

# Improved estimation of general cognitive ability and its neural correlates with a large battery of cognitive tasks

Liang Zhang<sup>1</sup>, Junjiao Feng<sup>2</sup>, Chuqi Liu<sup>1</sup>, Huinan Hu<sup>1</sup>, Yu Zhou<sup>1</sup>, Gangyao Yang<sup>1</sup>, Xiaojing Peng<sup>1</sup>, Tong Li<sup>1</sup>, Chuansheng Chen<sup>3</sup>, Gui Xue<sup>1,4,\*</sup>

<sup>1</sup>State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, PR China,

<sup>2</sup>Faculty of Psychology, Tianjin Normal University, Tianjin 300387, China,

<sup>3</sup>Department of Psychological Science, University of California, Irvine, CA 92697, USA,

<sup>4</sup>Chinese Institute for Brain Research, Beijing 102206, PR China

\*Corresponding author: Gui Xue, State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Xijiekouwai Street No.19, Beijing 100875, PR China. Email: [gxue@bnu.edu.cn](mailto:gxue@bnu.edu.cn)

Elucidating the neural mechanisms of general cognitive ability (GCA) is an important mission of cognitive neuroscience. Recent large-sample cohort studies measured GCA through multiple cognitive tasks and explored its neural basis, but they did not investigate how task number, factor models, and neural data type affect the estimation of GCA and its neural correlates. To address these issues, we tested 1,605 Chinese young adults with 19 cognitive tasks and Raven's Advanced Progressive Matrices (RAPM) and collected resting state and *n*-back task fMRI data from a subsample of 683 individuals. Results showed that GCA could be reliably estimated by multiple tasks. Increasing task number enhances both reliability and validity of GCA estimates and reliably strengthens their correlations with brain data. The Spearman model and hierarchical bifactor model yield similar GCA estimates. The bifactor model has better model fit and stronger correlation with RAPM but explains less variance and shows weaker correlations with brain data than does the Spearman model. Notably, the *n*-back task-based functional connectivity patterns outperform resting-state fMRI in predicting GCA. These results suggest that GCA derived from a multitude of cognitive tasks serves as a valid measure of general intelligence and that its neural correlates could be better characterized by task fMRI than resting-state fMRI data.

**Key words:** general cognitive ability; factor models; fMRI; RAPM; cognitive tasks.

## Introduction

One important aim of cognitive neuroscience is to understand the neural mechanisms of general intelligence or general cognitive ability (GCA; Deary et al. 2010; Barbey 2018). GCA or the *g* factor was proposed based on the well-documented finding that individuals' performances on various cognitive tests are positively correlated with one another (or the "positive manifold") (Spearman 1904; Jensen 1998). GCA has been found to play important roles in many aspects of everyday life. For example, GCA is reliably correlated with school achievement (Flores-Mendoza et al. 2021; Gustafsson and Balke 1993; Roth et al. 2015), work performance (Ree et al. 1994; Kuncel et al. 2004), and even physical health and longevity (Gottfredson and Deary 2004).

Although most of the cognitive neuroscience studies have focused on the neural mechanisms of specific cognitive domains, some recent studies have attempted to decipher the neural correlates of GCA (Deary et al. 2010; Barbey 2018; Duncan et al. 2020). However, these studies had two major weaknesses. First, the measurements of GCA were inconsistent across studies. Some studies used single high-*g*-loaded tasks, especially Raven's Advanced Progressive Matrices (RAPM; Raven et al. 1978), to measure GCA (Duncan 2000; Haier et al. 2003; Finn et al. 2015), while other studies adopted intelligence tests that included diverse tasks, such as Wechsler Adult Intelligence Scale (WAIS; Gignac and Bates 2017). Each test has its limitations: RAPM might tap only certain aspects of GCA, and WAIS is time-consuming and thus hard to be

administered to a large sample. Second, many existing structural and functional brain studies of GCA had inadequate sample sizes (e.g. less than 100) (Duncan 2000; Jung and Haier 2007; Colom et al. 2010; Cole et al. 2012; Schultz and Cole 2016), yet it has been shown that a large sample size is required for both the reliable estimation of GCA (Jensen 1998, p. 95) and reproducible brain-wide association with behavior (Marek et al. 2022).

To overcome the above two weaknesses, researchers have recently turned to large-scale cohort studies that include a large sample of subjects, multiple cognitive tasks, and both structural and functional imaging. They estimated GCA by extracting the principal component from the available tasks: eleven tasks (six of them from the NIH toolbox) in the Human Connectome Project (HCP; Dubois et al. 2018; Stammen et al. 2023), four cognitive tasks in the UK Biobank (Cox et al. 2019), and nine tasks (seven of them from the NIH toolbox) in the Adolescent Brain Cognitive Development (ABCD) project (Thompson et al. 2019; Sripatha, Rutherford, et al. 2020). Based on these large-scale studies, researchers found robust associations ( $r = 0.2\text{--}0.4$ ) between GCA and functional as well as structural indices of the brain, and identified widely distributed brain regions associated with GCA (Barbey 2018; Duncan et al. 2020).

Despite advances in this area of research, several methodological issues have yet to be examined closely. First, although several studies have shown that tasks from different intelligence batteries provide a highly consistent measure of general intelligence (Johnson et al. 2004; Johnson et al. 2008), verification is still

needed to see whether GCA could be validly estimated from cognitive tasks that were not developed to measure intelligence but rather to characterize specific cognitive processes (Roznowski et al. 2000). Second, previous behavioral studies have shown that the number of tasks significantly impacts the GCA estimation, with larger test batteries producing better estimates of GCA (Major et al. 2011). However, large-scale neuroimaging studies have not systematically examined the effect of task number on the accuracy of GCA estimation using both behavioral and neural data. This issue may be particularly important when cognitive tasks have relatively low reliability (Enkavi et al. 2019).

Third, it is still unclear how the choice of factor models affects the estimation of GCA and its neural correlates. So far, two major factor models have been used to estimate GCA: the first is based on the Spearman model to extract a single  $g$  factor, and the second incorporates a bifactor model with group factors (Jensen 1998). Jensen (Jensen and Weng 1994) initially named these two GCAs as “Spearman’s  $g$ ” and “psychometric  $g$ ,” respectively, and warned that the Spearman’s  $g$  could be biased (Humphreys 1989), but further studies have been inconclusive. Some studies show highly similar results between the two methods (Ree and Earles 1991; Jensen and Weng 1994), whereas other studies suggest that task loadings on the general factor are significantly influenced by the type of factor model (Floyd et al. 2009; Major et al. 2011). In the large-scale imaging studies, the Spearman model has been used to analyze the UK Biobank data (Cox et al. 2019) and the bifactor model has been fitted to the HCP dataset (Dubois et al. 2018; Sripada, Angstadt, et al. 2020; Stammen et al. 2023) and the ABCD dataset (Thompson et al. 2019; Sripada, Rutherford, et al. 2020). However, no imaging study has systematically examined how the choice of factor model impacts the estimation of GCA and its neural correlates.

Finally, it is still being debated as to which type of brain data could provide a better understanding of the neural mechanisms of GCA. Earlier studies utilized brain imaging data (e.g. Duncan 2000; Jung and Haier 2007; Rushton and Ankney 2009) or brain lesion data (Gläscher et al. 2009) to localize GCA to specific brain regions, but recent studies suggest that intelligence relies on the integrated functions of brain networks (Barbey 2018). In addition, while the majority of studies have used resting state functional data to predict GCA (Finn et al. 2015; Dubois et al. 2018), there is emerging evidence that the brain connectivity patterns based on task fMRI data might provide better prediction of individuals’ cognitive abilities (Greene et al. 2018; Jiang et al. 2020; Sripada, Angstadt, et al. 2020). Given that these studies are mainly based on Western populations, more studies from diverse demographic samples are required to address this important question.

To summarize, the current study aimed to further examine the neural substrates of GCA by systematically examining the roles of the three factors discussed above: the number of tasks, the choice of factor model, and the type of brain data. We collected a large dataset from young healthy Chinese adults ( $n = 1605$ , male = 667, mean age = 20.8 years), who performed 19 cognitive tasks as well as RAPM. Two types of brain imaging data (resting state and  $n$ -back task) were collected from a subsample ( $n = 683$ , male = 263, mean age = 21.0). Four sets of analyses were conducted. First, we compared the modeling fit and GCA–brain correlation between the two types of factor models: the Spearman model and the bifactor model. Second, we systematically compared task vs. resting fMRI and different preprocessing parameters on the prediction of GCA. Third, we estimated GCA by randomly sampling different numbers of tasks to examine whether more tasks would enhance the reliability and validity of GCA estimation, as well as greater and more reliable brain–GCA association. Finally, using

the most reliable estimate of GCA and the best neural data and preprocessing parameters obtained by an independent task (i.e. RAMP), we carefully examined the neural correlates of GCA. Our results suggest that the bifactor model from a larger number of cognitive tasks could yield reliable estimation of GCA and that task fMRI yielded stronger predictability and more interpretable neural correlates than did resting-state fMRI. Our results not only contribute to a better understanding of the neural correlates of GCA but also have significant methodological implications for future studies.

