Longitudinally stable, brain-based predictive models explain the relationships of childhood intelligence with socio-demographic, psychological and genetic factors

Narun Pornpattananangkul[a], http://orcid.org/0000-0003-1459-5255
Yue Wang[a],
Richard Anney[b], http://orcid.org/0000-0002-6083-407X
Lucy Riglin[b], http://orcid.org/0000-0002-5124-5230
Anita Thapar[b], https://orcid.org/0000-0002-3689-737X
Argyris Stringaris[c] http://orcid.org/0000-0002-6264-8377

[a]Department of Psychology, University of Otago, New Zealand
[b]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK, CF10 3AT
[c]Section on Clinical and Computational Psychiatry, National Institute of Mental Health, USA

Corresponding Author:
Narun Pornpattananangkul, PhD
Department of Psychology, University of Otago
William James Building
275 Leith Walk
Dunedin 9016, New Zealand
Email: narun.pat@otago.ac.nz

Acknowledgments:
Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, U24DA041147, U01DA041093, and U01DA041025. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/scientists/workgroups/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. We thank the developers of several R libraries, including semTools (Sunthud Pornprasertmanit), eNetXplorer (Julián Candia), and ggseg (Athanasia M. Mowinckel), for their technical advice.

Declaration of Interest: The authors declare no competing interests.

Abstract

Childhood intelligence is strikingly predictive of major life outcomes. Understanding the brain underpinnings of early-life intelligence is of prime clinical and public health importance, but has so far remained elusive. Here, we demonstrate that it is possible to arrive at models of the brain that are both predictive and explanatory of childhood intelligence. For this we leverage the unique power of the large-scale, longitudinal, multimodal MRI data from the Adolescent Brain Cognitive Development (ABCD) study (n ~11k) and create a novel predictive modeling framework that integrates machine-learning and structural-equation-based methodologies. Our predictive models combine six MRI modalities (task-fMRI from three tasks, resting-state fMRI, structural MRI, DTI) using machine-learning. In terms of prediction, our models achieve an unprecedented longitudinal association (r=.41) with childhood intelligence across two years in unseen data. We found fronto-parietal networks during a working-memory task to drive childhood-intelligence prediction. In terms of explanation, our models significantly explain variance in childhood intelligence due to (1) key socio-demographic and psychological factors (proportion mediated=18.65% [17.29%-20.12%]) and (2) genetic liability, as reflected by the polygenic score of cognitive ability (proportion mediated=15.6% [11%-20.7%]). In summary, our work shows that novel models of human multimodal neuroimaging data are powerful in helping us predict and explain variation in childhood intelligence.

Keywords:
general intelligence, multimodal neuroimaging, machine learning, polygenic score, longitudinal large-scale data

Introduction

General intelligence, known as the G-Factor[1], captures individual differences in the ability to perform across cognitive domains, such as language, mental flexibility, and memory. Children's G-Factor forecasts many meaningful life outcomes of children's future, including academic achievement[1], job attainment and performance[2], health[3], and mortality[4]. The G-Factor is a complex trait, associated with many social-demographic, psychological and genetic factors[5]. Yet, how exactly these associations are reflected in brain structure and function remains unclear. To address this question, we need a brain-based biomarker that robustly captures brain involvement in the G-Factor. Recently neuroscientists have been striving to identify such a brain-based biomarker of children's G-Factor using predictive modeling based on brain Magnetic Resonance Imaging (MRI)[6,7]. However, this endeavour so far has met with modest success due to several challenges detailed below. Our aim with this paper is to develop robust brain-based predictive models that can explain the relationships of children's G-Factor with socio-demographic, psychological and genetic factors.

Developing the brain-based predictive models of children's G-Factor faces several challenges. The first challenge is longitudinal prediction. Predictive models should not only be generalisable to out-of-sample (i.e. be predictive of children's G-Factor that were not part of the original sample) but also be longitudinally, and therefore developmentally, stable[7] in capturing the G-Factor[8]. Here we tackled this challenge by using-for the first time -to our knowledge --longitudinal, large-scale data in children, from the Adolescent Brain Cognitive Development (ABCD) study[9], to demonstrate the longitudinal stability of the prediction.

The second challenge is about missing data. Children's neuroimaging data are notoriously subject to movement artifacts[10]. For example, the ABCD study recommended a set of quality control variables for detecting noisy data from each modality[9,11], resulting in an exclusion of 17% to over 50% of data depending on a modality. If we were to exclude children who have noisy data from any single modality, we would have to exclude almost 80% of the data, strictly limiting the generalizability of our model to children with highly clean data (who are unlikely to be representative of the rest of the sample). We overcame this problem by using a recently developed, opportunistic stacking framework in machine learning[12]. This framework can deal with missingness in multimodal modeling, allowing us to keep the data as long as there is at least one modality available, leaving more data in the model building process and reducing the risk of missing-data bias.

The third challenge is multimodal integration. So far the brain-based predictive modeling has been mainly built from a single MRI modality without integrating different sources of information from different MRI modalities. It is possible, for instance, that the G-Factor depends not only on the activity of certain areas during certain cognitive tasks, such as working memory[13,14](task-based functional MRI; task-fMRI) but also on the intrinsic functional connectivity between different areas[15–17] (resting-state fMRI; rs-fMRI) as well as the anatomy of grey[18] (structural MRI; sMRI) and white[19,20] (Diffusion Tensor Imaging; DTI) matter. Here we built on a recent 'stacking' framework used in research with adults[21,22] that integrates information across MRI modalities into a stacked model. We then tested if the stacked model indeed enhanced predictive performance over single modalities in predicting children's G-Factor.

Beyond robustly predicting the G-Factor, the brain-based predictive models should be able to capture socio-demographic, psychological and genetic factors that are associated with the G-Factor. Several key socio-demographic and psychological factors are related to children's G-Factor. For instance, socio-economic status[23], mental health[24,25] and extracurricular activities[26] have previously been associated with the G-Factor. Likewise, genetics also play an important role in explaining children's general intelligence[27]. Genome-Wide Association Studies (GWAS) of cognitive ability have revealed the polygenic nature of general intelligence[28]. A recent GWAS, for instance, was able to identify common gene variants: single nucleotide polymorphisms (SNPs) that were significantly associated with cognitive ability among 257,841 adults[29]. Polygenic scores (PGS), a composite measure of common gene variants, provides an index of genetic liability thereby allowing researchers to quantify genetic liability for cognitive ability[29,30].

To develop brain-based predictive models for the G-Factor from children's MRI data of different modalities, we used the ABCD Release 3.0[9] that included full baseline data (age 9-10 years old) and half follow-up data (age 11-12 years old) from over 11,000 children. We first applied a higher-order confirmatory factor analysis (CFA) to extract the G-Factor from six cognitive tasks, collected outside of MRI. We then used the baseline data from children whose follow-up data have not been released yet as the training set, which was further split into first- and second-layer training sets. For the first-layer training set, we used Elastic Net[31] to compute modality-specific models that separately predicted the G-Factor, including three task-based fMRI (working-memory "N-Back", reward "Monetary Incentive Delay; MID" and inhibitory control "Stop Signal"), rs-fMRI, sMRI, and DTI. At the second-layer training set, we applied Random Forests[32] to integrate the predicted values from all modality-specific models and handled missing values with Opportunistic Stacking[12].

Using a longitudinal, predictive modeling approach (see Figure 1), we tested the predictive performance of modality-specific (using the first layer only) and stacked modeling (using both layers) on the unseen baseline and follow-up data. To determine the robustness of the predictive models, we specifically examined whether the predictive models were generalisable, longitudinally stable and robust against missing values. Next to demonstrate whether multimodal integration led to better predictive performance, we applied bootstrapping to compare the stacked model with the best-performing modality-specific model. To interpret the feature importance of the final models (i.e., determining brain features that contributed highly to the prediction of the G-Factor), we applied permutation tests (eNetXplorer[33] and conditional permutation importance[34]).

Using structural equation modeling (SEM) along with leave-one-site-out cross-validation, we then tested the extent to which the brain-based predictive models captured the relationships of children's G-Factor with socio-demographic, psychological and genetic factors. We first extracted predicted values of the G-Factor based on the stacked model from multimodal MRI data at each held-out data collection site. To quantify the relation between the G-Factor and key socio-demographic and psychological factors, we applied Elastic Net[31] to predict the G-Factor from 70 variables[26], covering children's and/or their parents' socio-demographics, mental health, personality, sleep, physical activity, screen use, drug use, developmental adversity and social interaction. Similar to the brain-based predictive models, we then extracted predicted values of the G-Factor based on these socio-demographic and psychological factors at each held-out data collection site. To quantify the genetic liability for

the G-Factor, we computed a PGS of cognitive ability based on the large-scale GWAS in adults[29]. We then used mediation analyses via SEM to demonstrate the extent to which the brain-based predictive models accounted for the relationships of children's G-Factor with these socio-demographic, psychological and genetic factors.

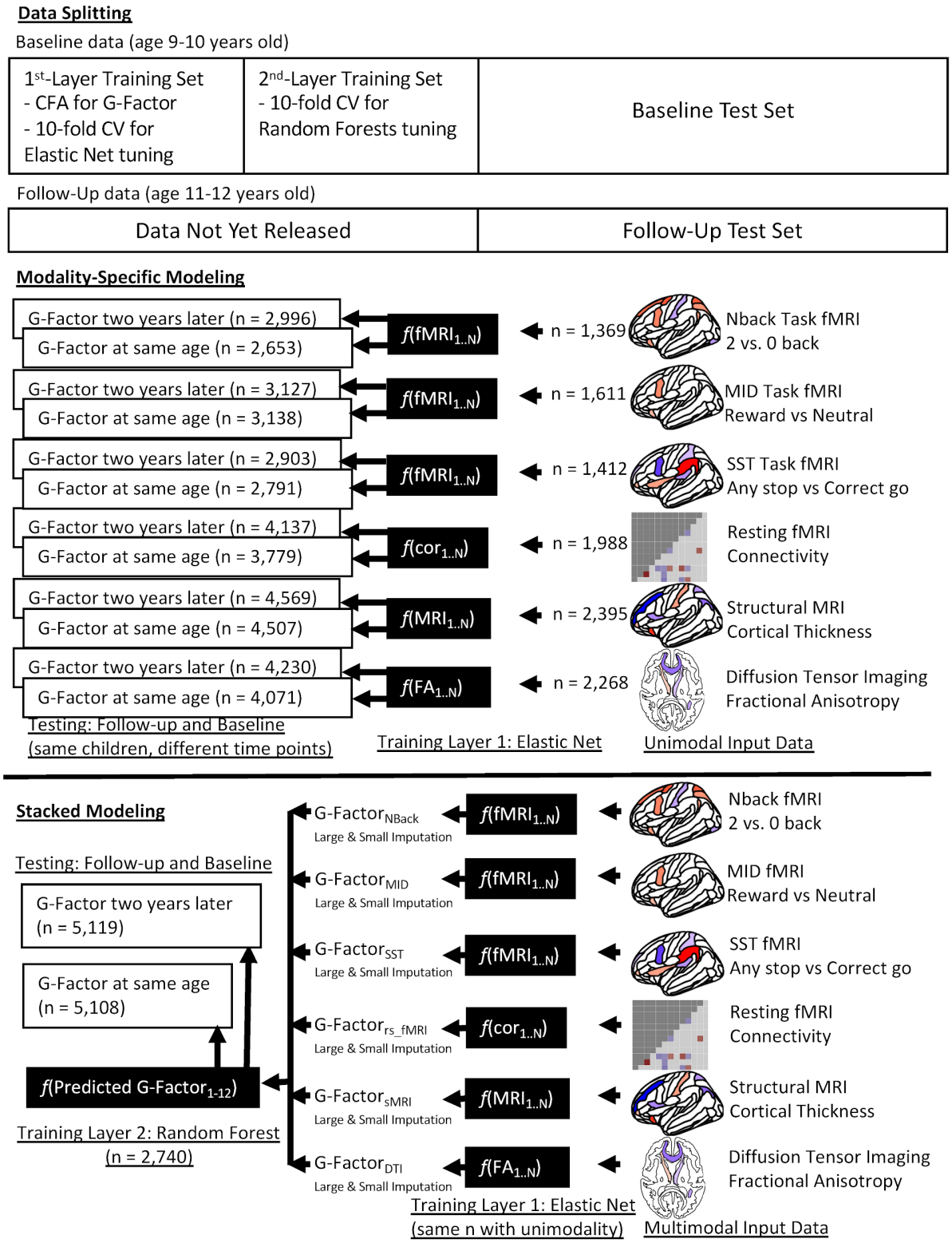


Figure 1. Longitudinal Predictive Modeling Approach. We split the data into four sets: first-layer training, second-layer training, baseline test, and follow-up test. We used the same

participants in the baseline test and follow-up test sets. Modality-specific modeling only used the first-layer training set, while stacked modeling used both training sets to combine predicted values across modalities. The number of observations was different depending on the quality control of data from each modality. Stacked modeling imputed missing value to a large number (1000) and a small number (-1000) in two separate features. "Data not yet released" reflects the fact that ABCD Release 3.0[9] only provided half of the follow-up data (age 11-12 years old), while providing the full baseline data (age 9-10 years old). CFA = Confirmatory Factor Analysis; CV = Cross-Validation; cor = correlation; FA = fractional anisotropy.

