Reliability modelling of resting-state functional connectivity

Jalmar Teeuw [a],*, Hilleke E. Hulshoff Pol [a], Dorret I. Boomsma [b], Rachel M. Brouwer [a]

[a] Brain Center Rudolf Magnus and Department of Psychiatry, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, Netherlands
[b] Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

ABSTRACT

Resting-state functional magnetic resonance imaging (rs-fMRI) has an inherently low signal-to-noise ratio largely due to thermal and physiological noise that attenuates the functional connectivity (FC) estimates. Such attenuation limits the reliability of FC and may bias its association with other traits. Low reliability also limits heritability estimates. Classical test theory can be used to obtain a true correlation estimate free of random measurement error from parallel tests, such as split-half sessions of a rs-fMRI scan.

We applied a measurement model to split-half FC estimates from the resting-state fMRI data of 1003 participants from the Human Connectome Project (HCP) to examine the benefit of reliability modelling of FC in association with traits from various domains. We evaluated the efficiency of the measurement model on extracting a stable and reliable component of FC and its association with several traits for various sample sizes and scan durations. In addition, we aimed to replicate our previous findings of increased heritability estimates when using a measurement model in a longitudinal adolescent twin cohort.

The split-half measurement model improved test-retest reliability of FC on average with +0.33 points (from +0.49 to +0.82), improved strength of associations between FC and various traits on average 1.2-fold (range 1.09–1.35), and increased heritability estimates on average with +20% points (from 39% to 59%) for the full HCP dataset. On average, about half of the variance in split-session FC estimates was attributed to the stable and reliable component of FC. Shorter scan durations showed greater benefit of reliability modelling (up to 1.6-fold improvement), with an additional gain for smaller sample sizes (up to 1.8-fold improvement).

Reliability modelling of FC based on a split-half using a measurement model can benefit genetic and behavioral studies by extracting a stable and reliable component of FC that is free from random measurement error and improves genetic and behavioral associations.

1. Introduction

Resting-state functional connectivity has become a popular method to study the functional organization of the human brain (Biswal et al., 1995; Greicius et al., 2002; van den Heuvel and Hulshoff Pol, 2010). Functional connectivity has shown promise as a potential biomarker for its association with neuropsychiatric and neurological disorders (Hager and Keshavan, 2015; Hohenfeld et al., 2018; Whitfield-Gabrieli and Ford, 2012; Zhang and Raichle, 2010; Fox et al., 2014). In addition, it is associated with various cognitive and behavioural traits (Vaidya and Gordon, 2013; Basten et al., 2015; Shen et al., 2018; Smith et al., 2015; Toschi et al., 2018), and functional connectivity is heritable to a certain extent (Ge et al., 2017; Colclough et al., 2017; Adhikari et al., 2018; Teeuw et al., 2019). However, for resting-state functional connectivity to become a biomarker, it needs to be a reli-

able and state-independent endophenotype (Gottesman and Todd, 2003; Beauchaine and Constantino, 2017).

Resting-state functional magnetic resonance imaging (rs-fMRI) has an inherently low signal-to-noise ratio (SNR) primarily due to thermal noise of the magnetic resonance (MR) scanner and physiological noise from the subject, such as head motion, cardiovascular and respiratory activity (Logothetis and Wandell, 2004; Bianciardi et al., 2009; Murphy et al., 2013; Liu et al., 2016). These sources of noise are mixed into the blood-oxygenation level dependant (BOLD) signal and may lead to biased estimates of resting-state functional connectivity (Birn et al., 2014; Siegel et al., 2017). Numerous strategies have been proposed to clean the BOLD signal from these unwanted sources of noise and recover the signals related to neural activity (Burgess et al., 2016; Caballero-Gaudes and Reynolds, 2017; Power et al., 2015). However, no perfect solution exists (Bright and Murphy, 2015; Varikuti et al., 2017; Parkes et al., 2018; Krishnamurthy et al., 2018; Lindquist, 2020). Noise