## Method

### Participants

The participants were from the Cognitive Neurogenetic Study of Chinese Young Adults Project (Feng et al. 2020, 2022), in which more than 2,500 Han Chinese young adults from Beijing and Chongqing were recruited and 2,236 of them had behavioral data. In this analysis, we only included participants with high-quality data, which were defined as (i) the correct rate for each task was above chance level (Feng et al. 2022) (see Behavioral Tasks Description in the [Supplementary Text](#) for some special treatments for certain tasks), and (ii) the performance score was not classified as an outlier based on the boxplot method (Tukey 1977). Based on the above criteria, 631 participants who had more than 20% missing data, i.e. 4 invalid/missing scores out of the 20 scores were excluded from final analysis. Of the 1,605 participants (male = 667; mean age = 20.8, SD = 2.1, range: [16.8, 30.8]) who had high-quality behavioral data, 682 (male = 263; mean age = 21.0, SD = 2.2, range: [17, 29]) also had high-quality resting-state and task-state fMRI data. All participants gave written consent to the study and were paid for their participation. This study was approved by the Institutional Review Boards of Beijing Normal University and Southwest University, China.

### Procedures and materials

#### Tasks used to measure GCA

We chose 19 commonly used tasks to tap a wide range of cognitive abilities, including working memory (WM), complex spans, attention, episodic memory, inhibition, shifting, and reaction time. One of these tasks (i.e. Penn Continuous Performance Test [PCPT]) produced two indicators (i.e. median reaction time and sensitivity index ( $d'$ )), resulting in a total of 20 task performance indicators. Additionally, the classic intelligence measure, RAPM, was utilized for validation purposes.

Specifically, six tasks from WM domain were orientation change detection task, letter 3-back task, spatial 2-back task, keep track task, operation complex span task, and symmetry complex span task (SSPAN); three tasks from inhibition domain were anti-saccade task, stop signal task, and Stroop task; three tasks from shifting domain were number-letter switching task, color-shape switching task, and size-life switching task (SLST); three tasks from episodic memory domain were face–name associative memory task (FN), Korean symbol recognition task (KS), and mnemonic similarity task; four tasks from speed and attention domain were simple Reaction Time, choice Reaction Time (CRT), PCPT, and Penn line orientation test. Detailed descriptions of all these tasks can be found in [Supplementary Text](#).

#### $N$ -back task during MRI scanning

During the scanning session, the participants performed a digit  $n$ -back task, spanning from 0-back to 3-back. In the 0-back blocks, the participants were required to evaluate whether the displayed number was 7 or not. For the 1-back to 3-back blocks, they had to

determine whether the presented number was identical to that shown “ $n$ ” trials ago ( $n = 1, 2, \text{ or } 3$ ). The overall run comprised four blocks of each type of  $n$ -back trials, resulting in a total of 16 blocks.

### Experimental procedures

The participants performed all behavioral tasks in a computer lab. The experiments were conducted under the supervision of 15–20 experimenters, with each experimenter overseeing 2 participants. The test lasted ~6 h and was divided into a morning and an afternoon session of roughly equal length.

### The estimation of GCA score

The “lavaan” (0.6–15) package (Rosseel 2012) and “psych” (2.3.6) package (Revelle 2023) in R (4.3.1) software (R Core Team 2023) were used to estimate the GCA scores. We fitted two types of factor models of GCA, i.e. the Spearman model and the bifactor model (Barbey et al. 2021, p. 8). For the bifactor model, we split the whole data into two equivalent subsamples using the “SOLOMON” method (Lorenzo-Seva 2021) to avoid potential biases introduced in the splitting process. The first subsample was then used to explore the number of group factors and the structure of factors. Parallel analysis and empirical BIC criterion were used to determine the number of factors. Based on the results of the exploratory factor structure, we did a confirmatory factor analysis on the second subsample to evaluate the goodness of model fitting. Since  $\chi^2$  statistic is almost always statistically significant for data with a large sample size, we assessed model fit using the following fit indices: Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), and Tucker–Lewis Index (TLI). Values of RMSEA and SRMR less than 0.05 and values of CFI and TLI greater than 0.95 are considered good (Hu and Bentler 1999). Considering missing data and estimation efficiency (Enders and Bandalos 2001), we used a full-information maximum likelihood (FIML) estimator so that factor scores could be estimated for all participants in the CFA subsample. Then, we used the “regression” method to estimate the latent factor scores from the fitted model, yielding GCA scores for all participants in this subsample. After confirming the adequacy of the bifactor model, we refitted it to the total sample to generate GCA scores for all participants. In addition, because the widely accepted Cattell–Horn–Carroll (CHC) theory of intelligence utilizes a high-order hierarchical model (Carroll 1993), we also fitted a high-order hierarchical model (CHC model) as a comparison to the bifactor model.

To examine the effect of the number of tasks on the reliability of GCA score estimation, we randomly split the tasks into two sets and sampled 3–10 tasks in each set, estimated the GCA scores using the Spearman model described above, and examined the reliability of the two estimates using Pearson correlation. (The bifactor model was not used in this analysis because of the reduced number of tasks for each analysis, which makes it impractical to include group factors.) This procedure was repeated 100 times. After that, we fitted the relationships between the magnitude of correlations and the number of tasks with a BoxBOD regression model (Box et al. 2005, Chapter 10), a curve with a saturation value of  $\beta_1$  and a slope value of  $\beta_2$

$$y = \beta_1 (1 - e^{-\beta_2 x})$$

fitted curve was used to predict the correlations if more tasks had been included.

## MRI data collection and processing

### MRI data acquisition

Neuroimaging data were collected using 3.0 T Siemens MRI Trio scanners at the Brain Imaging Centers of Beijing Normal University and Southwest University. During resting state scanning, participants were instructed to close their eyes and avoid engaging in any specific thoughts. During task state scanning, they were asked to perform an  $n$ -back task (see Behavioral task section for more details).

We used a one-shot  $T_2^*$ -weighted gradient-echo, echo-planar imaging (EPI) sequence for functional scanning. Parameters for Beijing and Chongqing sites were generally equivalent to each other. For resting state, the following parameters were used: repetition time = 2,000 ms; echo time = 30 ms; flip angle = 90°; field of view = 200 × 200 mm<sup>2</sup> (Beijing) and 220 × 220 mm<sup>2</sup> (Chongqing); 64 × 64 matrix size with a resolution of 3.1 × 3.1 mm<sup>2</sup> (Beijing) and 3.4 × 3.4 mm<sup>2</sup> (Chongqing), 3.0 mm (Beijing) and 3.5 mm (Chongqing) transverse slices. A total of 200 (Beijing) and 242 (Chongqing) brain volumes (time points) were acquired.

For the  $n$ -back task, the parameters were the same as those for the resting state scan except the in-plane resolution, i.e. 64 × 64 matrix size with a resolution of 3 × 3 mm<sup>2</sup> (Beijing) and 3.4 × 3.4 mm<sup>2</sup> (Chongqing). The slice thickness was 3.0 mm in both sites. A total of 226 brain volumes were acquired in both sites.

### fMRI data preprocessing and network construction

We used GRETNA tools (Wang et al. 2015) and the AFNI software (Cox 1996) to preprocess the fMRI data according to the standard steps, including deleting the first 10 EPI volumes, slice-timing correction, realigning, normalization, adjusting for the nuisance covariates, and removing linear trends using temporal filters (a band-pass of 0.01–0.1 Hz) in a single regression model. The included nuisance covariates were the global signal, the average signal of the white matter and the cerebrospinal fluid, and the 24 motion parameters. More detailed information about the imaging preprocessing can be found in our previous study (Feng et al. 2022). All the subsequent analyses and visualizations were performed in MATLAB (The MathWorks Inc. 2023), R, and Python (Python Core Team 2019).

### Functional parcellation and network construction

The Shen’s 268-node (Shen268) atlas (Shen et al. 2013) was applied to the preprocessed fMRI data. After parcellation, we calculated Pearson correlations between time series of each node-pair and the correlation coefficients were then Fisher transformed, resulting in two 268 × 268 connectivity matrices for each subject. We kept the upper triangular values only (i.e. the unique 35,778 edges) and standardized them within each participant to control for the differences in scanning parameters. We also included latent connectivity patterns, which integrated the  $n$ -back task and resting state conditions. Specifically, we extracted the first principal component from the two connectivity matrices of each edge generated from the task and resting state fMRI data, resulting in a latent connectivity pattern (McCormick et al. 2022). This method is preferred over directly averaging the two functional connectivity patterns because principal component analysis calculates weights for each pattern so as to provide a better summary of the two patterns. To avoid any bias introduced by scan durations, we truncated the Chongqing data by keeping the first 200 brain volumes only; and when comparing performances of different fMRI conditions, we also truncated the  $n$ -back task fMRI data by keeping the first 200 brain volumes. We also used Power’s

264-node (Power264) parcellation scheme (Power et al. 2011) and applied the same processing procedure to construct functional connectivity network in order to replicate the results.

For Shen268 atlas, we used the same spectral clustering algorithm to assign these 268 nodes to 8 canonical networks (Shen et al. 2013, 2017), and the subcortical–cerebellar network was further split into 3 networks (Noble et al. 2017), resulting in 10 networks. These networks are named based on their approximate correspondence to previously defined resting-state networks as medial frontal (MF), frontoparietal (FP), default mode (DMN), motor (Mot), visual I (VI), visual II (VII), visual association (VAs), limbic (Lim), subcortical (Sub), and cerebellum (CBL).