Results

See Supplementary for the confirmatory factor analysis (CFA) of the G-Factor. Briefly the CFA indicated a good fit of our G-Factor structure.

How robust are the brain-based predictive models?

We first applied the longitudinal predictive models for multimodal MRI (Figure 1) to examine the robustness of the brain-based predictive models. Figure 1 also shows the number of participants left in each modality per data split after quality control. To test the stacked model's performance, we split the test sets based on the presence of each modality. First, Stacked All required data with at least one modality present. This allowed us to examine the stacked model's performance when the missing values were all arbitrarily coded via Opportunistic Stacking[12] (see Method). Second, Stacked Complete required data with all modalities present. This represents the situation when the data were as clean as possible. Third, Stacked Best had the same missing values with the modality with the best prediction. This allowed us to make a fair comparison in performance between the stacked model and the model with the best modality, given their same noise level from missing value. We applied bootstrapping with 5,000 iterations to examine the difference in performance indices. Fourth, Stacked No Best did not have any data from the modality with the best prediction and had at least one modality present. For hyperparameters for the longitudinal predictive models for multimodal MRI, see Supplementary.

The stacked model for multimodal MRI generalised well to the unseen baseline and follow-up data.

Table 1 and Figure 2 summarise the predictive ability of the longitudinal predictive models. Performance of Stacked All, Stacked Complete and Stacked Best was among the top with Pearson's r over .4 and R^2 around .19. Importantly, the superior performance of stacked models was found across baseline and follow-up test sets at a similar magnitude, suggesting their longitudinal stability. Note that given that the N-Back task-based fMRI had the highest performance among modality-specific models, we set the missing values of the Stacked Best to be the same as those of the N-Back task-based fMRI.

The stacked model for multimodal MRI was robust against missing values.

Figure 2 shows the proportion of missing data in the two test sets. sMRI had the lowest missing observations, while the three task-based fMRI data had the highest. Missing observations in the stacked model with at least one modality present, Stacked All, were around 3-6%, while those in the stacked model with all modalities present, Stacked Complete, were up to 78.79%. Because performance of Stacked All was similar to that of Stacked Complete, the Opportunistic Stacking[12] algorithm that led to Stacked All was robust against missing values.

Table 1. Predictive Performance of the longitudinal predictive model and leave one-site out cross-validation for multimodal MRI. R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

Predictive Performance from the baseline test set

Models	Pearson's r	R squared	MAE	RMSE
Stacked_All	0.439	0.191	0.699	0.895
Stacked_Complete	0.429	0.183	0.61	0.78
Stacked_Best	0.442	0.195	0.62	0.798
Stacked_NoBest	0.296	0.085	0.783	0.987
NBack	0.402	0.072	0.664	0.857
SST	0.129	-0.033	0.744	0.95
MID	0.202	0.013	0.738	0.944
rs_fMRI	0.233	0.042	0.749	0.955
sMRI	0.208	0.04	0.763	0.969
DTI	0.19	0.033	0.757	0.972

Predictive Performance from the follow-up test set

Models	Pearson's r	R squared	MAE	RMSE
Stacked_All	0.414	0.166	0.719	0.913
Stacked_Complete	0.427	0.168	0.651	0.829
Stacked_Best	0.438	0.175	0.666	0.846
Stacked_NoBest	0.317	0.1	0.794	1
NBack	0.383	0.118	0.687	0.875
SST	0.145	-0.004	0.76	0.961
MID	0.148	-0.003	0.757	0.955
rs_fMRI	0.251	0.055	0.754	0.954
sMRI	0.226	0.049	0.764	0.965
DTI	0.212	0.045	0.771	0.978

Mean (SD) of Out-Of-Site Predictive Performance

Models	Pearson's r	R squared	MAE	RMSE
Stacked	0.46 (0.057)	0.21 (0.052)	0.698 (0.023)	0.888 (0.029)
NBack	0.408 (0.069)	0.167 (0.055)	0.718 (0.028)	0.91 (0.031)
MID	0.227 (0.096)	0.05 (0.05)	0.772 (0.021)	0.973 (0.025)
SST	0.139 (0.071)	0.019 (0.024)	0.783 (0.014)	0.988 (0.012)
rs_fMRI	0.255 (0.061)	0.064 (0.03)	0.765 (0.015)	0.966 (0.016)
sMRI	0.248 (0.092)	0.061 (0.046)	0.763 (0.024)	0.967 (0.024)
DTI	0.223 (0.076)	0.049 (0.037)	0.766 (0.016)	0.974 (0.019)

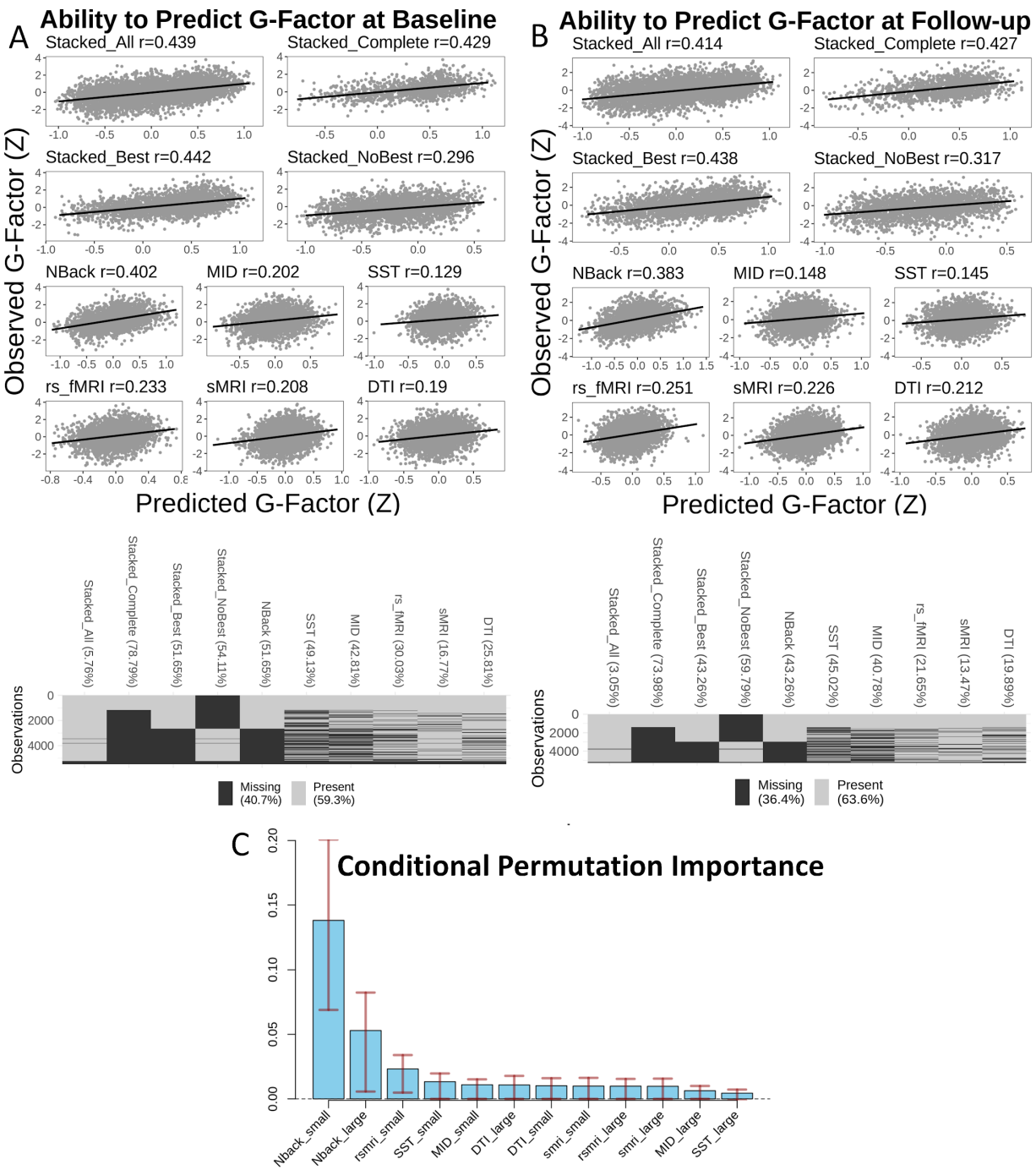


Figure 2. Predictive Ability of the Longitudinal Predictive Models and Missing Observations as a Function of Modalities in the Test Sets (2A for the baseline test set and 2B for the follow-up test set) and the Conditional Permutation Importance (CPI) of the Stacked Model (2C). Stacked All required data with at least one modality present. Stacked Complete required data with all modalities present. Stacked Best had the same missing values with the modality with the best prediction (N-Back task-based fMRI). Stacked No Best did not have any data from the modality with the best prediction and had at least one modality present. The CPI was computed based on the second-layer training set. Error bars in the bar plot show an interval between .25 and .75 quantiles of the CPI for each tree in the Random Forests. The "_large" and "_small" suffixes indicate whether the missing values were coded as a large (1000) or small (-1000) number, respectively.

The stacked model for multimodal MRI performed better than the best single modality.

N-Back task-based fMRI provided the best out-of-sample predictive ability for both baseline and follow-up test sets, relative to other modality-specific models. Figure 3 compared the predictive ability between the Stacked Best and N-Back task-based fMRI using bootstrapped differences. The Stacked Best had significantly higher performance in both baseline and follow-up test sets, reflected by higher Pearson's *r* and R^2 and lower MAE and RMSE. This indicates the boost in predictive performance when multiple modalities were integrated, at around 12% for the baseline data and 6% for the follow-up data.