---

in the BOLD signal has a negative impact on the reliability of the signal and results in the attenuation of functional connectivity estimates based on temporal correlation between two BOLD signals (Spearman, 1904; Wang, 2010; Birn et al., 2014; Mueller et al., 2015). Even with state-of-the-art procedures, test-retest reliability of functional connectivity is only poor to moderate for short scans (i.e. less than 30 min) (Noble et al., 2019; Chen et al., 2015; Shah et al., 2016), with large variation in reliability for the different connections measured in a single individual over an extended period of time (Choe et al., 2015). Although reliability of functional connectivity can be improved by increasing the scan duration up to 1.5 hour (Birn et al., 2013; Laumann et al., 2015), this approach is not always feasible due to the burden on the subject or the cost and availability of MRI. The limited reliability of functional connectivity for scan lengths that are typically used in resting-state fMRI studies puts an upper bound on the heritability estimates of functional connectivity and its association with traits (Vul et al., 2009; Neale and Cardon, 1992; Ge et al., 2017). Thereby making it difficult to reliably identify functional connections in the brain that are associated with particular traits (Geerligs et al., 2017; Kruschwitz et al., 2018).

In Classical Test Theory, the true score of a measure can be obtained from the observed score if the error term is known: *observed score = true score + error term* (Streiner, 2003; Miller, 1995). The error term can be approximated from 'parallel scores' (i.e. repeated measures). When it is not feasible to acquire two full measurements, two parallel half-score measures might be an option; e.g. an odd-even split in an event-related study design (van Baal et al., 1998; van Beijsterveldt et al., 2001). Resting-state fMRI data is uniquely suited for 'split-half' reliability modelling because of the temporal nature of the BOLD signal to create two parallel half-score measures by splitting scan session data into two or more parts (Brandmaier et al., 2018). For associations, such as functional connectivity based on the temporal correlation of two BOLD signals, the error term is defined by the reliability of the two measures (Spearman, 1904). The true association can be obtained by scaling the observed association with a factor inversely proportional to the reliability (Mueller et al., 2015; Golestani and Goodyear, 2011). However, classical disattenuation requires the correction factor to be known a priori and has the risk of overcorrecting the association (Wang, 2010). Instead, a structural equation measurement model can be applied to the half-score measures to derive a latent variable representing the trait of interest that is "free" of measurement error without the need of an a prior correction factor or the risk of overcorrection (Brandmaier et al., 2018; Cooper et al., 2019). Such measurement models have previously been applied in twin studies to separate the variance attributed to measurement error from genetic and environmental variance components to obtain a robust heritability estimate for the reliable part of the variation (van Baal et al., 1998; van Beijsterveldt et al., 2001; Ge et al., 2017; Teeuw et al., 2019). The measurement model is however suited for many types of measures and study designs, as long as some form of parallel scores can be obtained. This implies that reliability modelling can be useful for the typical cross-sectional resting-state fMRI dataset of unrelated individuals. However, little is known about the effectiveness of reliability modelling of resting-state MRI functional connectivity and its ability to uncover the true associations between functional connectivity and other traits.

Here, we examine the benefits of reliability modelling of functional connectivity in association with physiological, cognitive and behavioral traits. For that purpose, a measurement model was applied to functional connectivity estimates from the resting-state functional MRI scans of the Human Connectome Project Young Adult (HCP-YA) cohort (van Essen et al., 2013). The efficiency of reliability modelling was evaluated for various sample sizes and scan durations on (i) the ability to extract a stable and reliable component of functional connectivity and (ii) the improvement in the associations of functional connectivity with various traits. We also aimed to replicate our previous findings of increased heritability estimates for the reliable component of functional connectivity (Teeuw et al., 2019). Finally, we investigate the use of the measurement

Table 1

Demographics table for participants included in the extensively processed fMRI data package of the Human Connectome Project Young Adult cohort.

Measure	Statistic
Participants (N)	1003
Families (N)	429
Age range (min; max)	22–37 years
Age (mean±SD)	28.71±3.71 years
Sex ratio (female:male)	534:469

Abbreviations: SD = standard deviation of the mean.

model for a typical study that comprises a single resting-state fMRI session.