### Predicting GCA with functional connectivity patterns

Utilizing a connectome predictive modeling (CPM) protocol (Shen et al. 2017), we employed functional connectivity patterns to predict GCA score and investigated the underlying neural substrates. Briefly, we applied a 10-fold cross validation to predict GCA in the left-out participants. First, to select the edges, the strength of each edge was correlated with GCA with age, gender, and mean framewise displacement as covariates. This procedure generated the “positive correlated network” (Pos-Cor Networks) using the edges showing high positive correlations with GCA, and the “anti-correlated network” (Anti-Cor Networks) using the edges showing high negative correlations with GCA. Following Greene et al. (2018), we tested different thresholds in this step. Two different thresholding methods (“alpha” level of correlation and “sparsity”) and different levels ( $P < 0.05$ ,  $P < 0.01$ ,  $P < 0.005$ ,  $P < 0.001$ ; 1%, 2.5%, 5%, 10%) were included.

Next, separately for Pos-Cor and Anti-Cor Networks, the edge strengths were first standardized (to avoid biasing the test fold) across participants and then summed for each participant in the training folds, yielding two summary statistics for each participant in each fold. Linear regression models were utilized to evaluate the correlation between network strength and GCA in the training folds, and 3 such models were built: one for Pos-Cor networks, one for Anti-Cor networks, and one for both networks (i.e. two predictors, named as Combined model).

Finally, for participants in the testing fold, the edge strengths were also standardized based on the corresponding data in training fold and then the same summary statistics were calculated for each participant. These summarized statistics were submitted to the corresponding models to predict GCA for these participants.

Model performance was evaluated by Pearson correlation between the model-estimated GCA and true GCA. To improve the robustness of results, this procedure was repeated 20 times for each model and condition, and the mean performance scores (Fisher transformed before averaging and transformed back after averaging) are reported.

### The stability of brain–GCA association using dice similarity

To examine the stability of brain–GCA association in the CPM analysis, we examined whether the contributing edges are consistent across the two splits of tasks (see above). Because we did 10-fold cross-validation for each model for 20 times, there were a total of 200 edge selections. Consequently, the contributing edges were determined based on the probability (i.e. 0.5, 0.8, 0.95, 0.99) of being selected across the 200 selections. For example, if the probability is set at 0.95, then an edge must be selected at least 190 times to be included in the contributing edges. For simplicity,

we kept the threshold level at 0.01 during model construction in this analysis.

We then calculated the dice similarity (Feng et al. 2022) between the contributing edges across the two randomly split task sets. For the two sets of selected edges  $X$  and  $Y$ , dice similarity can be calculated with the following formula:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|},$$

where  $|X \cap Y|$  represents the number of overlapping edges between  $X$  and  $Y$ , and  $|X|$  and  $|Y|$  are the numbers of edges in  $X$  and  $Y$ , respectively. We calculated them separately for Pos-Cor Networks and Anti-Cor Networks and averaged across these two networks.

### The enrichment pattern of the contributing edges

To delineate the neural substrates of GCA, we next sought to understand the enrichment pattern of networks to which the contributing edges belong, as well as the brain regions where the connected nodes were located. These results for GCA were then compared with those for RAPM.

First, we sought to understand how the contributing edges were distributed among the 10 canonical networks. Considering that the number of nodes for networks differs significantly (ranging from 9 nodes in VII network to 51 nodes in Mot network, see Table S2 for the number of nodes in each network), we calculated the enrichment value for each pair of networks (Greene et al. 2018). Specifically, for each pair of networks, we calculated the proportion of contributing edges belonging to this pair to the number of whole contributing edges, then normalized it by the fraction of total edges belonging to this pair in the whole Shen268 atlas. Formally, for network pair A and B:

$$\text{enrich}_{A,B} = \frac{n_{A,B}/n_{\text{total}}}{E_{A,B}/E_{\text{total}}},$$

where  $\text{enrich}_{A,B}$  is the enrichment value of network pair A and B,  $n_{A,B}$  is the number of contributing edges between network A and B,  $n_{\text{total}}$  is the total number of contributing edges,  $E_{A,B}$  is the number of edges between network A and B in the atlas, and  $E_{\text{total}}$  is the total number of edges in the atlas. The results are visualized as matrices with upper triangle in Fig. 5A.

Next, we quantified that the number of times each node was connected by the contributing edges, separately for GCA and RAPM, resulting in 2 (i.e. Pos-Cor and Anti-Cor Networks) edge counts vectors of length 268. We visualized them as markers in a glass brain by “nilearn” package (Abraham et al. 2014) in Python (Fig. 5B), with darker color representing a larger number of edges. Considering that the edge counts vectors were not normally distributed, we calculated Spearman correlation coefficients to estimate the similarity of edge counts vectors between GCA and RAPM.

## Results

### Descriptive statistics, correlations between task indicators, and factor models

Descriptive statistics of the 20 indicators of task performance are provided in Table S1. The scores pertaining to reaction times (such as median reaction time, inverse efficiency score, etc., see Methods for details of each task) were transformed by reversing the sign (e.g.  $-1$  to  $1$  and vice versa), so that higher

**Table 1.** Main fit indices for factor analysis models involving all 20 task indicators.

Model	$\chi^2$	df	RMSEA	SRMR	CFI	TLI
Spearman	1,616.8	170	0.07	0.07	0.64	0.60
Bifactor	257.4	150	0.03	0.03	0.95	0.93
CHC	472.5	166	0.05	0.05	0.85	0.83

Notes:  $\chi^2$ : chi square, df: degree of freedom, RMSEA: Root Mean Square Error of Approximation, SRMR: Standardized Root Mean Square Residual, CFI: Comparative Fit Index, TLI: Tucker–Lewis Index.

scores indicate better performance. All task indicators exhibited normal distribution (Fig. S1) and their internal consistency reliabilities ranged from satisfactory to high (0.62 to 0.98) except that of SLST ( $\alpha = 0.53$ ) (Table S1).

Considering the missing data (see Fig. S2 for visualization of the missing pattern), we presented pairwise correlations among task indicators. The 190 correlations had a mean of 0.12 and ranged from  $-0.06$  to  $0.50$ , and only six of them (3%) were slightly but not significantly smaller than 0, indicating the presence of a “positive manifold” (Fig. 1A) (Spearman 1904). Based on this, we fitted a one- $g$ -factor model with no group factors (i.e. Spearman model) using the FIML method, which allowed us to derive the factor scores of the latent factor as GCA scores using regression-based weights. The FIML method also allowed us to attribute GCA scores to participants with missing data. The GCA factor from the Spearman model, namely “GCA-S,” accounted for a substantial proportion, i.e. 71.1% ( $\omega_h$ ) (McDonald 1999; Zinbarg et al. 2006), of the total variance (Fig. 1C). However, the fit indices for the above Spearman model were not ideal (first row of Table 1): RMSEA = 0.07, SRMR = 0.07, CFI = 0.64, TLI = 0.60.

To test the bifactor model, we utilized empirical BIC criteria to explore potential underlying structure with group factors for all 20 task indicators. The exploratory factor analysis on the first subsample demonstrated that a four-factor structure fit the data well (Fig. 1B). Subsequently, the confirmatory bifactor analysis on the second subsample using the same structure from exploratory factor analysis revealed that a model comprising a general factor, i.e. “GCA-B,” and four group factors provided a satisfactory fit to the data (second row of Table 1, see visualization in Fig. 1D). These four factors were identified and interpreted as follows: processing speed (Speed), WM, learning and memory (Memory), and mental shifting (Shifting). Using the bifactor model, the fitting parameters improved significantly: RMSEA = 0.03, SRMR = 0.03, CFI = 0.95, and TLI = 0.93. The GCA factor, represented by “GCA-B,” accounted for 55.1% ( $\omega_h$ ) of the total variance across all the task indicators. Notably, the correlation between GCA-S scores and GCA-B scores was 0.939, suggesting that different factor models provide highly consistent estimation of GCA.

As a comparison to the bifactor model, an alternative hierarchical model (the CHC high-order hierarchical model) was run using the same grouping structure as the bifactor model. The CHC model also showed better fit indices than did the Spearman model (third row of Table 1, see visualization in Fig. S3), but it showed significantly poorer than did the bifactor model: RMSEA = 0.05, SRMR = 0.05, CFI = 0.85, and TLI = 0.83. The  $g$  factor of the CHC model accounted for 60.7% ( $\omega_h$ ) of the total variance across all the task indicators. The correlation between the  $g$  factor scores based on this model and GCA-B was 0.947, demonstrating again the consistency among different GCA estimation methods. Considering that the bifactor model and the CHC model are both hierarchical factor models and that the bifactor model showed better fit indices, we reported the results from the bifactor model only in the rest of the article to avoid redundancy.