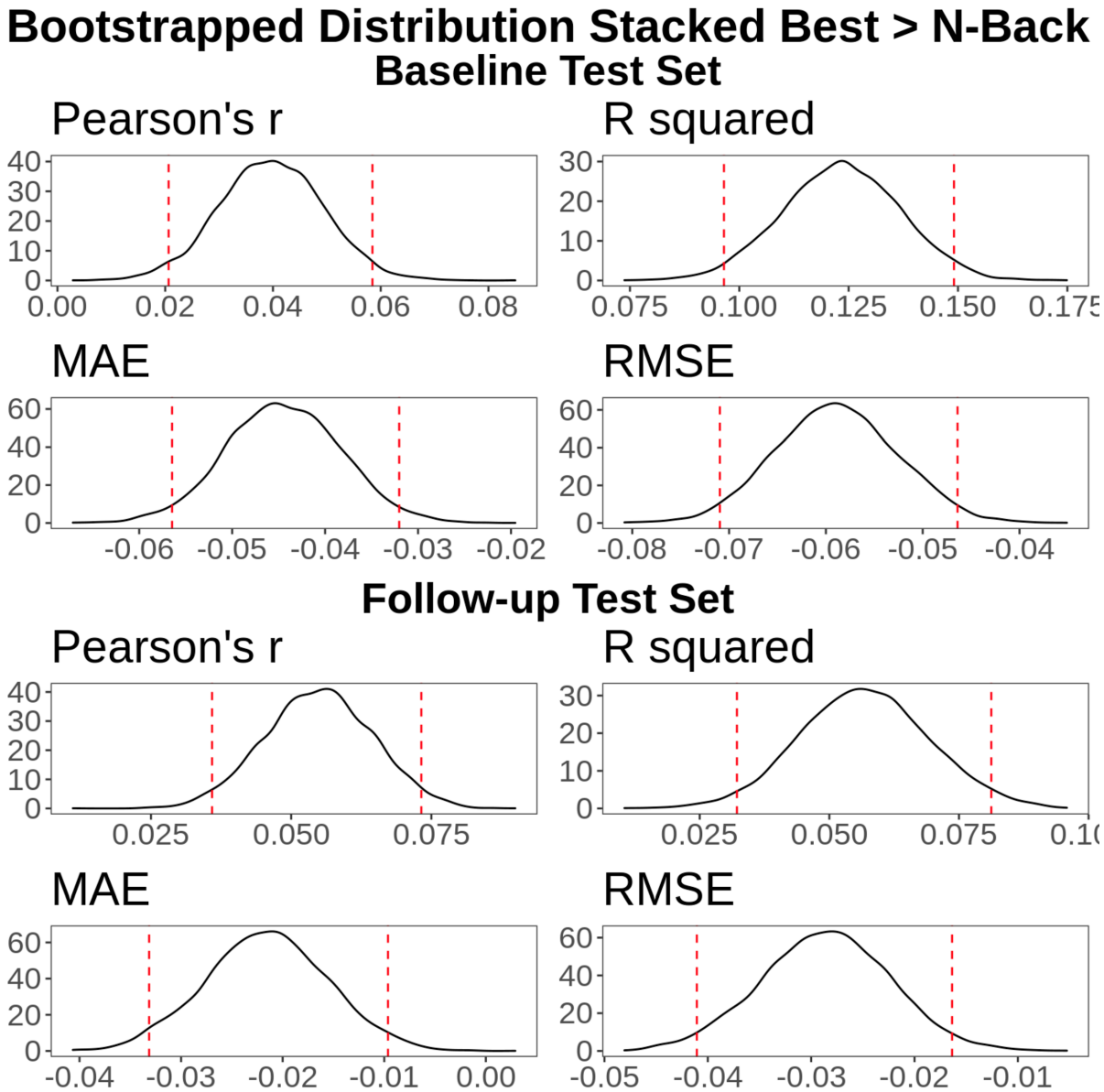


Figure 3. The Bootstrapped Distribution of the Differences in Predictive Ability between the Stacked Best and N-Back Task-Based fMRI. Dotted lines indicate 95% confidence intervals. We subtracted the predictive ability of N-Back Task-Based fMRI from that of the Stacked Best. R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

Feature Importance for multimodal MRI

To interpret the feature importance of the final models, we first applied the Conditional Permutation Importance[34] to examine which of the modalities contributed strongly on the prediction of the stacked model (Figure 2C). N-Back task-based fMRI had the highest importance score. We then applied eNetXplorer[33] to show brain features that significantly (empirical $p < .05$) contributed to the prediction of each modality-specific model (Figure 4). For N-Back task-based fMRI, the G-Factor prediction was driven by activity in areas, such as the precuneus, sulcus intermedius primus, superior frontal sulci, and dorsal cingulate. For MID task-based fMRI, the prediction was driven by activity in several areas in the parietal, frontal and temporal regions. For SST, the prediction was contributed by activity in areas such as the supramarginal gyrus and inferior precentral sulcus. For rs-fMRI, the prediction was driven by connectivity within cinguloparietal and sensory-motor-hand as well as between networks that were connected with frontoparietal, default-mode, and sensory-motor-hand networks. For sMRI, the prediction was driven by the volume/thickness at several areas, such as the insula, middle frontal gyrus, and lingual sulcus. For DTI, the prediction was driven by FA at several white matter tracts, such as the superior longitudinal fasciculus, forceps minor, uncinate and parahippocampal cingulum.

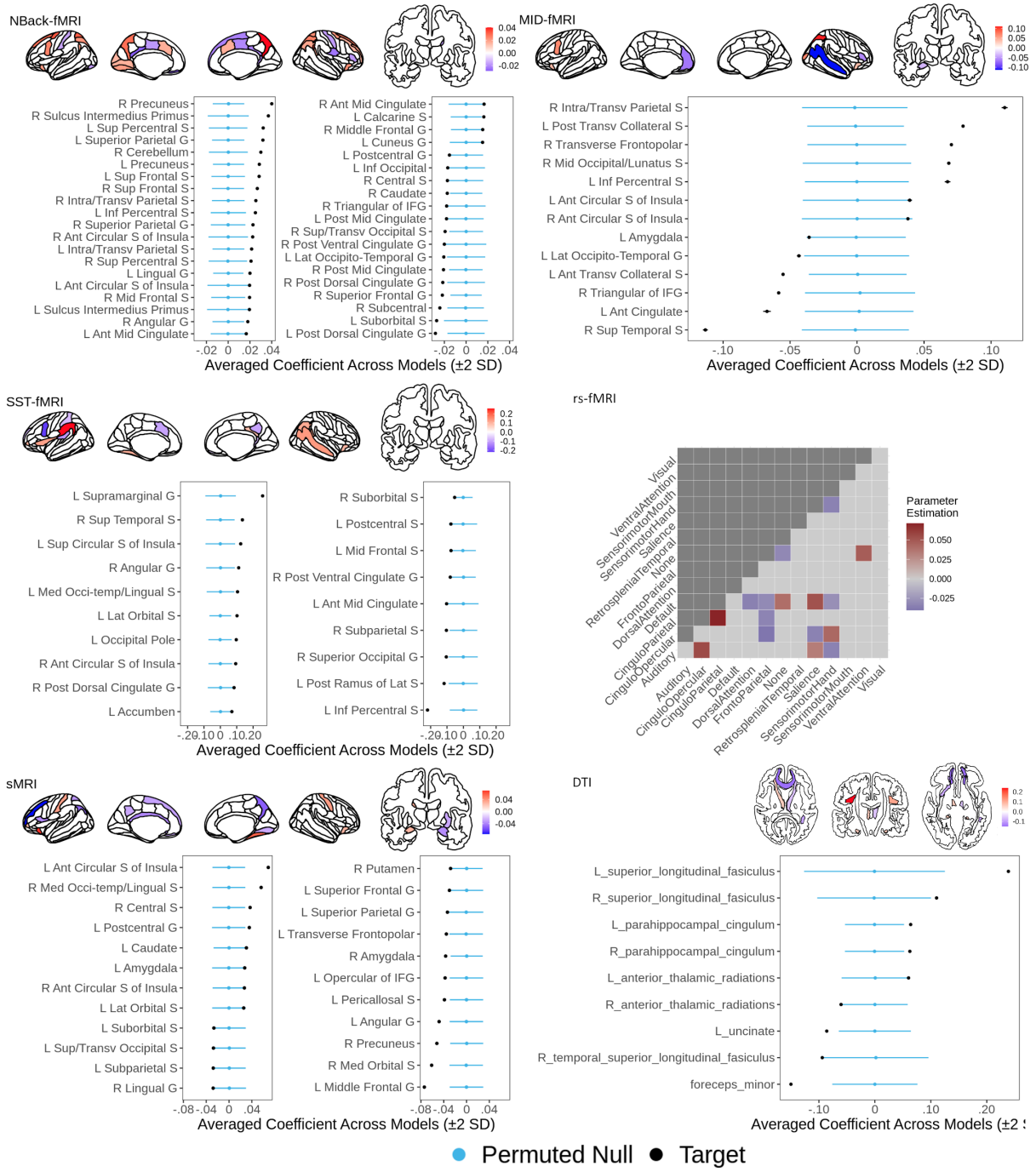


Figure 4. Feature importance for modality-specific models. We only plotted brain features with empirical p < .05 based on eNetXplorer[33] permutation. Sup=Superior; Ant=Anterior; Lat=lateral; Med=Medial; S=Sulcus; G=Gyrus; IFG=Inferior Frontal Gyrus; L=left; R=right.

<u>Did the brain-based predictive models capture the relationships of the G-Factor with socio-demographic, psychological and genetic factors?</u>

<u>Multimodal MRI</u>

We applied leave-one-site-out cross-validation to multimodal MRI from the baseline data (see Table 1 and Figure 5). This enabled us to extract predicted values of the G-Factor based on multimodal MRI data at each held-out site. Across 21 sites, the out-of-site predictive ability of the stacked model was highest, explaining on-average 21% (SD =5.2) of

the variance in the G-Factor. This confirmed the generalisability of the stacked model and ensured its use for subsequent mediation analyses.

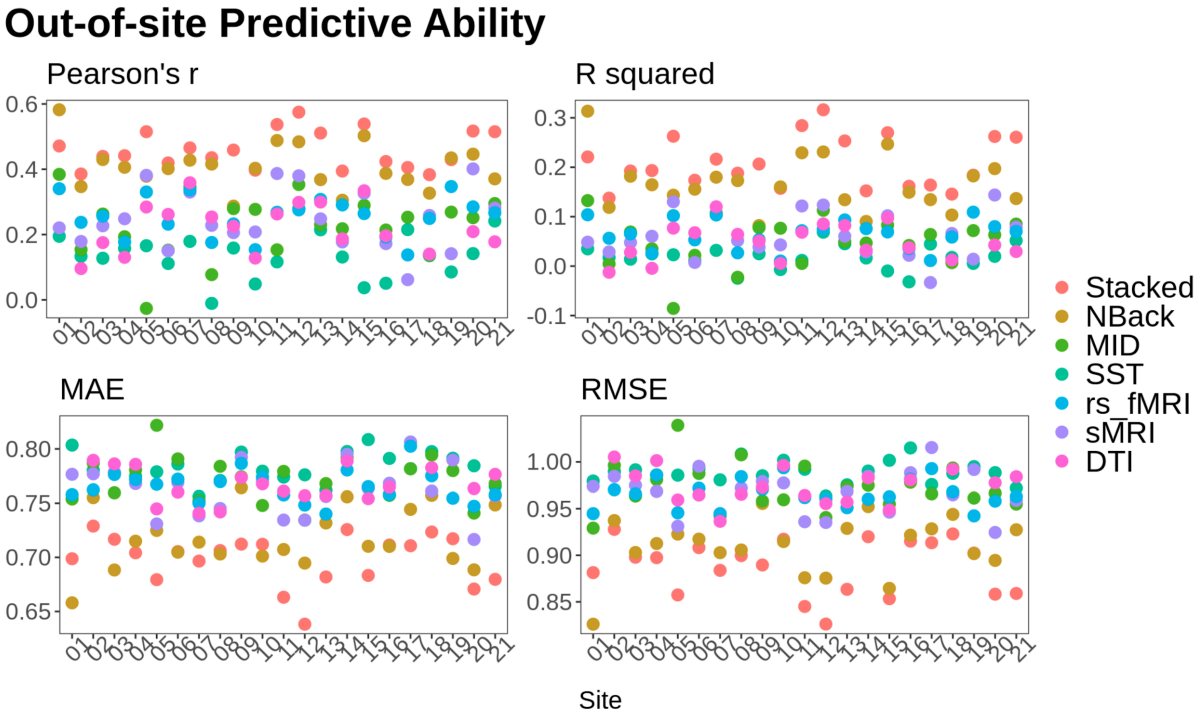


Figure 5. Leave-one-site-out cross-validation for multimodal MRI. We evaluated out-of-site predictive ability between predicted vs. observed G-Factor in the held-out site. Note that DTI data were not available from 3 sites (sites 1, 17, and 19). R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

Key socio-demographic and psychological factors

We applied leave-one-site-out cross-validation on socio-demographic and psychological factors (see Figure 6). Across sites, socio-demographic and psychological factors explained on-average 29.7% (SD =8.1) of the variance in the G-Factor. The top features in the Elastic-Net models that had the magnitude of their standardised coefficients over .1 included parents' education and income along with child's attention and social problems as well as extracurricular activities.