2. Materials and methods

2.1. Human connectome project

We utilized data from the publicly available *extensively processed fMRI data* package[1] that is part of the Human Connectome Project Young Adult cohort (van Essen et al., 2013). The package provides data for 1003 related individuals (siblings, including monozygotic- and dizygotic twins; aged 22 to 37 years) from 429 families with four complete runs of resting-state fMRI scans (**Table 1**) and consists of precomputed denoised and centered BOLD signal time series for nodes in the brain based on group-ICA decomposition of the data at various decomposition levels. The acquisition parameters and processing of this data have been described elsewhere[2] (Smith et al., 2013; Glasser et al., 2013; see supplementary methods for summary). All analyses were performed using the group-ICA decomposition with 50 nodes (**Supplementary Figure S1**).

2.2. Functional connectivity

Functional connectivity estimates were obtained by calculating the temporal correlation coefficient between the BOLD time series of two nodes using Pearson correlation (Biswal et al., 1995; van den Heuvel and Hulshoff Pol, 2010). Functional connectivity was estimated for different temporal blocks of the time series data to provide full-, half-, and quarter-score estimates of functional connectivity for the purpose of reliability modelling (**Fig. 1**A). All statistical and mathematical operations on functional connectivity were performed on Fisher's r-to-Z transformed functional connectivity estimates.
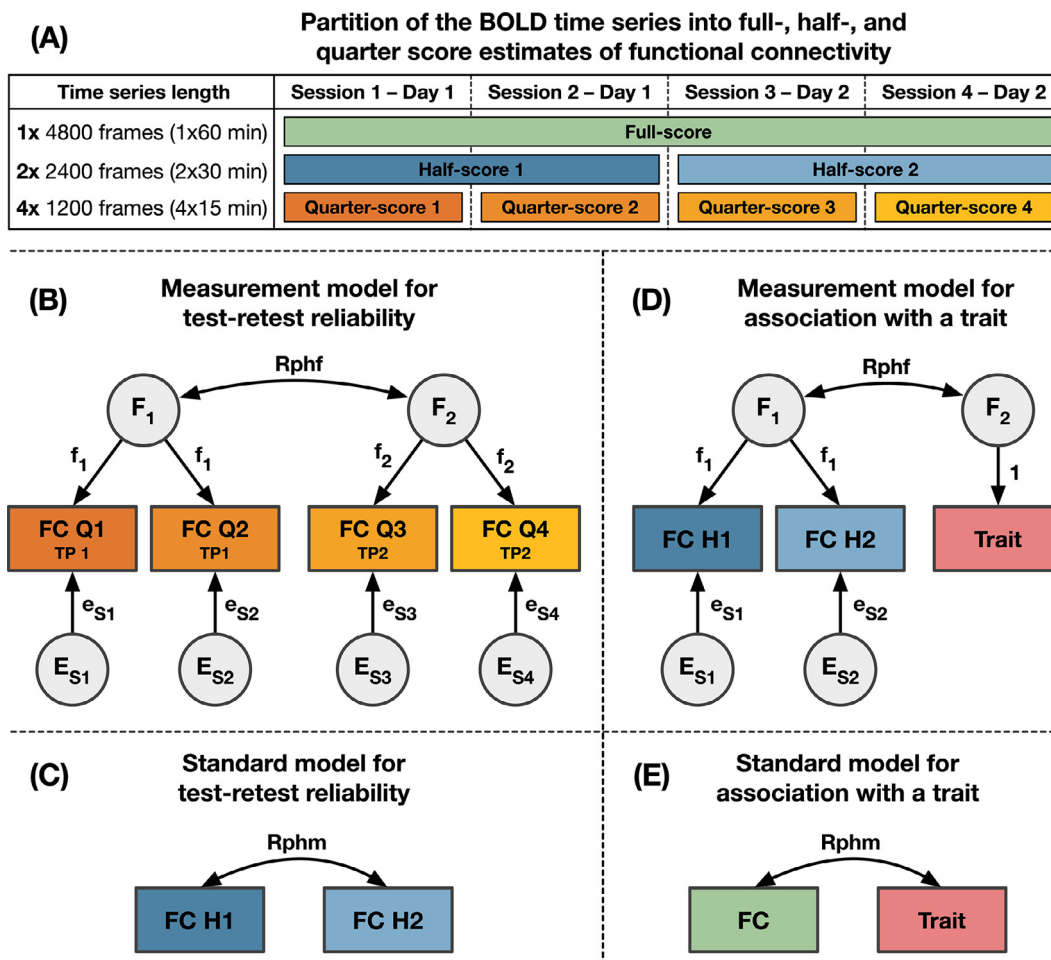
2.3. Reliability modelling of functional connectivity