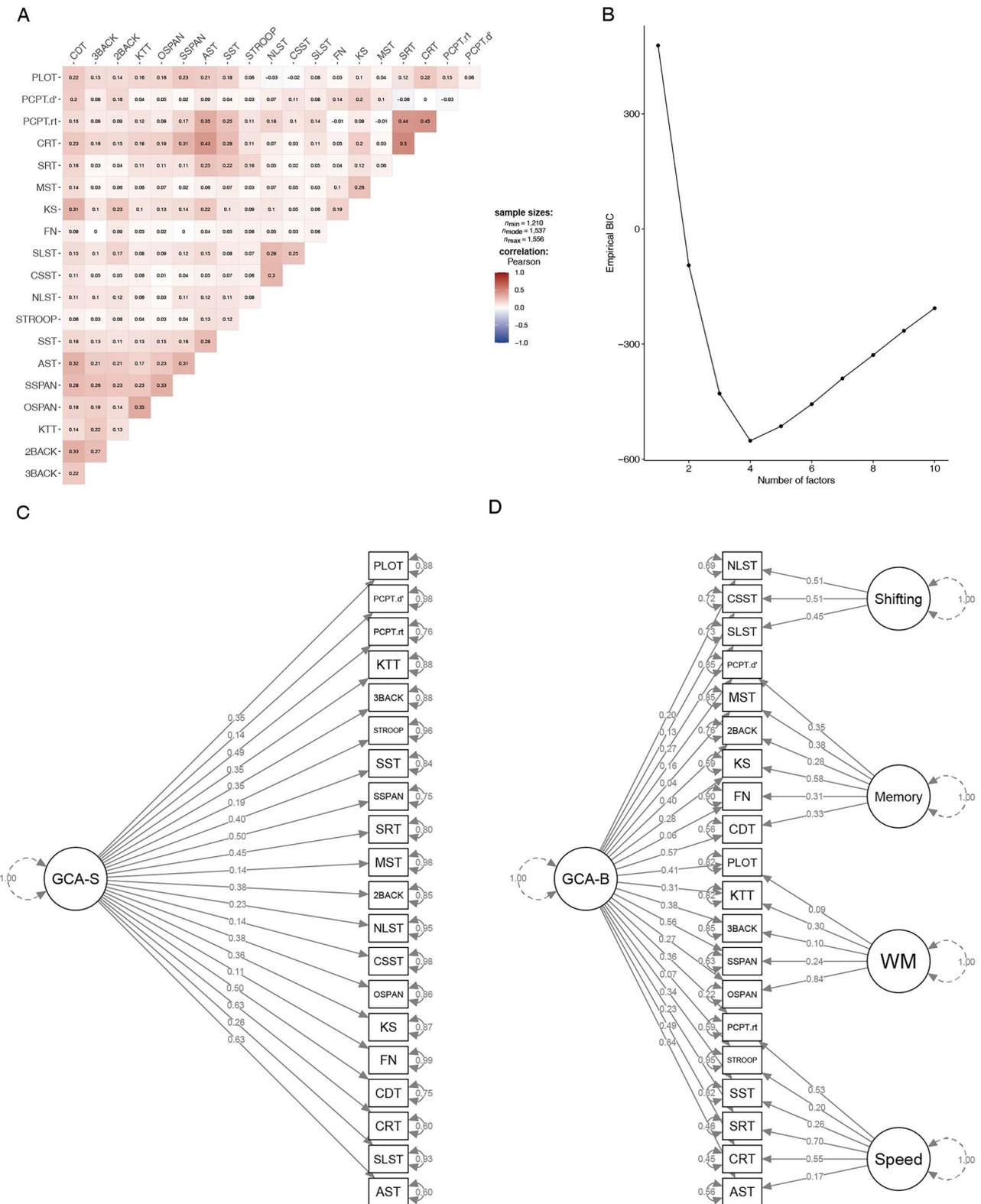
## Improved estimation of GCA with more tasks

The above analysis provides compelling evidence that the selected tasks are well-suited for the  $g$  factor model. Next, we investigated whether incorporating more tasks yielded better estimation of GCA, i.e. higher reliability and validity. To assess the reliability, we randomly split the task indicators into two sets and estimated GCA in each set with varying numbers of tasks, ranging from 3 to 10. The results revealed that the median Pearson correlation coefficient between the GCAs of the two split samples showed a remarkable increase from 0.23 (SD = 0.114) with three tasks to 0.57 (SD = 0.073) with 10 tasks, indicating an average gain of 0.045 for each additional task in the estimation (Fig. 2A). Using a BoxBOD regression model, we found that with about 40 tasks, we can achieve a peak reliability of 0.8.

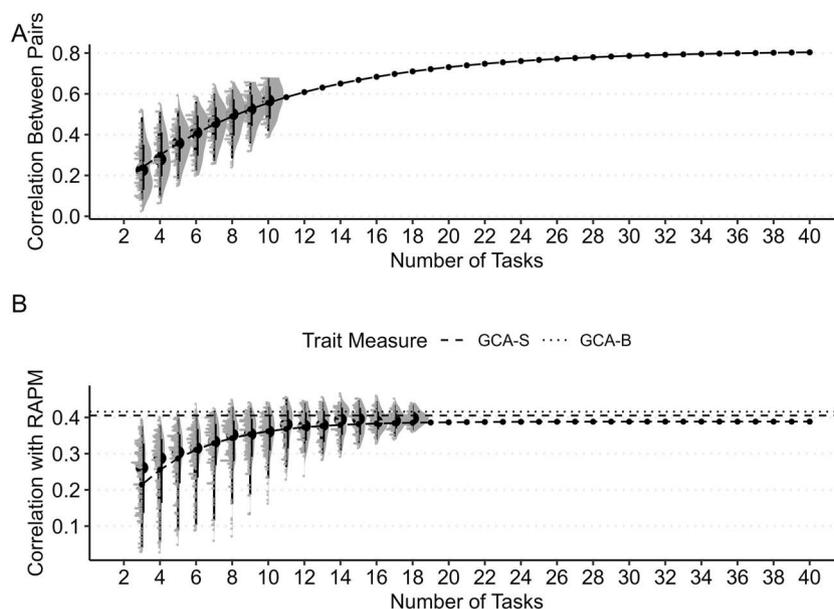
To examine the validity of GCA estimate, we investigated whether the GCA estimated from more tasks would show increased correlation with the RAPM performance. We generated the GCA-S scores based on 3–18 task indicators, each with 100 times of random selections. The median correlation coefficient between the GCA-S scores and RAPM performance increased from 0.25 (SD = 0.091) with 3 tasks to 0.36 (SD = 0.054) with 10 tasks, and to 0.39 (SD = 0.020) with 18 tasks. When all 20 task indicators were used, the correlations were 0.40 between GCA-S and RAPM and 0.42 between GCA-B and RAPM (Fig. 2B). Using a BoxBOD regression model, we found that the predicted peak correlation between GCA and RAPM was 0.39. In other words, there appear to have little additional gains in the strength of GCA’s association with RAPM beyond 18 tasks. Such a modest level of association seems to suggest that GCA and RAPM reflect overlapping but different cognitive functions.

In addition to using RAPM as the criterion variable, we also conducted additional analyses using the GCA estimated from all 20 task indicators (both GCA-S and GCA-B) as the criterion variables. The correlations between GCA scores from select task indicators and those from the total 20 task indicators increased steadily with increasing number of task indicators (Fig. S4). The correlations with GCA-S of 20 task indicators were generally higher than those with GCA-B, replicating the pattern found with RAPM as the criterion variable. Finally and not surprisingly, these correlations (relying on the same pool of task indicators) were also higher than those between GCA scores from select task indicators and RAPM (which was an independent criterion variable).

It should be noted that the above analyses did not consider factor loadings when selecting task indicators. Tasks such as CRT, anti-saccade, and symmetry complex span had the highest loadings on GCA (Fig. S5A). Indeed, if we used the 5 tasks with the highest loadings, they would account for 84.5% of the total variance of GCA, whereas if we used the 5 tasks with the lowest loadings, they would account for only 7.5% of the total variance of GCA (Fig. S5B). In our analyses using the top 10 tasks (with the highest factor loadings), we achieved an averaged reliability of 0.52 with just 5 of those tasks, whereas using the bottom 10



**Fig. 1.** Latent structure of all 20 task indicators. A) Correlation matrix of all task indicators. Darker color indicates larger differences from 0. A pairwise method was used because of missing values. The meaning of task abbreviations can be found in [Methods](#) section. We included two indicators for the PCPT task and used “.” to separate the task name and the indicator name. B) Empirical BIC for different numbers of factors suggested a 4-factor solution. C) Structure of the Spearman model (i.e. no group factors). D) Structure of the bifactor model, with a general factor, i.e. GCA-B, and four group factors, i.e. speed: processing speed; WM: working memory; memory: learning and memory; shifting: mental shifting.



**Fig. 2.** The influence of the number of tasks on GCA estimation. Results were based on the 100 times resampling, and the black dots with long-dashed line were the predictions based on BoxBOD regression. A) Each gray dot represents the Pearson correlation between the GCA scores estimated from each pair of resampled tasks. B) Each gray dot represents the Pearson correlation between the GCA-S score and RAPM. For 3–10 task indicators, there were 200 dots because of the paired sampling. For 11 or more task indicators, there were 100 dots. Horizontal dashed lines represent the correlations between the two types of GCA scores and RAPM scores. GCA-B: GCA estimated from all task indicators by the bifactor model; GCA-S: GCA estimated from all task indicators by the Spearman model.

tasks, we could only achieve an averaged reliability of 0.21 with 5 tasks (Fig. S5C).

In sum, these findings suggest that the inclusion of more tasks would lead to significantly enhanced reliability and validity of GCA estimation and that using tasks with high factor loadings on the  $g$  factor would also improve the validity of GCA estimation.

### Determining the best parameters for brain–GCA association

The above analyses based on behavioral data showed that more tasks led to more reliable and valid estimation of GCA. In the following analyses, we further examined this issue with neural data. That is, if the inclusion of more cognitive tasks could increase the validity and reliability of GCA, we should find greater, more reliable, and more interpretable brain–GCA associations.