Polygenic scores (PGS)

Figure 8a,b show the relationship between the G-Factor and the PGS of cognitive ability at different thresholds. While the G-Factor was significantly related to the PGS of cognitive ability across thresholds, the relationship at the $p<.01$ PGS threshold was the numerically strongest ($r=.21$, $p<.001$ [CI95%=.18-.24]). Accordingly, we used PGS at the $p<.01$ PGS threshold for the mediation analyses.

Mediation analyses

We tested whether the brain-based, stacked models captured the relationships of the G-Factor with socio-demographic, psychological and genetic factors using mediations. We found significant indirect effects (1) when the socio-demographic and psychological factors were the sole independent variable (Figure 6d proportion mediated = 19.1%), (2) when the genetic factor was the sole independent variable (Figure 7c, proportion mediated = 15.6%)

and (3) when both the socio-demographic and psychological factors (Figure 8, proportion mediated = 15%) and the genetic factor (Figure 8, proportion mediated = 10.75%) were the covariated independent variables.

a Mean (SD) of Out-Of-Site Predictive Performance of the Elastic-Net Model Predicting G-Factor from Key Socio-Demographic And Psychological Factors

Pearson's r	R squared	MAE	RMSE
.546(.072)	.297(.081)	.646(.039)	.820(.048)

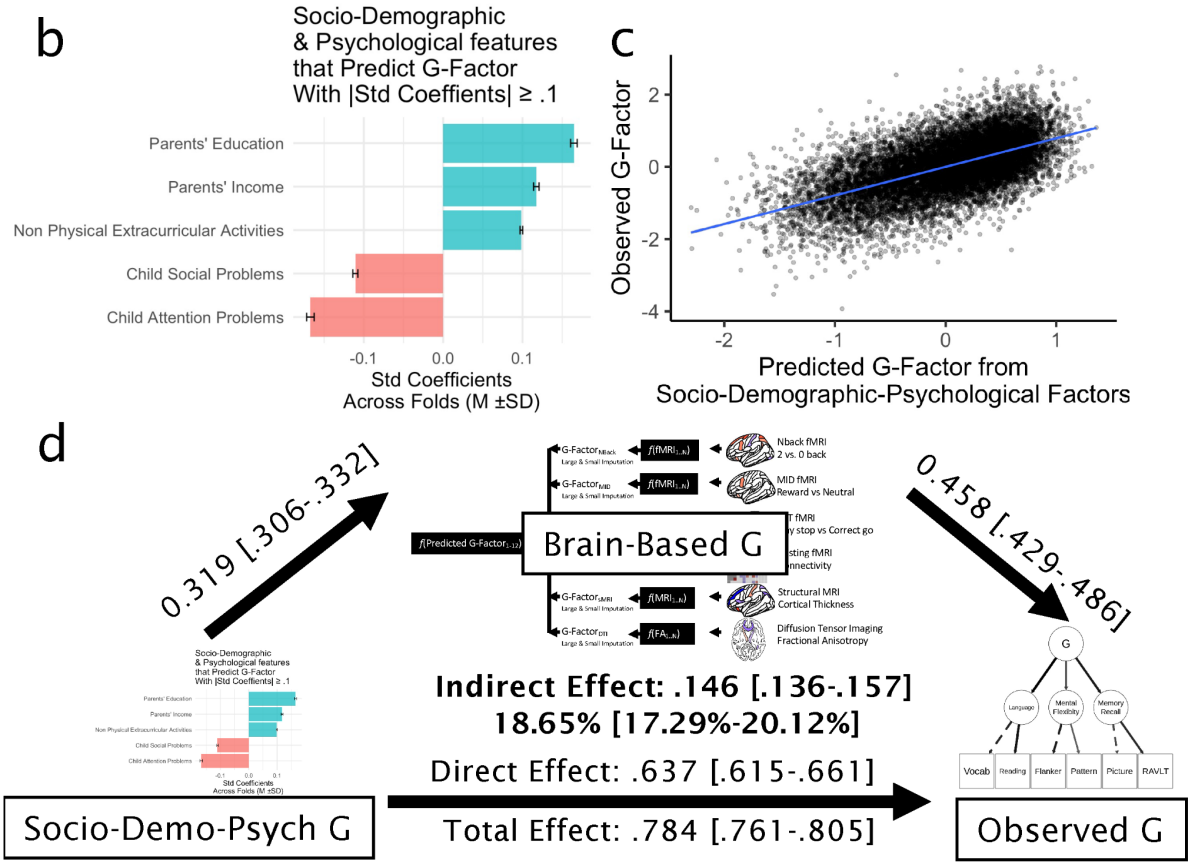


Figure 6. Leave-one-site-out cross-validation for key socio-demographic and psychological factors. 6a shows out-of-site predictive performance of the Elastic-Net model predicting the G-Factor from key socio-demographic and psychological factors. 6b shows the top socio-demographic and psychological features with the magnitude of standardised coefficient over .1 based on the Elastic-Net model. Blue indicates a positive relationship while red indicates a negative relationship. 6c shows a scatter plot between out-of-site predicted values of the G-Factor based on key socio-demographic and psychological factors and the observed (i.e., real) values of the G-Factor. 6d shows a mediation analysis where (1) the predicted values of the G-Factor based on the key socio-demographic and psychological factors at all held-out sites is the independent variable, (2) the predicted values of the G-Factor based on the brain-based, stacked predictive models at all held-out sites is the mediator and (3) the observed G-Factor is the dependent variable. % under the indirect effect indicates proportion mediated. [ ] indicates a 95% confidence interval based on bootstrapping. R squared = coefficient of determination; MAE = mean absolute error; RMSE = root mean squared error.

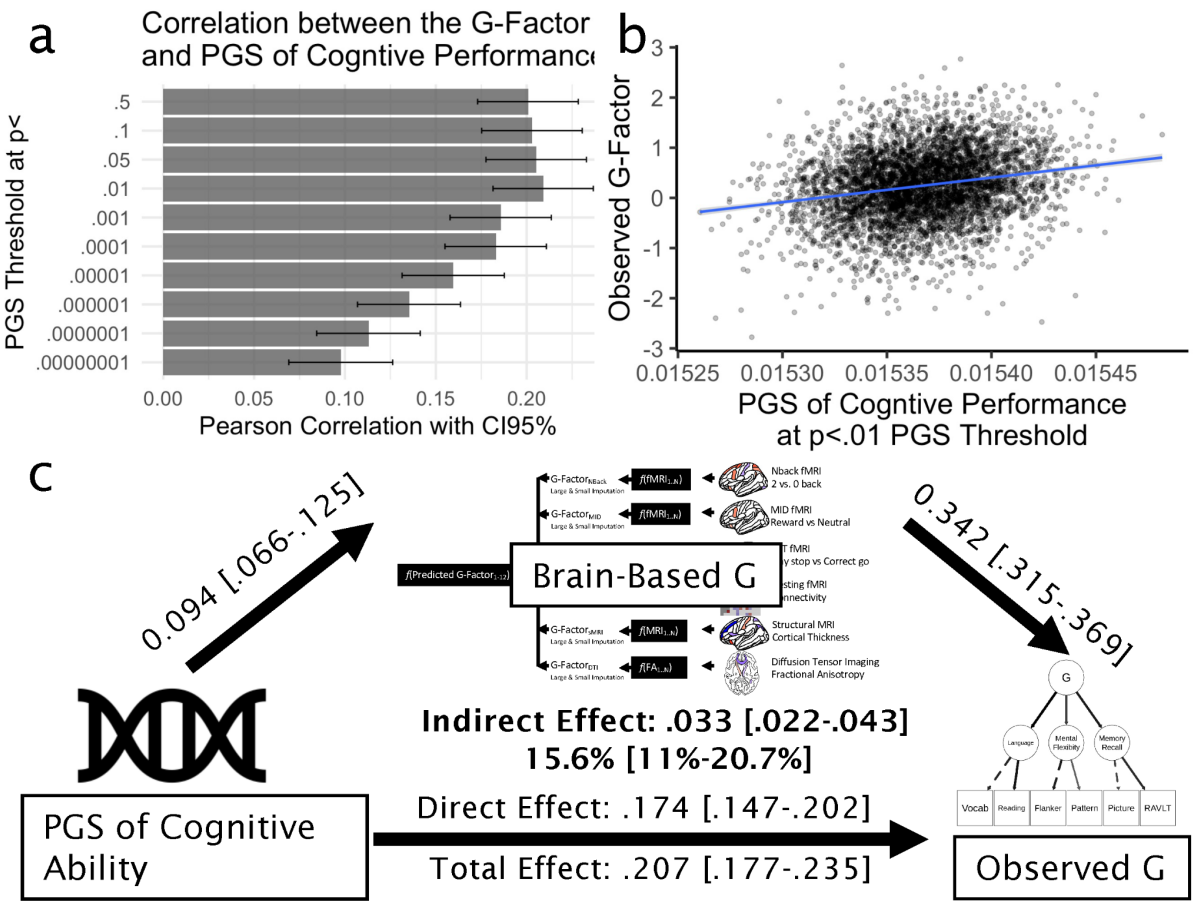


Figure 7. Polygenic Scores (PGS) of Cognitive Ability. 7a shows Pearson's correlations between the G-Factor and PGS of cognitive ability at different PGS thresholds. 7b shows a scatter plot between the PGS of cognitive ability at the p<.01 PGS threshold and the observed (i.e., real) values of the G-Factor. 7c shows a mediation analysis where (1) the PGS of cognitive ability at the p<.01 PGS threshold is the independent variable, (2) the predicted values of the G-Factor based on the brain-based, stacked predictive models at all held-out sites is the mediator and (3) the observed G-Factor is the dependent variable. Not shown in the figure are four principal components included as the control variables to adjust for population stratification. % under the indirect effect indicates proportion mediated. [ ] indicates a 95% confidence interval based on bootstrapping.