A measurement model is applied to half-score measures of functional connectivity to extract a reliable component of functional connectivity represented by the latent variable F_i (**Fig. 1**). Variance shared between half-score measures that can be attributed to the latent variable is quantified by the path coefficient (f_i), which are constrained to be equal in proportion of variance for both half score measures. Any residual variance of the half-score measures, which includes measurement error, is considered noise, and is represented by the measurement-specific latent variables Es_i quantified by the path coefficients es_i.

To demonstrate the ability of the measurement model to extract a stable and reliable component of functional connectivity, test-retest reliability is estimated between two latent variables F_1 and F_2 that each

---

[1] https://www.humanconnectome.org/study/hcp-young-adult/document/extensively-processed-fmri-data-documentation

[2] https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP1200-DenseConnectome+PTN+Appendix-July2017.pdf

Fig. 1. Partitioning of the BOLD time series and the measurement models used in the analyses. (**A**) Decomposition of the BOLD time series data into temporal blocks for the estimation of full-, half-, and quarter-score functional connectivity based on the original full-length time series with 4800 frames (approximately 1 h) of resting-state functional MRI data. (**B**) A measurement model with two latent variables each representing a reliable component of functional connectivity across a set of half-score measures of functional connectivity. The model is used to estimate *corrected* test-retest reliability between the two reliable components of functional connectivity at different scan sessions (*Rphf*) quantified by the correlation between the two latent variables. (**C**) A standard association model is used to estimate the *uncorrected* test-retest reliability between the observed half-score measures of functional connectivity (*Rphm*) quantified by the correlation between the two observed variables. (**D**) A measurement model with one latent variable representing the reliable component of functional connectivity of two half-score measures and another latent variable representing the trait. This model is used to estimate the *corrected* association between the reliable component of functional connectivity and the trait (*Rphf*) quantified by the correlation between the two latent factors. (**E**) A standard association model is used to estimate the *uncorrected* correlation between the observed full-score measure of functional connectivity and the trait.

represent a reliable component of functional connectivity based on two half-score measures of functional connectivity from independent scan sessions acquired on different days (**Fig. 1**B). The *corrected* test-retest reliability (*Rphf*) is estimated as the correlation between the two latent variables. Test-retest reliability from the measurement model is compared to the *uncorrected* test-retest reliability of functional connectivity estimated as the correlation between the two half-score measures of functional connectivity (**Fig. 1**C). Improvement in corrected test-retest reliability is determined by the pointwise difference compared to the uncorrected test-retest reliability estimate (*Rphm*).

The measurement model can be adapted to estimate the *corrected* association between the reliable component of functional connectivity and a trait (**Fig. 1**D). For comparison, the *uncorrected* association between the full-score measure of functional connectivity and the trait is estimated and compared over all connections (**Fig. 1**E). The overall improvement in association strength (i.e. the average improvement factor) is determined by the slope coefficient of the linear regression of the association strengths from the measurement model onto the association strengths from the standard association model.

All models included fixed effects of sex, age, head motion and HCP processing pipeline version as covariates on the means of the functional connectivity estimates, and fixed effects of sex and age as covariates on the means of the trait. Head motion was approximated by mean framewise displacement (Power et al., 2012; **Supplementary Figure 2**). All models were specified in the *OpenMx* structural equation modelling (SEM) software package (Neale et al., 2016; Boker et al., 2018) for *R* (R Core Team, 2018). The definition of the measurement model in OpenMx is provided as a supplementary data file (**Supplementary Data File F1**), and is available on GitHub with an example (Teeuw, GitHub 2020).