First, because the method we employed to measure brain–GCA associations (i.e. connectome-based predictive modeling or CPM) involves two different threshold methods (“alpha” and “sparsity”), we first needed to determine the best method and threshold (see Methods). We used an independent behavioral index (i.e. RAPM) for this purpose. The results indicated that the two threshold methods were generally comparable to each other and that the predictive performance improved slightly when using a less conservative threshold (i.e. more edges included) (Fig. 3A). Considering both predictive performance and the conventions in this field, we ultimately chose the “alpha” method with a level of  $P < 0.01$ .

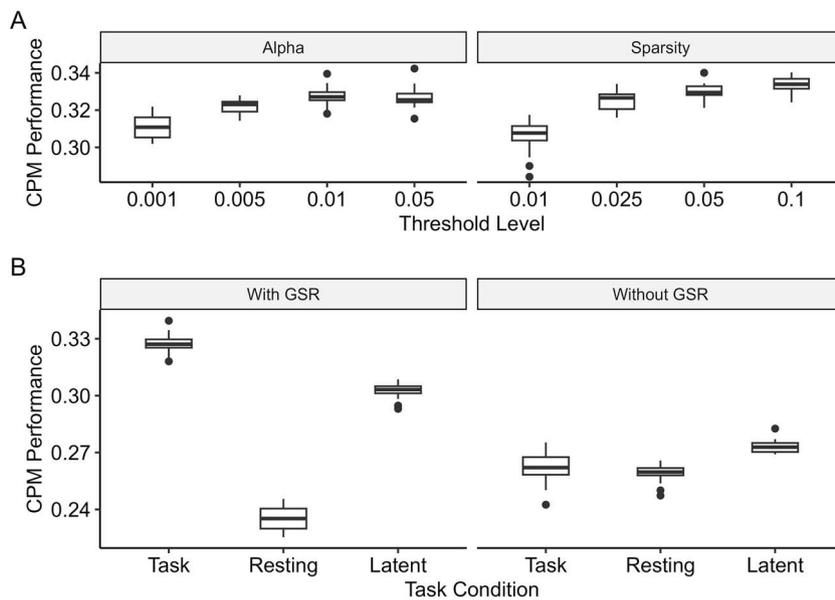
Second, we used RAPM to determine the best neural data and processing parameters (i.e. global signal regression or GSR) that could predict cognitive performance. Previous studies have shown that both resting-state functional connectivity (FC) and task-state FC could predict domain-specific and general cognitive abilities, with generally higher prediction from task-state FC. We aimed to replicate these findings with our data. Besides task-state and resting-state FC, we also estimated the latent FC pattern (i.e. labeled as the latent condition), which integrated information

from both task and resting fMRI (see Methods). We found that the  $n$ -back task outperformed all other conditions, including the latent condition (Fig. 3B). In addition, GSR generally enhanced the performance of models (Fig. 3B). These results were replicated when using Power’s 264-node atlas, although the average correlations based on Power’s 264-node atlas were lower than those based on the Shen268 atlas (Fig. S6). Based on these results, the following analysis will use  $n$ -back fMRI data with GSR and the Shen268 atlas, with an “alpha” threshold of  $P < 0.01$ , except otherwise noted.

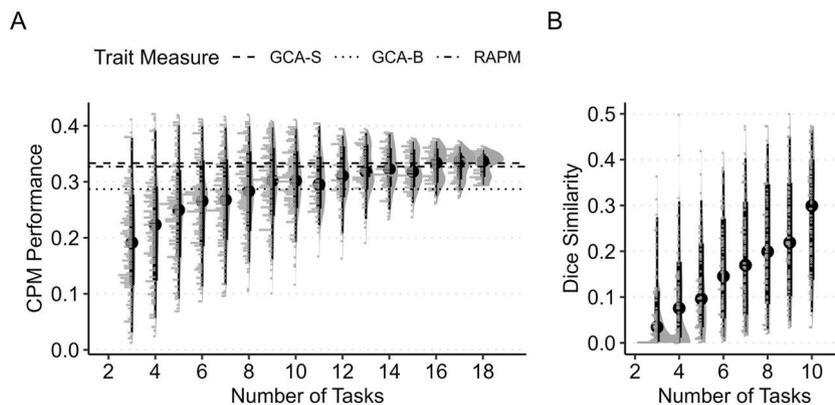
### More tasks were associated with greater and more reliable brain–GCA association

For each sample of the 3–18 task indicators, we evaluated the neural predictability of the GCA scores using CPM. The results suggest that as more task indicators were used to estimate the GCA-S (Fig. 4A), the median scores increased from 0.19 (SD = 0.09, ranging from 0.02 to 0.41) with 3 task indicators, to 0.30 (SD = 0.06, ranging from 0.12 to 0.41) with 10 task indicators, to 0.34 (SD = 0.02, ranging from 0.29 to 0.36) with 18 task indicators. Notably, the GCA-S scores estimated from all task indicators showed comparable predictability of the RAPM score, while GCA-B scores exhibited relatively lower predictability ( $r_{\text{GCA-S}} = 0.333$ ,  $r_{\text{GCA-B}} = 0.287$ ,  $r_{\text{RAPM}} = 0.327$ ).

Furthermore, we examined the reliability of brain–GCA association by calculating the dice similarity of the “contributing edges” from the two sets of tasks. For each set, our CPM protocol generated a total of 200 edge-selection results (see Methods). The contributing edges were defined as those that were selected 190 times (i.e.  $P$  of 0.95) or above across the 200 selections. As predicted, the dice similarity also increased with more tasks. Specifically, the median dice similarity increased from 0.03 (SD = 0.08) with 3 task indicators to 0.30 (SD = 0.12) with 10 task indicators. These results were also replicated when using different brain parcellation schemes and different edge selection criteria (Fig. S7).



**Fig. 3.** Determining the best CPM hyperparameters and fMRI data parameters for brain-GCA association using RAPM. Results were based on 20 times cross-validation using the Shen268 atlas. A) Boxplot of the CPM performance with different feature selection threshold hyperparameters using functional connectivity patterns from task-state with GSR. B) Boxplot of the CPM performance with different task conditions and different GSR manipulations. Results were based on the “alpha” threshold of  $P < 0.01$ .



**Fig. 4.** More tasks were associated with greater and more reliable brain-GCA association. A) Distribution of brain-GCA association based on different numbers of task indicators. The horizontal lines represent the CPM performance for the two types of GCA scores and RAPM scores. B) The degree of overlapping contributing edges (dice similarity) between pairs of models with different numbers of task indicators. The dice similarity was averaged across Pos-Cor networks and Anti-Cor networks (see [Methods](#)).

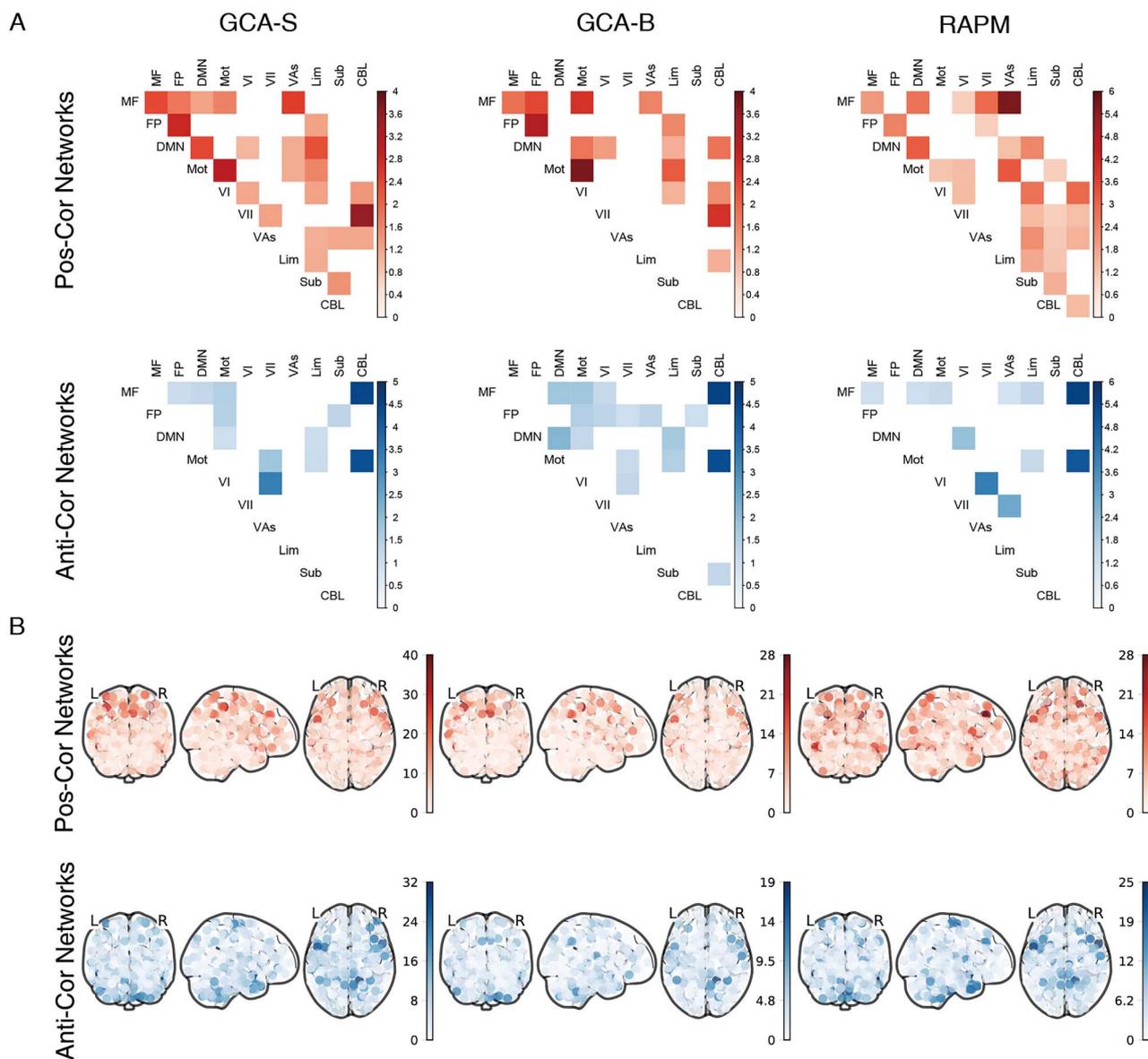
### Comparing the neural substrates of GCA between different models

Having identified the best parameters for brain-GCA association and shown that more tasks led to greater and more reliable brain-GCA association, we then included all tasks and used the best parameters to examine the neural correlates of GCA, by comparing the Spearman and bifactor models and relating these results to RAPM. Specifically, we inspected the “contributing edges” (i.e. edges that were selected 95% times) determined by the CPM scheme to understand how different brain networks contributed to GCA.