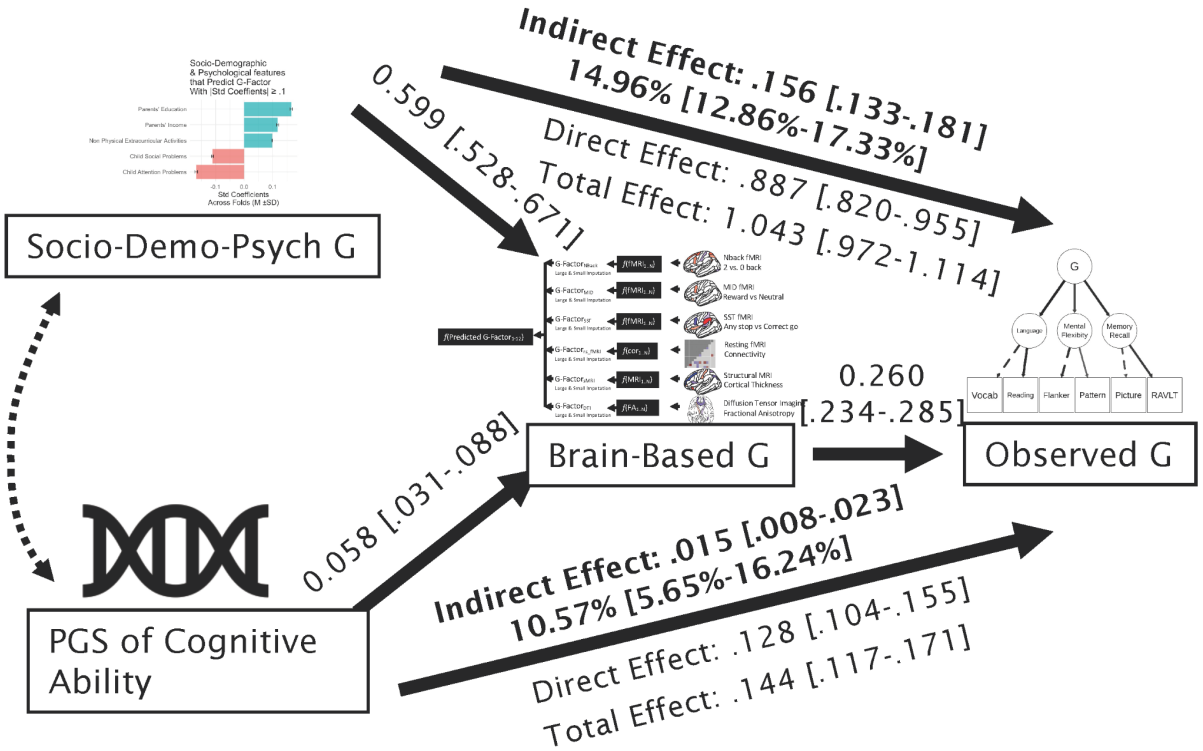


Figure 8. Mediation analysis with both key socio-demographic and psychological factors as well as genetic factors as independent variables. Specifically, this model treated (1) the predicted values of the G-Factor based on the key socio-demographic and psychological factors at all held-out sites and (2) the PGS of cognitive ability at the p<.01 PGS threshold as two separate independent variables. It treated the predicted values of the G-Factor based on the brain-based, stacked predictive models at all held-out sites as the mediator and the observed G-Factor as the dependent variable. Not shown in the figure are four principal components included as the control variables to adjust for population stratification. % under the indirect effect indicates proportion mediated. [ ] indicates a 95% confidence interval based on bootstrapping. The dotted, double arrowed line indicates covariation between the two independent variables.

Discussion

We aim to develop robust brain-based predictive models for childhood G-Factor that can capture the relationships of the G-Factor with socio-demographic, psychological and genetic factors. Here we showed that incorporating data from different MRI modalities into stacked models allowed for robust brain-based predictive models. They were stable across years and generalisable to different sites while being able to handle missing values. Moreover, we demonstrated that the brain-based, stacked models significantly explained the relationships of the G-Factor with socio-demographic, psychological and genetic factors.

<u>The brain-based, stacked predictive models for the G-Factor were robust.</u>
We developed longitudinal predictive models for children's G-Factor from MRI data of different modalities. We built models from the baseline MRI data and tested them on unseen children at the same age and two years older. We found similar predictive ability across these two test sets for all modality-specific and stacked models. That is, the models that had high out-of-sample prediction on same-age children also had high out-of-sample prediction on older children, suggesting the longitudinal stability of MRI for many modalities. The best model across all performance indices (Pearson's r, R^2, MAE, and RMSE) was the stacked model that incorporated all six modalities, which was followed closely by the N-back task-related fMRI model. Apart from the SST task-related fMRI model, other models (including the MID tasked-related, rs-fMRI, sMRI, and DTI) performed moderately well. We also found a similar magnitude of predictive ability based on leave-one-site-out cross-validation, suggesting the generalisability of MRI not only across ages but also across data collection sites. Overall, the stacked model predicted around 20% of the variance in the children's G-Factor. This made the stacked model the most generalisable to out-of-sample children and the most longitudinally stable.

Beyond generalisability across ages and sites, the stacked model also allowed us to handle missingness in the data[12]. This is especially important for children's MRI data given high levels of noise in certain modalities[10]. If we were to use data only from children with all modalities present (i.e., the Stacked Complete), the model would not apply to around 80% of the children. The Opportunistic Stacking allowed us to use the data as long as one modality was present (i.e., the Stacked All), leaving the exclusion to just around 5%. Importantly, the predictive performance of Stacked Complete and Stacked All were both relatively high, ensuring the ability of Opportunistic Stacking to deal with the missing data. Moreover, in the case when the best modality was not available, using the stacked model (i.e., the Stacked No Best) could be helpful. While the predictive ability of the Stacked No Best was not as strong as the Stacked Complete, Stacked All, and Stacked Best, its performance measures of variance (Pearson's r and R squared) appeared stronger in magnitude than any other non-optimal modalities by itself. Accordingly, in settings when not all of the modalities are available, researchers/practitioners can still take advantage of the boosted predictive ability of the stacked models over unimodal models.

The stacked model improved predictive ability over and above the best modality, which was the N-Back task-based fMRI. This is based on bootstrapping of the difference in performance indices between the N-Back task-based fMRI and stacked model with the same participants (i.e., the Stacked Best). Accordingly, combining across modalities, as long as they are available, in one model seems to capture additional variance in the test data that

cannot be achieved even by the best modality in the model. This is consistent with previous studies showing the enhanced predictive power of the stacked mode[12,21].

Our use of permutation allowed easy-to-interpret models, highlighting the neurobiological bases of children's G-Factor. Conditional Permutation Importance[34] allowed us to infer that prediction from the stacked model was driven primarily by N-back task-related fMRI. This indicates the important role of working memory. eNetXplorer permutation[33] further showed us that contribution from fMRI activity in the parietal and frontal areas during the N-back task drove the prediction. These areas were similar to the areas previously found in a recent study in adults[35]. Similarly, we also found brain indices from other modalities, from activity during other tasks to the cortical thickness and white matter density, that contributed to the prediction of the G-Factor, albeit with lower predictive performance.

Unlike previous unimodal[13–20] and multimodal studies[21,22], we were able to compare the ability of task-based fMRI with other modalities in predicting the G-Factor. We found that one of the three task-based fMRI models, the N-Back, performed exceptionally well. Based on the Conditional Permutation Importance[34], the N-Back task-related fMRI appeared to drive the prediction of the stacked model. This finding is consistent with recent study showing superior performance of the N-Back task in predicting the G-Factor, compared to resting-state fMRI[35]. Showing that task-based fMRI from a certain task could capture cognitive ability across a two-year gap provided a promising outlook for the use of task-based fMRI as a predictive tool.

It is important to note that not all fMRI tasks were suitable for predicting certain targets. The N-Back task and SST, for instance, were designed to capture working memory[36,37] and inhibitory control[36,38], respectively. Accordingly, both should be related to the G-Factor, especially on memory recall and mental flexibility portions of the G-Factor. Yet, only the N-Back task showed good predictive ability. This may be due to different cognitive processes in each task (i.e., working memory vs. inhibitory control) or to different task configurations. It is entirely possible, for instance, that the block design used in the N-Back, as opposed to the event-related design used in the SST, allowed the N-Back to have higher predictive power. Accordingly, while task-based fMRI can have high predictive power, systematic comparisons are required in future research to better understand the characteristics of some tasks that make them more suitable for predicting the G-Factor and other individual differences.

<u>The brain-based, stacked predictive models for the G-Factor explained the influences of socio-demographic, psychological and genetic factors.</u>

Here we first built a predictive model of the G-Factor using 70 socio-demographic and psychological features[26]. We found relatively high performance of this model, accounting for around 30% of the G-Factor. Moreover, the top contributing features are consistent with previous studies, including socio-demographics[23] (e.g., parents' education and income) along with children's mental health[24,25] (e.g., attention and social problems) and children's extracurricular activities[26]. More importantly, our mediation analysis showed that the brain-based, stacked predictive models captured approximately 19% of the relationships of children's G-Factor with these socio-demographic and psychological features.

As for the genetic factor, we first showed that the polygenic score (PGS) based on adults' cognitive ability[29] was related to children's G-Factor, consistent with a recent study[30]. Similar

to the socio-demographic and psychological factors, our mediation analysis showed that the brain-based, stacked predictive models captured approximately 16% of the relationship of children's G-Factor with the PGS of cognitive ability. In fact, when having socio-demographic and psychological factors and the PGS of cognitive ability together as independent variables, the brain-based, stacked predictive models significantly captured the relationship of the G-Factor with both variables. Altogether, this demonstrates the utility of the brain-based, stacked predictive models in explaining the brain involvement in the relationships of the G-Factor with socio-demographic, psychological and genetic factors, in line with the multifacet framework of the G-Factor[5].

Our study is not without limitations. We relied on the ABCD study's curated, preprocessed data[9,11,36]. This provided certain advantages. For instance, given its standardization, other studies that wish to apply our model of G-Factor to the ABCD data can readily do so without concerns about differences in preprocessing steps. Preprocessed data also enabled us to apply the manual quality control done by the study, a process that required time and well-trained labour[9,11,36]. Preprocessing large-scale multi-modal data ourselves would not only demand significant computer power and time but is prone to error. However, using the preprocessed data only allowed us to follow the choices of processing done by the study. For example, ABCD Release 3 only provided Freesurfer's parcellation[39,40] for task-based fMRI. While this popular method allowed us to interpret task-based activity on subject-specific anatomical landmarks, the regions are relatively large compared to other parcellations. Future studies will need to examine if smaller and/or different parcellations would improve predictive performance.

In conclusion, we developed brain-based stacked, predictive models for children's G-Factor that were longitudinally stable, generalisable and robust against missingness. More importantly, our brain-based models were able to explain the relationships of childhood G-Factor with the socio-demographic, psychological and genetic factors. Accordingly, our approach should pave the way for future researchers to employ multimodal MRI as robust predictive tools for children's G-Factor.

Online Methods

We employed the ABCD Study Curated Annual Release 3.0[9], which included 3T MRI data and cognitive tests from 11,758 children (female=5,631) at the baseline (9-10 years old) and 5693 children (female = 2,617) at the two-year follow-up (11-12 years old). The study recruited the children from 21 sites across the United States[41] according to the ethical oversight as detailed elsewhere[42]. We further excluded 54 children based on Snellen Vision Screener[43,44]. These children either could not read any line, could only read the 1st (biggest) line, or could read up to the 4th line but indicated difficulty in reading stimuli on the iPad used for administering cognitive tasks (see below).

Testing the robustness of the brain-based predictive models:

Data Splitting

We split the data into 4 parts (Figure 1): 1) first-layer training set (n=3,041), 2) second-layer training set (n=3,042), 3) baseline test set (n=5,622) and 4) follow-up test set (n=5,656). Especially noteworthy is that children who were in the baseline test set were also in the follow-up test set. In other words, none of the children in the first-layer and second-layer training sets were in either of the test sets.

Target: the G-Factor

We modeled the G-Factor using children's performance from six cognitive tasks. These six tasks, collected on an iPad during a 70-min in-session visit outside of MRI[44,45], were available in both baseline and follow-up datasets. First, the Picture Vocabulary measured vocabulary comprehension and language[46]. Second, the Oral Reading Recognition measured reading and language decoding[47]. Third, the Franker measured conflict monitoring and inhibitory control[48]. Fourth, the Pattern Comparison Processing measured the speed of processing[49]. Fifth, the Picture Sequence Memory measured episodic memory[50]. Sixth, the Rey-Auditory Verbal Learning measured auditory recognition, recall, and learning[51].

Similar to the previous work[45,52,53], we applied a higher-order G-Factor model using confirmatory factor analysis (CFA) to encapsulate the G-Factor as the latent variable underlying performance across cognitive tasks. More specifically, in our higher-order G-Factor model (Figure 2), we had the G-Factor as the 2nd-order latent variable. We also had three 1st-order latent variables in the model: language (capturing the Picture Vocabulary and Oral Reading Recognition), mental flexibility (capturing the Franker and Pattern Comparison Processing), and memory recall (capturing the Picture Sequence Memory and Rey-Auditory Verbal Learning).

