ss. Mean average of the 64 participants are plotted for the behavioral (blue), neuroimaging (green) and combined models (red). Error bars represent standard error of the mean. \*  $p < .05$

**Supplementary Figure 5.**

**a. Eight Variables Per Model**



**b. Three Variables Per Model**



**Predicting Time2 WA-ss Matching the Number of Variables in Each Model.** Leave-one-out cross-validation analyses were performed and deviation of predicted values from the actual Time2-WAss were calculated when including 8 variables (a) or 3 variables (b) per model. Mean average of the 64 participants are plotted for the behavioral (in blue), neuroimaging (in green) and combined models (in red). Error bars represent standard error of the mean. \*  $p < .05$