As there was a large correlation between GCA-S and GCA-B, the model-selected edges between them also largely overlapped (Pos-Cor Networks:  $DSC = 0.57$ , Anti-Cor Networks:  $DSC = 0.55$ ). However, although GCA-B showed a slightly larger correlation with RAPM than did GCA-S behaviorally, the overlap of selected edges between GCA-B and RAPM (Pos-Cor Networks:  $DSC = 0.15$ , Anti-Cor Networks:  $DSC = 0.24$ ) was smaller than that between GCA-S and RAPM (Pos-Cor Networks:  $DSC = 0.30$ , Anti-Cor Networks:  $DSC = 0.35$ ).

Additionally, utilizing 10 functional networks derived from previous studies ([Methods](#)), we inspected the enrichment patterns of the contributing edges within and across the canonical functional networks. We quantified the enrichment patterns of the contributing edges in a way that a value larger than 1 indicated a disproportionate (i.e. enriched) contribution of the network pair ([Methods](#)) ([Fig. 5A](#)). For Pos-Cor Networks, the top 10% overrepresented network pairs in the two GCA models were highly consistent, including those within the Mot network, the FP network, and between VII network and CBL network. In contrast, the RAPM involved the network pairs between the MF network and VAs network, between the Mot network and VAs network, and within DMN network. This indicates that the GCA models show a larger involvement of frontal regions and less involvement of visual regions than did the RAPM model.

For the Anti-Cor Networks, the three models showed very consistent top 10% overrepresented network pairs, including those between the MF network and CBL network, and between the Mot network and CBL network.



**Fig. 5.** Contributing edges were widely distributed among different networks and showed comparable results among three intelligence/GCA measures. A) Visualization of proportional contribution for each pair of assigned networks. See functional parcellation and network construction for full names of the networks. B) Visualization of edge counts distribution of all nodes with markers on the whole brain.

In another analysis, we examined the nodes that contributed to the prediction of GCA and RAMP (See *Methods*). The high similarity between models of GCA-S and GCA-B was again confirmed by the high correlation of edge counts of nodes (Fig. 5B) (Pos-Cor Networks:  $\rho = 0.79$ , Anti-Cor Networks:  $\rho = 0.81$ ). The nodes pattern was more similar between GCA-S and RAMP than between GCA-B and RAMP (Pos-Cor Networks:  $\rho_{GCA-S} = 0.49$ ,  $\rho_{GCA-B} = 0.27$ , one-tailed  $P_{diff} = 0.002$ ; Anti-Cor Networks:  $\rho_{GCA-S} = 0.54$ ,  $\rho_{GCA-B} = 0.42$ , one-tailed  $P_{diff} = 0.03$ ). Additionally, across the three measures, the hub nodes, i.e. nodes with the most contributing edges, for Pos-Networks were mostly from the frontal and parietal regions, while the hub nodes for Anti-Networks were mostly from the temporal and cerebellar regions (see Table S3 for top 10 nodes).

## Discussion

Since most cohort studies incorporate multiple cognitive tasks but not standard intelligence tests, researchers have tried to use GCA to index general intelligence by fitting factor models on various

cognitive tasks (Cox et al. 2019; Dubois et al. 2018; Thompson et al. 2019). In current research, we aimed to verify whether GCA is a valid index of general intelligence (which was measured by RAMP) and investigate the effects of the number of cognitive tasks and modeling methods on GCA estimation accuracy, using both behavioral and neural data.

Our results provide important evidence that the cognitive tasks that were designed to identify specific cognitive processes could measure GCA with significant reliability, but the accuracy of GCA estimation is influenced by number of tasks (an average reliability of 0.5 for 10 tasks, and an expected average reliability of 0.8 for 40 tasks). These results extend previous studies which have shown that a large battery of IQ test could lead to better measurement of  $g$  (Floyd et al. 2009; Major et al. 2011) and that greater cross-battery correlations has been found for batteries with more tasks (Johnson et al. 2004, 2008).

More importantly, by integrating neural data, we found that GCA estimates from more tasks showed greater and more reliable brain-GCA association results. This finding provides neural

evidence that good GCA estimation depends on a large battery size and that the reliability of GCA measurement affects the effect size (and its stability) of the brain–behavior association (Gignac and Bates 2017).

Moreover, using a highly accepted measure of general intelligence (i.e. RAPM) (Burke and Bingham 1969; Roth 2021), we found that criterion-related validity of the GCA estimates also increases as battery size increases. However, the highest correlations between GCA and RAPM are in the moderate level ( $r \approx 0.4$ ), suggesting RAPM and GCA might measure partially overlapping but also different aspects of human cognitive abilities (Gignac 2015). Consequently, we also used the GCA scores estimated from all 20 task indicators as a criterion to evaluate the validity of GCA estimates. As expected, the validity indices of GCA estimates in such analyses were generally higher than those based on RAPM (an independent measure) as the criterion variable. Importantly, in the new analysis, the validity indices increased as the battery size increased, just as in the analyses using RAPM as the criterion variable. One potential caveat of this result is that our task battery might not be large and diversified enough to achieve a close estimation of the true GCA.

In addition, although it is predicted that 40 tasks (somewhat impractical for an intelligence battery) are required to achieve a reliability of 0.8 when choosing tasks at random, the number of tasks required to achieve that level of reliability may decrease when we consider the factor loadings of the tasks. We found that some tasks, such as CRT, anti-saccade, and SSPAN, had higher factor loadings on GCA than other tasks and that the use of these tasks would yield more reliable estimation of GCA. Furthermore, like many other large-sample cohort study, the current study mainly chose tasks based on classic behavioral paradigms in cognitive neuroscience to examine their underlying neural correlates. These tasks were notably different from those used in the classical CHC theory of intelligence. Future studies should further examine how to optimize the task selections to better estimate GCA and its neural correlates.

We found that the two factor models (Spearman and bifactor) showed highly correlated GCA estimates and also highly overlapping neural correlates, extending previous behavioral studies (Ree and Earles 1991; Jensen and Weng 1994). We further found that compared with the Spearman model, the bifactor model showed better model fit and a larger correlation with RAPM. This is consistent with recent studies showing that general factor loadings are largely influenced by factor-extraction method (Floyd et al. 2009; Major et al. 2011) and suggests that the Spearman model introduces certain contamination when the correlation matrix of all variables has more than one common factor (Jensen and Weng 1994).

Another question of great interest is whether resting state or task fMRI has better prediction of GCA. Recent studies consistently found that task-induced brain state could better predict behavior trait than could resting-state (Greene et al. 2018; Jiang et al. 2020). This finding is consistent with the “treadmill test” hypothesis that just as a treadmill test is preferred when measuring cardiac function, cognitively demanding tasks activate the brain in a way that would yield systematic individual differences (Sripada, Angstadt, et al. 2020). Our results confirm that task-induced brain state shows better prediction of GCA. However, contrary to previous reports (e.g. Gao et al. 2019; McCormick et al. 2022), our study found that the integration of task-induced and resting state neural data did not improve the prediction of GCA. More research is needed to examine when and how the integration of different types of neural data can improve the brain–cognition correlations.

Using the best CPM parameters and functional connectivity data from the *n*-back test, we further examined the neural correlates of GCA. First, by examining the contributing edges of the two types of GCAs (GCA-B and GCA-S) and RAPM, we found a large overlap among them, consistent with Deary et al.’s (2010) claim that “biological associations studies with different measures, regardless of single high *g*-loaded task or derived from a battery of tests, are generally similar to each other.” Second, consistent with previous reports (Duncan 2000; Santarnecchi et al. 2017; Dubois et al. 2018; Greene et al. 2018), GCA-related networks are widely distributed across the whole brain, suggesting distributed neural correlates of GCA (Barbey 2018). Third, the hub regions revealed by the edge counts pattern are mainly located in the fronto-parietal regions, in agreement with the parieto-frontal integration theory (P-FIT) of intelligence (Jung and Haier 2007). Finally, there is also significant involvement of cerebellar regions, emphasizing its prominent role in cognitive functions (Yoon et al. 2017; King et al. 2019; Schmahmann 2019).