To prevent data leakage, we fit the CFA model to the observations in the first-layer training set and then computed factor scores of the G-Factor on all training and test sets. We fixed latent factor variances to one and applied robust maximum likelihood estimation (MLR) with robust (Huber-White) standard errors and scaled test statistics. To demonstrate model fit, we used scaled and robust comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean squared error of approximation (RMSEA) with 90% CI and internal consistency, OmegaL2[54], of the G-Factor (see Supplementary). To impletment the CFA, we used lavaan[55] (version=.6-6) and semTools[54] along with semPlot[56] for visualization.

Features: multimodal MRI

We used MRI data from six modalities: three task-based fMRI, rs-fMRI, sMRI, and DTI. The ABCD Study provided detailed procedures on data acquisition and MRI image processing elsewhere[9,11,36]. We strictly followed their recommended exclusion criteria based on automated and manual QC review for each modality, listed under the *abcd_imgincl01* table[9]. These criteria involved not only the image quality but also MR neurological screening, behavioral performance, number of TRs left after scrubbing, the integrity of task presentation among others. We removed participants with an exclusion flag at any MRI indices, separately for each modality. We also applied the three IQR rule with listwise deletion to remove observations with outliers in any indices within each modality. Additionally, to adjust for between-site variability, we used an Empirical Bayes method, ComBat[57,58]. We applied ComBat to all modalities except for task-based fMRI, given that between-site variability was found to be negligible for task-based contrasts[58]. To prevent data leakage, we applied the IQR rule and Combat separately for first-layer training, second-layer training, baseline test, and follow-up test sets.

1-3) *Three Task-Based fMRI*

We used task-based fMRI from three tasks. First, in the working-memory "N-Back" task[36,37], children saw pictures of houses and emotional faces. Depending on the blocks, children reported if a picture matched either: (a) a picture that is shown 2 trials earlier (2-back), or (b) a picture that is shown at the beginning of the block (0-back). To focus on working-memory-related activity, we use the [2-back vs 0-back] linear contrast (i.e., high vs. low working memory load). Second, in the Monetary Incentive Delay (MID) task[36,59], children needed to respond before the target disappeared. And doing so would provide them with a reward, if and only if the target followed the "reward cue" (but not in "neural cue").To focus on reward anticipation-related activity, we used the [Reward Cue vs Neutral Cue] linear contrast. Third, in the Stop-Signal Task (SST)[36,38], children needed to withhold or interrupt their motor response to a "Go" stimulus when it is followed unpredictably by a Stop signal. To focus on inhibitory control-related activity, we used the [Any Stop vs Correct Go] linear contrast. Note that, for the SST, we used two additional exclusion criteria, *tfmri_sst_beh_glitchflag,* and *tfmri_sst_beh_violatorflag,* to address glitches in the task as recommended by the study[60,61]. For all tasks, we used the average contrast values across two runs. These values were embedded in the brain parcels based on FreeSurfer[62]'s Destrieux[40] and ASEG[39] atlases (148 cortical surface and 19 subcortical volumetric regions, resulting in 167 features for each task-based fMRI task).

4) *Resting-State fMRI (rs-fMRI)*

During rs-fMRI collection, the children viewed a crosshair for 20 minutes. The ABCD's preprocessing strategy has been published elsewhere[11]. Briefly, the study parcellated regions into 333 cortical-surface regions[63] and correlated their time-series[11]. They then grouped these correlations based on 13 predefined large-scale networks[63]: auditory, cingulo-opercular, cingulo-parietal, default-mode, dorsal-attention, frontoparietal, none, retrosplenial-temporal, salience, sensorimotor-hand, sensorimotor-mouth, ventral-attention, and visual networks. Note that "none" refers to regions that do not belong to any networks. After applying Fisher r to z transformation, the study computed mean correlations between pairs of regions within each large-scale network (n=13) and between large-scale networks (n=78) and provided these mean correlations in their Releases[9]. This resulted in 91 features for the rs-fMRI. Given that the correlations between (not within) large-scale networks were

highly collinear with each other (e.g., the correlation between auditory and cingulo-opercular was collinear with that between auditory and default-mode), we further decorrelated them using partial correlation. We first applied inverse Fisher r to z transformation, then partial correlation transformation, and then reapplied Fisher r to z transformation.

5) *Structural MRI (sMRI)*

The ABCD study processed sMRI, including cortical reconstruction and subcortical volumetric segmentation, using FreeSurfer[62]. Here we considered FreeSurfer-derived Destrieux[40] regional cortical thickness measures (n=148 cortical surface) and ASEG[39] regional subcortical volume measures (n=19), resulting in 167 features for sMRI. We also adjusted regional cortical thickness and volumetric measures using mean cortical thickness and total intracranial volume, respectively.

6) *Diffusion Tensor Imaging (DTI)*

Here we focused on fractional anisotropy (FA)[64]. FA characterizes the directionality of the distribution of diffusion within white matter tracts, which can indicate the density of fiber packing[64]. The ABCD study segmented major white matter tracts using AtlasTrack[11,65]. Here we considered FA of 23 major tracks, 10 of which were separately labeled for each hemisphere. These tracks included corpus callosum, forceps major, forceps minor, cingulate and parahippocampal portions of cingulum, fornix, inferior frontal occipital fasciculus, inferior longitudinal fasciculus, pyramidal/corticospinal tract, superior longitudinal fasciculus, temporal lobe portion of superior longitudinal fasciculus, anterior thalamic radiations and uncinate. This left 23 features for DTI.

Predictive Model Fitting

We started with the first-layer training set. Here we used standardized features from each modality to separately predict the G-Factor via the Elastic Net algorithm[31] from the glmnet package[66] (see Figure 1). As a general form of penalized regression, the Elastic Net requires two hyperparameters. First, the 'penalty' determines how strong the feature's slopes are regularised. Second, the 'mixture' determines the degree to which the regularisation is applied to the sum of squared coefficients (known as Ridge) vs. to the sum of absolute values of the coefficients (known as LASSO). We tuned these two hyperparameters using a 10-fold cross-validation grid search and selected the model with the lowest Mean Absolute Error (MAE). In the grid, we used 200 levels of the penalty from 10^-10 to 10, equally spaced on the logarithmic-10 scale and 11 levels of the mixture from 0 to 1 on the linear scale.

Once we obtained the final modality-specific models from the first-layer training set, we fit these models to data in the second-layer training set. This gave us six predicted values of the G-Factor from six modalities, and these are the features to predict the G-Factor in the second-layer training set. To handle missing observations when combining these modality-specific features, we applied an Opportunistic Stacking approach[12] by creating duplicates of each modality-specific feature. After standardization, we coded missing observations in one as an arbitrarily large value of 1000 and in the other as an arbitrarily small value of -1000, resulting in 12 features. That is, as long as a child had at least one modality available, we would be able to include this child in Stacked modeling.

We then used the Random Forests algorithm[32] from the ranger package[67] to predict the G-Factor from these 12 features[12,68]. Random Forests use a multitude of decision trees on

various sub-samples of the data and implement averaging to enhance prediction and to control over-fitting. We used 1000 trees and turned two hyperparameters. First 'mtry' is the number of features randomly sampled at each split. Second 'min_n' is the minimum number of observations in a node needed for the node to be split further. We implemented a 10-fold cross-validation grid search and selected the model with the lowest Root Mean Squared Error (RMAE). In the grid, we used 12 levels of the mtry from 1 to 12, and 101 levels of the min_n from 1 to 1000, both on the linear scale. This resulted in the "stacked" model that incorporated data across modalities.

We examined the predictive ability of the models between predicted vs. observed G-Factor, using Pearson's correlation (*r*), coefficient of determination (R^2, calculated using the sum of square definition), mean absolute error (MAE), and root mean squared error (RMSE). To investigate the prediction of the modality-specific models, we used the models tuned from the first-layer training set. To investigate the prediction of the stacked model, we used the model tuned from the second-layer training set. We used the baseline test set for out-of-sample, same-age predictive ability, while we used the follow-up test sets for out-of-sample, longitudinal predictive ability.

To test the stacked model's performance, we further split the test sets based on the presence of each modality. First, Stacked All required data with at least one modality present. This allowed us to examine the stacked model's performance when the missing values were all arbitrarily coded. Second, Stacked Complete required data with all modalities present. This represents the situation when the data were as clean as possible. Third, Stacked Best had the same missing values with the modality with the best prediction. This allowed us to make a fair comparison in performance between the stacked model and the model with the best modality, given their same noise level from missing value. We applied bootstrapping with 5,000 iterations to examine the difference in performance indices. Fourth, Stacked No Best did not have any data from the modality with the best prediction and had at least one modality present. This represents the highest level of noise possible. For the machine learning workflow, we used 'tidymodels' (www.tidymodels.org).

Feature Importance

To understand which features contribute to the prediction of the modality-specific, Elastic Net model, we applied permutation from the eNetXplorer[69] package to the first-layer training set. We first chose the best mixture from the previously run grid and fit two sets of several Elastic Net models. The first "target" models used the true G-Factor as the target, while the second "null" models used the randomly permuted G-Factor as the target. eNetXplorer split the data into 10 folds 100 times/runs. For each run, eNetXplorer performed cross-validation by repeatedly training the target models on 9 folds and tested on the leftover fold. Also in each cross-validation run, eNetXplorer trained the null models 25 times. eNetXplorer then used the mean of non-zero model coefficients across all folds in a given run as a coefficient for each run, k^r. Across runs, eNetXplorer weighted the mean of a model coefficient by the frequency of obtaining a non-zero model coefficient per run. Formally, we defined an empirical p-value as:

$$p = \frac{1}{1 + nr*np}\left\{1 + \sum_{r=1}^{nr}\sum_{p=1}^{np} \Theta\left(|k|_{null}^{r,p} - |k|_{target}^{r}\right)\right\}, (1)$$

where p is an empirical p-value, r is a run index, nr is the number of runs, p is a permutation index, np is the number of permutation, Θ is the right-continuous Heaviside step function, and $|k|$ is the magnitude of feature' coefficient. That is, to establish statistical significance for each feature, we used the proportion of runs in which the null models performed better than the target models. We plotted the target models' coefficients with $p < .05$ on the brain images using the ggseg[70] package.

To understand which modalities contributed strongly to the prediction of the stacked, Random Forests model, we applied conditional permutation importance (CPI) to the second-layer training set using the 'permimp' package[71]. The implementation was documented in detail elsewhere[71]. Briefly, the original permutation importance[32] shuffled the observations of one feature at a time while holding the target and other features in the same order. Researchers then examined decreases in prediction accuracy in the out-of-bag observations due to the permutation of some features. Stronger decreases are then assumed to reflect the importance of such features. However, this method has shown to be biased when there are correlated features[72]. CPI corrected for this bias by constraining the feature permutation to be within partitions of other features, which was controlled by the threshold 's' value. We used the default s value at .95, which assumed dependencies among features[71].

Testing whether the brain-based predictive models captured the relationships of the G-Factor with socio-demographic, psychological and genetic factors

Multimodal MRI

In addition to our analyses on longitudinal, brain-based predictive models, we also applied leave-one-site-out cross-validation to the baseline data. This enabled us to extract predicted values of the G-Factor based on multimodal MRI data at each held-out site.
Leave-one-site-out cross-validation also allowed us to examine the generalisability of the modality-specific and stacked models on different data collection sites. Different sites involved different MRI machines, experimenters as well as demographics across the US[41]. Moreover, using leave-one-site-out cross-validation also prevented having the participants from the same family in the training and test sets. Here, we first removed data from one site that only recruited 34 participants and removed participants from six families who were separately scanned at different sites. We then held out data from one site as a test set and divided the rest to be first- and second-layer training sets. We cross-validated predictive ability across these held-out sites. We applied the same modeling approach with the longitudinal predictive models, except for two configurations to reduce the amount of ram used and computational time. Specifically, in our grid search, we used 100 levels of penalty (as opposed to 200) for Elastic Net and limited the maximal min_n to 500 (as opposed to 1000) for Random Forests. For the stacked model, we tested its predictive ability on children with at least one modality (i.e., stacked all). We examined the out-of-site prediction between predicted vs. observed G-Factor at each held-out site using r, R^2, MAE, and RMSE.