Besides the similarities of neural correlates of the three measures (GCA-B, GCA-S, and RAPM), we also revealed several major differences. First, the GCA estimated from the bifactor model (i.e. GCA-B) shows relatively poorer prediction performance than does either GCA-S or RAPM. One possibility is that, compared with the general factor in the Spearman model, the general factor in the bifactor model accounted for a smaller proportion of the total variance, resulting in decreased correlations with brain networks. Second, we found relatively larger contributions of frontal regions to GCA-S and GCA-B than to RAPM, which is consistent with findings about the key role of the frontal cortex in GCA (Duncan 2000, 2010). Finally, we found relatively larger contributions of the visual networks to RAPM than to GCA, which might be due to the fact that the RAPM task mainly involves visual patterns.

Several questions need to be further examined in future studies. First, our study used connectome-based predictive modeling (CPM) as the tool to predict behaviors, which is easy to use and facilitates interpretation of the modeling networks. Nevertheless, the cross-validation used by this method might introduce some bias because the GCA scores were first estimated based on all participants (rather than only on the training and testing data separately), which may lead to inflated prediction performance. Additionally, the CPM method does not take graph-theory-based measures into account, thus cannot examine how network properties (e.g. small-world) correlate differently among the measures (Barbey 2018). In addition, although we included a relatively large number of tasks in this study, our results suggest that they are not sufficient to yield highly reliable GCA estimates. As predicted in Fig. 2A and mentioned earlier, the correlation between test batteries would go up to 0.8 with 40 or more tasks. Future work could use a larger number of cognitive tasks from more cognitive domains to generate more reliable GCA estimates and to further examine the underlying neural correlates.

To conclude, our results demonstrated that GCA can be reliably estimated by multiple cognitive tasks and that the quality of measurement can be enhanced by adding more tasks and fitting the bifactor model rather than the Spearman model. These results have significant theoretical and methodological implications for our understanding of the neural correlates of GCA.

## Author contributions

Liang Zhang (Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Software, Visualization, Writing—original draft, Writing—review & editing), Junjiao Feng (Data curation, Formal analysis, Methodology, Software), Chuqi

Liu (Conceptualization, Data curation, Validation), Huinan Hu (Conceptualization), Yu Zhou (Conceptualization, Validation), Gangyao Yang (Conceptualization, Validation), Xiaojing Peng (Conceptualization, Validation), Tong Li (Conceptualization, Validation), Chuansheng Chen (Conceptualization, Investigation, Supervision, Writing—review & editing) and Gui Xue (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing)

## Supplementary material

Supplementary material is available at *Cerebral Cortex* online.

## Funding

This work was funded by the STI 2030—Major Projects 2021ZD0200500, the National Natural Science Foundation of China (32330039), and the Sino-German Collaborative Research Project “Cross-Modal Learning” (NSFC 62061136001/DFG TRR169).

Conflict of interest statement. None declared.

## References

- Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*. 2014;8. <https://doi.org/10.3389/fninf.2014.00014>.
- Barbey AK. Network neuroscience theory of human intelligence. *Trends Cogn Sci*. 2018;22(1):8–20. <https://doi.org/10.1016/j.tics.2017.10.001>.
- Barbey AK, Karama S, Haier RJ. *The Cambridge handbook of intelligence and cognitive neuroscience*. 1st ed. Cambridge, New York: Cambridge University Press; 2021. <https://doi.org/10.1017/9781108635462>.
- Box GEP, Hunter JS, Hunter WG. *Statistics for experimenters: design, innovation, and discovery*. 2nd ed. Hoboken, NJ: Wiley-Interscience; 2005.
- Burke HR, Bingham WC. Raven's Progressive Matrices: More on Construct Validity. *The Journal of Psychology*. 1969;72(2):247–251. <https://doi.org/10.1080/00223980.1969.10543505>.
- Carroll JB. A theory of cognitive abilities: the three-stratum theory. In: *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge, New York: Cambridge University Press; 1993. pp. 631–655.
- Cole MW, Yarkoni T, Repovs G, Anticevic A, Braver TS. Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *J Neurosci*. 2012;32(26):8988–8999. <https://doi.org/10.1523/JNEUROSCI.0536-12.2012>.
- Colom R, Karama S, Jung RE, Haier RJ. Human intelligence and brain networks. *Dialogues Clin Neurosci*. 2010;12(4):489–501. <https://doi.org/10.31887/DCNS.2010.12.4/rcolom>.
- Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance Neuroimages. *Comput Biomed Res*. 1996;29(3):162–173. <https://doi.org/10.1006/cbmr.1996.0014>.
- Cox SR, Ritchie SJ, Fawns-Ritchie C, Tucker-Drob EM, Deary IJ. Structural brain imaging correlates of general intelligence in UK biobank. *Intelligence*. 2019;76:101376. <https://doi.org/10.1016/j.intell.2019.101376>.
- Deary IJ, Penke L, Johnson W. The neuroscience of human intelligence differences. *Nat Rev Neurosci*. 2010;11(3):201–211. <https://doi.org/10.1038/nrn2793>.
- Dubois J, Galdi P, Paul LK, Adolphs R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philos Trans R Soc Lond B Biol Sci*. 2018;373(1756):20170284. <https://doi.org/10.1098/rstb.2017.0284>.
- Duncan J. A neural basis for general intelligence. *Science*. 2000;289(5478):457–460. <https://doi.org/10.1126/science.289.5478.457>.
- Duncan J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn Sci*. 2010;14(4):172–179. <https://doi.org/10.1016/j.tics.2010.01.004>.
- Duncan J, Assem M, Shashidhara S. Integrated intelligence from distributed brain activity. *Trends Cogn Sci*. 2020;24(10):838–852. <https://doi.org/10.1016/j.tics.2020.06.012>.
- Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Struct Equ Model Multidiscip J*. 2001;8(3):430–457. [https://doi.org/10.1207/S15328007SEM0803\\_5](https://doi.org/10.1207/S15328007SEM0803_5).
- Enkavi AZ, Eisenberg IW, Bissett PG, Mazza GL, MacKinnon DP, Marsch LA, Poldrack RA. Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proc Natl Acad Sci*. 2019;116(12):5472–5477. <https://doi.org/10.1073/pnas.1818430116>.
- Feng J, Chen C, Cai Y, Ye Z, Feng K, Liu J, Zhang L, Yang Q, Li A, Sheng J, et al. Partitioning heritability analyses unveil the genetic architecture of human brain multidimensional functional connectivity patterns. *Hum Brain Mapp*. 2020;41(12):3305–3317. <https://doi.org/10.1002/hbm.25018>.
- Feng J, Zhang L, Chen C, Sheng J, Ye Z, Feng K, Liu J, Cai Y, Zhu B, Yu Z, et al. A cognitive neurogenetic approach to uncovering the structure of executive functions. *Nat Commun*. 2022;13(1):4588. <https://doi.org/10.1038/s41467-022-32383-0>.
- Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci*. 2015;18(11):1664–1671. <https://doi.org/10.1038/nn.4135>.
- Flores-Mendoza C, Ardila R, Gallegos M, Reategui-Colareta N. General intelligence and socioeconomic status as strong predictors of student performance in Latin American schools: evidence from PISA items. *Front Educ*. 2021;6:632289. <https://doi.org/10.3389/educ.2021.632289>.
- Floyd RG, Shands EI, Rafael FA, Bergeron R, McGrew KS. The dependability of general-factor loadings: the effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence*. 2009;37(5):453–465. <https://doi.org/10.1016/j.intell.2009.05.003>.
- Gao S, Greene AS, Constable RT, Scheinost D. Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage*. 2019;201:116038. <https://doi.org/10.1016/j.neuroimage.2019.116038>.
- Gignac GE. Raven's is not a pure measure of general intelligence: implications for g factor theory and the brief measurement of g. *Intelligence*. 2015;52:71–79. <https://doi.org/10.1016/j.intell.2015.07.006>.
- Gignac GE, Bates TC. Brain volume and intelligence: the moderating role of intelligence measurement quality. *Intelligence*. 2017;64:18–29. <https://doi.org/10.1016/j.intell.2017.06.004>.
- Gläscher J, Tranel D, Paul LK, Rudrauf D, Rorden C, Hornaday A, Grabowski T, Damasio H, Adolphs R. Lesion mapping of cognitive abilities linked to intelligence. *Neuron*. 2009;61(5):681–691. <https://doi.org/10.1016/j.neuron.2009.01.026>.
- Gottfredson LS, Deary IJ. Intelligence predicts health and longevity, but why? *Curr Dir Psychol Sci*. 2004;13(1):1–4. <https://doi.org/10.1111/j.0963-7214.2004.01301001.x>.
- Greene AS, Gao S, Scheinost D, Constable RT. Task-induced brain state manipulation improves prediction of individual