Key socio-demographic and psychological factors

Similar to leave-one-site-out cross-validation for multimodal MRI, we also performed leave-one-site-out cross-validation to predict the G-Factor from key socio-demographic and psychological factors on the baseline data. This enabled us to extract predicted values of the G-Factor based on key socio-demographic and psychological factors at each held-out site. Here we applied the similar modeling approach with leave-one-site-out cross-validation for

multimodal MRI, except that we used only one layer of Elastic Net tuned with 200 levels of the penalty (from 10^-10 to 10) and 11 levels of the mixture (from 0 to 1). For pre-processing, we first imputed missing values of the categorical features via mode replacement and then converted them to dummy variables. We next normalised these dummy variables and all numerical features and the G-Factor. At the last pre-processing step, we used k-nearest neighbor with five neighbors to impute the missing values of the normalised, numerical features. We included 70 features[26] collected at the baseline (9-10 years old): child's mental health based on symptom scales in Child Behavioral Checklist[73] (8 features), primary caretaker's mental health based on personal strengths and symptom scales in Aseba Adult Self Report[73] and General Behavior Inventory-Mania[74] (9 features), child's personality based on Behavioral Inhibition System/Behavioral Activation System[75] and the UPPS-P Impulsive Behavior Scale[76] (9 features), child's sleep problems based on Sleep Disturbance Scale[77] (8 features), child's physical activities based on Youth Risk Behavior Survey[78] (4 features), child screen use[79] (4 features), parental use of alcohol, tobacco and marijuana after pregnancy based on Developmental History Questionnaire[80,81] (3 features), child developmental adversity (prematurity, birth complications and pregnancy complications) based on Developmental History Questionnaire[80,81] (3 features), child socio-demographics[82] including sex, race, bilingual use[83], parental marital status, parental education, parental income, household size, economic insecurities, area deprivation index[84], lead risk[85], crime reports[86], neighbourhood safety[87] and school environment, involvement and disengagement[88] (17 features) and child social interactions based on Parental Monitoring Scale[89], Child Report of Behavior Inventory[90], Strengths and Difficulties Questionnaire[91] and Moos Family Environment Scale[92] (5 features).

Polygenic scores

To extract predicted values of the G-Factor based on genetics, we used polygenic scores (PGS) for adult cognitive ability[29]. The ABCD study provided details on genotyping elsewhere[93]. Briefly, the study took saliva and whole blood samples and genotyped them using Smokescreen™ Array. The ABCD applied quality control based on calling signals and variant call rates, ran the Ricopili pipeline and imputed the data with TOPMED. We excluded problematic plates and data with a subject-matching issue. We further quality controlled the data using the genotypeqc function (https://github.com/ricanney/stata). We removed individuals with minimal or excessive heterozygosity, disproportionate levels of individual missingness (>2%) or insufficient sample replication (IBD <0.8). We also excluded SNPs based on minor allele frequency (<5%), call rate (<98%) or evidence for violations of Hardy-Weinberg equilibrium (P < 1E-10). We only included children with low genetic relatedness (more than 3rd-degree relative pairs; IBD ≥0.0422). We defined alleles associated with the G-Factor as those related to cognitive ability in a large-scale discovery GWAS (N = 257,841)[29]. Given the biases associated with generating PGS in samples that are ancestrally diverse, we restricted all analyses related to PGS to children of European ancestry[94] to match with the discovery samples[29]. We considered children to be genetically similar to the ancestry reference if they were within 4 standard deviations of the mean of the top 4 principal components (PCs) for the super-population individuals in the 1000 genomes Phase 3 reference genotypes. After extracting children of the ancestry reference, we recomputed multidimensional scaling. The final sample for PGS included 4,814 children (2,263 females; M_{age}= 9.94 (SD=.61) years). Using PLINK 1.9 via the summaryqc2score function (https://github.com/ricanney/stata), we computed PGS as the Z-scored across 10 PGS thresholds from p<.5 to .00000001, weighted mean number of risk alleles in approximate linkage equilibrium, derived from imputed autosomal SNPs. We focused on the PGS threshold that demonstrated the strongest relationship with the G-Factor.

Mediation analyses

To examine the extent to which brain-based, stacked predictive models of the G-Factor capture the relationship of the G-Factor with the socio-demographic, psychological and genetic factors, we applied mediation analyses[95]. Here, we created three separate mediation models: (1) for the key socio-demographic and psychological factors, (2) for the genetic factor and (3) for the two combined. All models treated the predicted values of the G-Factor based on the brain-based, stacked predictive models at all held-out sites from the leave-one-site-out cross validation as the mediator. Additionally, they also treated the observed (i.e., real) values of the G-Factor as the dependent variable. The difference between the three models was on the independent variable. For the socio-demographic and psychological factors, we used the predicted values of the G-Factor based on the 68 socio-demographic and psychological features at all held-out sites from the leave-one-site-out cross validation as the independent variable. Conversely, for the genetic factor, we used the PGS of cognitive ability[29] as the independent variable. Lastly, for the models that combined the two, we use both (1) the predicted values of the G-Factor based on the socio-demographic and psychological features and (2) the PGS of cognitive ability as two separate independent variables that were co-variated with each other.
To control for population stratification in genetics, we also included four PCs as control variables in the mediation analysis for the genetic factor and for the combined model.

To implement the mediation analyses, we used structural equation modeling (SEM) with 5,000 bootstrapping iterations via lavaan[55]. We specifically calculated the *indirect effects* to show whether the relationships of the G-Factor with the socio-demographic, psychological and genetic factors were significantly explained by the brain-based, stacked predictive models. Along with the indirect effects, we also computed the *proportion mediated* to demonstrate the proportion of variance accounted for by the brain-based models.

References

1. Jensen, A. R. *The g factor: the science of mental ability*. (Praeger, 1998).
2. Schmidt, F. L. & Hunter, J. General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology* **86**, 162–173 (2004).
3. Der, G., Batty, G. D. & Deary, I. J. The association between IQ in adolescence and a range of health outcomes at 40 in the 1979 US National Longitudinal Study of Youth. *Intelligence* **37**, 573–580 (2009).
4. Batty, G. D., Deary, I. J. & Gottfredson, L. S. Premorbid (early life) IQ and Later Mortality Risk: Systematic Review. *Annals of Epidemiology* **17**, 278–288 (2007).
5. Flynn, J. R. *What is intelligence? beyond the Flynn effect*. (Cambridge University Press, 2009).
6. Rosenberg, M. D., Casey, B. J. & Holmes, A. J. Prediction complements explanation in understanding the developing brain. *Nature Communications* **9**, 589 (2018).
7. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biological Psychiatry* **88**, 818–828 (2020).
8. Tucker-Drob, E. M. Differentiation of cognitive abilities across the life span. *Developmental Psychology* **45**, 1097–1118 (2009).
9. Yang, R. & Jernigan, Terry; Adolescent Brain Cognitive Development Study (ABCD) - Annual Release 3.0. doi:10.15154/1519007.
10. Fassbender, C., Mukherjee, P. & Schweitzer, J. B. Minimizing noise in pediatric task-based functional MRI; Adolescents with developmental disabilities and typical development. *NeuroImage* **149**, 338–347 (2017).
11. Hagler, D. J. *et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage* **202**, 116091 (2019).
12. Engemann, D. A. *et al.* Combining magnetoencephalography with magnetic resonance

imaging enhances learning of surrogate-biomarkers. *eLife* **9**, e54055 (2020).

13. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nature Neuroscience* **6**, 316–322 (2003).
14. Waiter, G. D. *et al.* Exploring possible neural mechanisms of intelligence differences using processing speed and working memory tasks: An fMRI study. *Intelligence* **37**, 199–206 (2009).
15. Sripada, C. *et al.* Prediction of neurocognition in youth from resting state fMRI. *Molecular Psychiatry* **25**, 3413–3421 (2020).
16. Pamplona, G. S. P., Santos Neto, G. S., Rosset, S. R. E., Rogers, B. P. & Salmon, C. E. G. Analyzing the association between functional connectivity of the brain and intellectual performance. *Front. Hum. Neurosci.* **9**, (2015).
17. Dubois, J., Galdi, P., Paul, L. K. & Adolphs, R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 20170284 (2018).
18. Narr, K. L. *et al.* Relationships between IQ and Regional Cortical Gray Matter Thickness in Healthy Adults. *Cerebral Cortex* **17**, 2163–2171 (2007).
19. Góngora, D., Vega-Hernández, M., Jahanshahi, M., Valdés-Sosa, P. A. & Bringas-Vega, M. L. Crystallized and fluid intelligence are predicted by microstructure of specific white-matter tracts. *Human Brain Mapping* **41**, 906–916 (2020).
20. Genç, E. *et al.* Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence. *Nature Communications* **9**, 1905 (2018).
21. Rasero, J., Sentis, A. I., Yeh, F.-C. & Verstynen, T. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *bioRxiv* 2020.09.01.278747 (2020) doi:10.1101/2020.09.01.278747.
22. Jiang, R. *et al.* Multimodal data revealed different neurobiological correlates of intelligence between males and females. *Brain Imaging and Behavior* **14**, 1979–1993 (2020).
23. Farah, M. J. *et al.* Childhood poverty: Specific associations with neurocognitive

development. *Brain Research* **1110**, 166–174 (2006).

24. Biederman, J. *et al.* Impact of Executive Function Deficits and Attention-Deficit/Hyperactivity Disorder (ADHD) on Academic Outcomes in Children. *Journal of Consulting and Clinical Psychology* **72**, 757–766 (2004).
25. Goodall, J. *et al.* Neurocognitive Functioning in Depressed Young People: A Systematic Review and Meta-Analysis. *Neuropsychol Rev* **28**, 216–231 (2018).
26. Kirlic, N. *et al.* Extracurricular Activities, Screen Media Activity, and Sleep May Be Modifiable Factors Related to Children's Cognitive Functioning: Evidence From the ABCD Study®. *Child Development* **n/a**, (2021).
27. Plomin, R. & Deary, I. J. Genetics and intelligence differences: five special findings. *Molecular Psychiatry* **20**, 98–108 (2015).
28. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry* **16**, 996–1005 (2011).
29. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* **50**, 1112–1121 (2018).
30. Allegrini, A. G. *et al.* Genomic prediction of cognitive traits in childhood and adolescence. *Mol Psychiatry* **24**, 819–827 (2019).
31. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **67**, 301–320 (2005).
32. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
33. Candia, J. & Tsang, J. S. eNetXplorer: an R package for the quantitative exploration of elastic net families for generalized linear models. *BMC Bioinformatics* **20**, 189 (2019).
34. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307 (2008).
35. 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. *Human Brain Mapping* **41**, 3186–3197 (2020).

36. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience* **32**, 43–54 (2018).
37. Barch, D. M. *et al.* Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169–189 (2013).
38. Whelan, R. *et al.* Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience* **15**, 920–925 (2012).
39. Fischl, B. *et al.* Whole Brain Segmentation. *Neuron* **33**, 341–355 (2002).
40. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* **53**, 1–15 (2010).
41. Garavan, H. *et al.* Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience* **32**, 16–22 (2018).
42. Clark, D. B. *et al.* Biomedical ethics and clinical oversight in multisite observational neuroimaging studies with children and adolescents: The ABCD experience. *Developmental Cognitive Neuroscience* **32**, 143–154 (2018).
43. Snellen, H. Letterproeven Tot Bepaling Der Gezictsscherpte, Utrecht, Weyers (Dutch Edition). *Also published in many languages as Optotypi ad Visum Determinandum* (1862).
44. Luciana, M. *et al.* Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. *Developmental Cognitive Neuroscience* **32**, 67–79 (2018).
45. Thompson, W. K. *et al.* 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. *Developmental Cognitive Neuroscience* **36**, 100606 (2019).
46. Gershon, R. C. *et al.* Language measures of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society* **20**, 642–651 (2014).
47. Bleck, T. P., Nowinski, C. J., Gershon, R. & Koroshetz, W. J. What is the NIH Toolbox, and what will it mean to neurology? *Neurology* **80**, 874–875 (2013).
48. Eriksen, B. A. & Eriksen, C. W. Effects of noise letters upon the identification of a target

letter in a nonsearch task. *Perception & Psychophysics* **16**, 143–149 (1974).

49. Carlozzi, N. E., Tulsky, D. S., Kail, R. V. & Beaumont, J. L. Vi. Nih Toolbox Cognition Battery (cb): Measuring Processing Speed. *Monographs of the Society for Research in Child Development* **78**, 88–102 (2013).
50. Bauer, P. J. *et al.* Iii. Nih Toolbox Cognition Battery (cb): Measuring Episodic Memory. *Monographs of the Society for Research in Child Development* **78**, 34–48 (2013).
51. Daniel, M. H. & Wahlstrom, D. Equivalence of Q-interactive™ and Paper Administrations of Cognitive Tasks: WISC®–V. 13 (2014).
52. Ang, Y.-S., Frontero, N., Belleau, E. & Pizzagalli, D. A. Disentangling vulnerability, state and trait features of neurocognitive impairments in depression. *Brain* awaa314 (2020) doi:10.1093/brain/awaa314.
53. Pornpattananangkul, N. *et al.* Motivation and Cognitive Abilities as Mediators between Polygenic Scores and Psychopathology in Children. *medRxiv* 2020.06.08.20123877 (2020) doi:10.1101/2020.06.08.20123877.
54. Jorgensen, T. D. *et al.* semTools: Useful tools for structural equation modeling. *R package version 0.5-1* (2018).
55. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Soft.* **48**, (2012).
56. Epskamp, S. semPlot: Unified visualizations of structural equation models. *Structural Equation Modeling* **22**, 474–483 (2015).
57. Fortin, J.-P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161**, 149–170 (2017).
58. Nielson, D. M. *et al.* Detecting and harmonizing scanner differences in the ABCD study - annual release 1.0. *bioRxiv* 309260 (2018) doi:10.1101/309260.
59. Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage* **12**, 20–27 (2000).
60. Bissett, P. G., Hagen, M. P. & Poldrack, R. A. A cautionary note on stop-signal data from the Adolescent Brain Cognitive Development [ABCD] study. *bioRxiv* 2020.05.08.084707

(2020) doi:10.1101/2020.05.08.084707.

61. Garavan, H. *et al.* The ABCD Stop Signal Data: Response to Bissett et al. *bioRxiv* 2020.07.27.223057 (2020) doi:10.1101/2020.07.27.223057.
62. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage* **9**, 179–194 (1999).
63. Gordon, E. M. *et al.* Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cereb. Cortex* **26**, 288–303 (2016).
64. Alexander, A. L., Lee, J. E., Lazar, M. & Field, A. S. Diffusion tensor imaging of the brain. *Neurotherapeutics* **4**, 316–329 (2007).
65. Hagler, D. J. *et al.* Automated white-matter tractography using a probabilistic diffusion tensor atlas: Application to temporal lobe epilepsy. *Human Brain Mapping* **30**, 1535–1547 (2009).
66. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1–22 (2010).
67. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1–17 (2017).
68. Josse, J., Prost, N., Scornet, E. & Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv:1902.06931 [cs, math, stat]* (2020).
69. Candia, J. & Tsang, J. S. ENetXplorer: An R package for the quantitative exploration of elastic net families for generalized linear models. *BMC Bioinformatics* **20**, 1–11 (2019).
70. Mowinckel, A. M. & Vidal-Piñeiro, D. Visualization of Brain Statistics With R Packages *ggseg* and *ggseg3d*. *Advances in Methods and Practices in Psychological Science* **3**, 466–483 (2020).
71. Debeer, D. & Strobl, C. Conditional permutation importance revisited. *BMC Bioinformatics* **21**, 307 (2020).
72. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25 (2007).

73. Achenbach, T. M., Ivanova, M. Y. & Rescorla, L. A. Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry* **79**, 4–18 (2017).
74. Youngstrom, E. A., Frazier, T. W., Demeter, C., Calabrese, J. R. & Findling, R. L. Developing a 10-Item Mania Scale From the Parent General Behavior Inventory for Children and Adolescents. *J Clin Psychiatry* **69**, 0–0 (2008).
75. Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology* **67**, 319–333 (1994).
76. Zapolski, T. C. B., Stairs, A. M., Settles, R. F., Combs, J. L. & Smith, G. T. The Measurement of Dispositions to Rash Action in Children. *Assessment* **17**, 116–125 (2010).
77. Bruni, O. *et al.* The Sleep Disturbance Scale for Children (SDSC) construction and validation of an instrument to evaluate sleep disturbances in childhood and adolescence. *Journal of Sleep Research* **5**, 251–261 (1996).
78. YRBSS | Youth Risk Behavior Surveillance System | Data | Adolescent and School Health | CDC. https://www.cdc.gov/healthyyouth/data/yrbs/index.htm (2020).
79. Bagot, K. S. *et al.* Current, future and potential use of mobile and wearable technologies and social media data in the ABCD study to increase understanding of contributors to child health. *Developmental Cognitive Neuroscience* **32**, 121–129 (2018).
80. Kessler, R. C. *et al.* Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). *International Journal of Methods in Psychiatric Research* **18**, 69–83 (2009).
81. Merikangas, K. R., Avenevoli, S., Costello, E. J., Koretz, D. & Kessler, R. C. National Comorbidity Survey Replication Adolescent Supplement (NCS-A): I. Background and Measures. *Journal of the American Academy of Child & Adolescent Psychiatry* **48**, 367–379 (2009).
82. Zucker, R. A. *et al.* Assessment of culture and environment in the Adolescent Brain and

Cognitive Development Study: Rationale, description of measures, and early data. *Developmental Cognitive Neuroscience* **32**, 107–120 (2018).

83. Dick, A. S. *et al.* No evidence for a bilingual executive function advantage in the ABCD study. *Nature Human Behaviour* **3**, 692–701 (2019).
84. Kind, A. J. H. *et al.* Neighborhood Socioeconomic Disadvantage and 30-Day Rehospitalization. *Ann Intern Med* **161**, 765–774 (2014).
85. Frostenson, S., April 6, S. K. on, 2016 & ET, 8:50 a m. Where is the lead exposure risk in your community? *Vox.com* http://www.vox.com/a/lead-exposure-risk-map.
86. United States Department Of Justice. Office Of Justice Programs. Federal Bureau Of Investigation. Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data, 2010: Version 2. (2012) doi:10.3886/ICPSR33523.V2.
87. Echeverria, S. E., Diez-Roux, A. V. & Link, B. G. Reliability of self-reported neighborhood characteristics. *J Urban Health* **81**, 682–701 (2004).
88. Stover, P. J., Harlan, W. R., Hammond, J. A., Hendershot, T. & Hamilton, C. M. PhenX: a toolkit for interdisciplinary genetics research. *Current Opinion in Lipidology* **21**, 136–140 (2010).
89. Chilcoat, H. D. & Anthony, J. C. Impact of Parent Monitoring on Initiation of Drug Use through Late Childhood. *Journal of the American Academy of Child & Adolescent Psychiatry* **35**, 91–100 (1996).
90. Schaefer, E. S. A configurational analysis of children's reports of parent behavior. *Journal of Consulting Psychology* **29**, 552–557 (1965).
91. Goodman, R., Meltzer, H. & Bailey, V. The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. *International Review of Psychiatry* **15**, 173–177 (2003).
92. Moos, R. H. & Humphrey, B. *Preliminary manual for family environment scale, work environment scale, group environment scale*. (Consulting Psychologists Press, 1974).
93. Uban, K. A. *et al.* Biospecimens and the ABCD study: Rationale, methods of collection, measurement and early data. *Developmental Cognitive Neuroscience* **32**, 97–106

(2018).

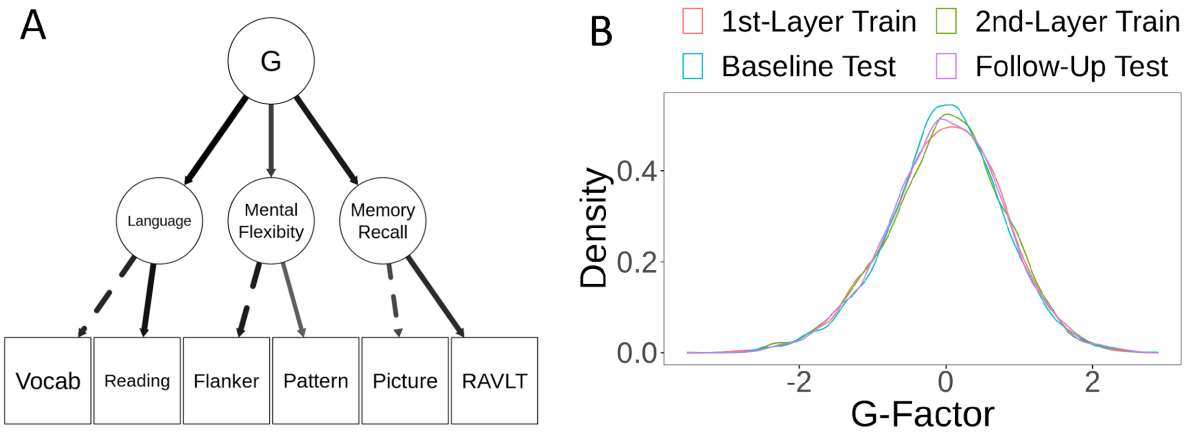
94. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10**, 3328 (2019).

95. MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. Mediation Analysis. *Annu. Rev. Psychol.* **58**, 593–614 (2007).

Supplementary

<u>The G-Factor</u>

Supplementary Figure 1 shows the CFA of the G-Factor. The higher-order model of the G-Factor shows a good fit (*scaled, robust CFI*=.995*, TLI*=.988 and *RMSE*= .029 (*90%CI*=.015-.043), and the G-Factor has high internal consistency (*OmegaL2=.78*). The distribution of the G-Factor factor scores was similar across data splits.



Supplementary Figure 1. Confirmatory Factor Analysis (CFA) of the G-Factor. 2A shows the higher-order model for the G-Factor. Line thickness reflects the magnitude of standardized parameter estimates. The dotted lines indicate marker variables that were fixed to 1. Vocab = Picture Vocabulary; Reading = Oral Reading Recognition; Pattern = Pattern Comparison Processing; Picture = Picture Sequence Memory; RAVLT = Rey-Auditory Verbal Learning. 2B shows the distribution of the G-Factor factor score across the four data splits.

Hyperparameters for the Longitudinal Predictive Models for Multimodal MRI

From the first-layer training set, we found the best-tuned Elastic Net's mixture was 0 for N-Back task-based fMRI, rs-fMRI and DIT, .1 for sMRI, and 1 for SST and MID task-based fMRI. Thus, whether to have brain features shrunk together (the Ridge solution, mixture close to 0) or to have certain features from all features selected (the Lasso solution, mixture close to 1) depended on the modality. The averaged Elastic Net's penalty was .339 (SD = .45). From the second-layer training set, we found the best-tuned Random Forests' mtry (the number of features randomly sampled at each path split) at 4 and min_n (the minimum number of observations in a node per path split) at 170.