- traits. *Nat Commun.* 2018;9(1):2807. <https://doi.org/10.1038/s41467-018-04920-3>.
- Gustafsson J-E, Balke G. General and specific abilities as predictors of school achievement. *Multivar Behav Res.* 1993;28(4):407–434. [https://doi.org/10.1207/s15327906mbr2804\\_2](https://doi.org/10.1207/s15327906mbr2804_2).
- Haier RJ, White NS, Alkire MT. Individual differences in general intelligence correlate with brain function during nonreasoning tasks. *Intelligence.* 2003;31(5):429–441. [https://doi.org/10.1016/S0160-2896\(03\)00025-4](https://doi.org/10.1016/S0160-2896(03)00025-4).
- Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J.* 1999;6(1):1–55. <https://doi.org/10.1080/10705519909540118>.
- Humphreys LG. The first factor extracted is an unreliable estimate of Spearman's "g": the case of discrimination reaction time. *Intelligence.* 1989;13(4):319–323. [https://doi.org/10.1016/S0160-2896\(89\)80003-0](https://doi.org/10.1016/S0160-2896(89)80003-0).
- Jensen AR. *The g factor: the science of mental ability.* Westport, Conn: Praeger; 1998.
- Jensen AR, Weng L-J. What is a good g? *Intelligence.* 1994;18(3):231–258. [https://doi.org/10.1016/0160-2896\(94\)90029-9](https://doi.org/10.1016/0160-2896(94)90029-9).
- Jiang R, Zuo N, Ford JM, Qi S, Zhi D, Zhuo C, Xu Y, Fu Z, Bustillo J, Turner JA, et al. Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *NeuroImage.* 2020;207:116370. <https://doi.org/10.1016/j.neuroimage.2019.116370>.
- Johnson W, Nijenhuis JTE, Bouchard TJ. Still just 1 g: consistent results from five test batteries. *Intelligence.* 2008;36(1):81–95. <https://doi.org/10.1016/j.intell.2007.06.001>.
- Johnson W, Bouchard TJ, Krueger RF, McGue M, Gottesman II. Just one g: consistent results from three test batteries. *Intelligence.* 2004;32(1):95–107. [https://doi.org/10.1016/S0160-2896\(03\)00062-X](https://doi.org/10.1016/S0160-2896(03)00062-X).
- Jung RE, Haier RJ. The Parieto-frontal integration theory (P-FIT) of intelligence: converging neuroimaging evidence. *Behav Brain Sci.* 2007;30(2):135–154. <https://doi.org/10.1017/S0140525X07001185>.
- King M, Hernandez-Castillo CR, Poldrack RA, Ivry RB, Diedrichsen J. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nat Neurosci.* 2019;22(8):1371–1378. <https://doi.org/10.1038/s41593-019-0436-x>.
- Kuncel NR, Hezlett SA, Ones DS. Academic performance, career potential, creativity, and job performance: can one construct predict them all? *J Pers Soc Psychol.* 2004;86(1):148–161. <https://doi.org/10.1037/0022-3514.86.1.148>.
- Lorenzo-Seva U. SOLOMON: a method for splitting a sample into equivalent subsamples in factor analysis. *Behav Res Methods.* 2021;54(6):2665–2677. <https://doi.org/10.3758/s13428-021-01750-y>.
- Major JT, Johnson W, Bouchard TJ. The dependability of the general factor of intelligence: why small, single-factor models do not adequately represent g. *Intelligence.* 2011;39(5):418–433. <https://doi.org/10.1016/j.intell.2011.07.002>.
- Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Hendrickson TJ, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature.* 2022;1–7(7902):654–660. <https://doi.org/10.1038/s41586-022-04492-9>.
- McCormick EM, Arnemann KL, Ito T, Hanson SJ, Cole MW. Latent functional connectivity underlying multiple brain states. *Netw Neurosci.* 2022;6(2):570–590. [https://doi.org/10.1162/netn\\_a\\_00234](https://doi.org/10.1162/netn_a_00234).
- McDonald RP. *Test theory: a unified treatment.* L. Mahwah, NJ: Erlbaum Associates; 1999.
- Noble S, Spann MN, Tokoglu F, Shen X, Constable RT, Scheinost D. Influences on the test-retest reliability of functional connectivity MRI and its relationship with Behavioral utility. *Cereb Cortex.* 2017;27(11):5415–5429. <https://doi.org/10.1093/cercor/bhx230>.
- Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL, et al. Functional network Organization of the Human Brain. *Neuron.* 2011;72(4):665–678. <https://doi.org/10.1016/j.neuron.2011.09.006>.
- Python Core Team. Python: a dynamic, open source programming language. Python Software Foundation. 2019. <https://www.python.org/>.
- R Core Team. (2023). R: a language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Raven JC, Court JH, Raven JE. *Manual for Raven's progressive matrices and vocabulary scales.* Rev. ed. London: H.K. Lewis; 1978.
- Ree MJ, Earles JA. The stability of g across different methods of estimation. *Intelligence.* 1991;15(3):271–278. [https://doi.org/10.1016/0160-2896\(91\)90036-D](https://doi.org/10.1016/0160-2896(91)90036-D).
- Ree MJ, Earles JA, Teachout MS. Predicting job performance: not much more than g. *J Appl Psychol.* 1994;79(4):518–524. <https://doi.org/10.1037/0021-9010.79.4.518>.
- Revelle W. (2023). *Psych: procedures for psychological, psychometric, and personality research [manual].* Evanston, Illinois: Northwestern University. <https://CRAN.R-project.org/package=psych>.
- Rosseel Y. Lavaan: an R package for structural equation Modeling. *J Stat Softw.* 2012;48(2):1–36. <https://doi.org/10.18637/jss.v048.i02>.
- Roth I. *Introduction to psychology.* Vol. 1. 1st ed. London, England: Psychology Press; 2021.
- Roth B, Becker N, Romeyke S, Schäfer S, Domnick F, Spinath FM. Intelligence and school grades: a meta-analysis. *Intelligence.* 2015;53:118–137. <https://doi.org/10.1016/j.intell.2015.09.002>.
- Roznowski M, Dickter DN, Hong S, Sawin LL, Shute VJ. Validity of measures of cognitive processes and general ability for learning and performance on highly complex computerized tutors: is the g factor of intelligence even more general? *J Appl Psychol.* 2000;85(6):940–955. <https://doi.org/10.1037/0021-9010.85.6.940>.
- Rushton JP, Ankney CD. Whole brain size and general mental ability: a review. *Int J Neurosci.* 2009;119(5):692–732. <https://doi.org/10.1080/00207450802325843>.
- Santarecchii E, Emmendorfer A, Tadayan S, Rossi S, Rossi A, Pascual-Leone A. Network connectivity correlates of variability in fluid intelligence performance. *Intelligence.* 2017;65:35–47. <https://doi.org/10.1016/j.intell.2017.10.002>.
- Schmahmann JD. The cerebellum and cognition. *Neurosci Lett.* 2019;688:62–75. <https://doi.org/10.1016/j.neulet.2018.07.005>.
- Schultz DH, Cole MW. Higher intelligence is associated with less task-related brain network reconfiguration. *J Neurosci.* 2016;36(33):8551–8561. <https://doi.org/10.1523/JNEUROSCI.0358-16.2016>.
- Shen X, Tokoglu F, Papademetris X, Constable RT. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage.* 2013;82:403–415. <https://doi.org/10.1016/j.neuroimage.2013.05.081>.
- Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, Constable RT. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat Protoc.* 2017;12(3):506–518. <https://doi.org/10.1038/nprot.2016.178>.
- Spearman C. "General intelligence," objectively determined and measured. *Am J Psychol.* 1904;15(2):201. <https://doi.org/10.2307/1412107>.

- Sripada C, Angstadt M, Rutherford S, Taxali A, Shedden K. Toward a “treadmill test” for cognition: improved prediction of general cognitive ability from the task activated brain. *Hum Brain Mapp.* 2020;41(12):3186–3197. <https://doi.org/10.1002/hbm.25007>.
- Sripada C, Rutherford S, Angstadt M, Thompson WK, Luciana M, Weigard A, Hyde LH, Heitzeg M. Prediction of neurocognition in youth from resting state fMRI. *Mol Psychiatry.* 2020;25(12):3413–3421. <https://doi.org/10.1038/s41380-019-0481-6>.
- Stammen C, Fraenz C, Grazioplene RG, Schlüter C, Merhof V, Johnson W, Güntürkün O, DeYoung CG, Genç E. Robust associations between white matter microstructure and general intelligence. *Cereb Cortex.* 2023;33(11):6723–6741. <https://doi.org/10.1093/cercor/bhac538>.
- The MathWorks Inc. (2023). *MATLAB version: 9.14.0 (R2023a)* [Computer software]. Natick, Massachusetts, United States: The MathWorks Inc. <https://www.mathworks.com>.
- Thompson WK, Barch DM, Bjork JM, Gonzalez R, Nagel BJ, Nixon SJ, Luciana M. The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: findings from the ABCD study’s baseline neurocognitive battery. *Dev Cogn Neurosci.* 2019;36:100606. <https://doi.org/10.1016/j.dcn.2018.12.004>.
- Tukey JW. *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub Co; 1977.
- Wang J, Wang X, Xia M, Liao X, Evans A, He Y. GRETNA: a graph theoretical network analysis toolbox for imaging connectomics. *Front Hum Neurosci.* 2015;9. <https://doi.org/10.3389/fnhum.2015.00386>.
- Yoon YB, Shin W-G, Lee TY, Hur J-W, Cho KIK, Sohn WS, Kim S-G, Lee K-H, Kwon JS. Brain structural networks associated with intelligence and Visuomotor ability. *Sci Rep.* 2017;7(1) Article 1:2177. <https://doi.org/10.1038/s41598-017-02304-z>.
- Zinbarg RE, Yovel I, Revelle W, McDonald RP. Estimating generalizability to a latent variable common to all of a Scale’s indicators: a comparison of estimators for  $\omega_h$ . *Appl Psychol Meas.* 2006;30(2): 121–144. <https://doi.org/10.1177/0146621605278814>.