The suitability of applying a measurement model to the data was assessed with the goodness of fit metrics Comparative Fit Index (CFI) and root-mean-square error of approximation (RMSEA). Model fits with a CFI > 0.95 and RMSEA < 0.05 were deemed a good fit, model fits with a CFI > 0.90 and RMSEA < 0.08 were deemed an acceptable fit, and the remaining models (CFI < 0.90 or RMSEA > 0.08) were deemed an unsuitable fit (Hu and Bentler, 1999; Browne and Cudeck, 1992). All models with an unsuitable fit were excluded from statistical analyses.

Table 2

Parameters and their values used in the evaluation of the minimal requirements on the input dataset for the measurement model.

Parameter	Values	Unit
Total scan duration	5, 7.5, 10, 15, 20, 30, 45, 60	Minutes
Four sessions across two	———————————————	——–
different days	400, 600, 800, 1200, 1600, 2400, 3600, 4800	Volumes
Total scan duration	5, 7.5, 10, 15, 20, 25, 30	Minutes
Two sessions on the same day	———————————————	——–
	400, 600, 800, 1200, 1600, 2000, 2400	Volumes
Total scan duration	5, 7.5, 10, 12.5, 15	Minutes
Single session scan data	———————————————	——–
	400, 600, 800, 100, 1200	Volumes
Number of participants per sample	25, 50, 75, 100, 150, 200, 250, 300, 400, 500	Count

Total number of combinations of parameter values per connection (e.g. for four sessions across two different days): 8 scan durations x 10 sample sizes x 100 sample draws = 8000 models per connection.

2.3.1. Physiological, cognitive, and behavioral traits

The Human Connectome Project provides rich phenotypic information on the participants.[3] Because of the computational complexity of the measurement models, we used the results from the standard association model (**Fig. 1**E) applied to 110 continuous and normally distributed measures (**Supplementary Data File F2**) to identify seven representative traits from different domains for extensive analysis (see Supplementary Materials for details; **Supplementary Table S1**): five traits (*BPDiastolic* – diastolic blood pressure levels; *CogTotalComp_AgeAdj* – total composite score on cognition adjusted for age; *WM_Task_2bk_Acc* – accuracy on all condition in the 2-back working memory task; *Emotion_Task_Median_RT* – median response time for each condition in the emotion task; and *PicVocab_AgeAdj* – picture vocabulary test score adjusted for age) were among the most strongly associated measures with functional connectivity at any individual connection, and two traits (*Gambling_Task_Reward_Perc_Larger* – percentage of trials that received a 'larger' prediction in the gambling task; and *Taste_AgeAdj* – score on the taste intensity test adjusted for age) were chosen because they were only weakly associated with functional connectivity. For these seven traits, we estimated the association with all functional connectivity measures.

For the remaining 103 traits, measurement models were computed only for the top 20 connections most strongly associated with functional connectivity and the 5 connections with the weakest associated with functional connectivity (i.e. near zero association) based on the results from the standard association model. The connections were selected independently for each trait (**Supplementary Data File F2**). This sampling scheme provides a good approximation of the actual improvement factor in association strength for the seven fully sampled traits, with mean absolute difference in improvement factor 2% (range from 0% to 5%; **Supplementary Table S2**).

2.3.2. Heritability of functional connectivity

To emphasize that the measurement model can be applied to the typical dataset consisting of unrelated individuals, up to this point we assumed that the subjects were independent. However, the Human Connectome Project cohort includes families with monozygotic and dizygotic twins and their siblings, and families with non-twin siblings (**Supplementary Table S3**). We aimed to replicate our previous findings of increased heritability estimates for the reliable and stable component of functional connectivity in a longitudinal adolescent twin cohort (Teeuw et al., 2019). In brief, genetic modelling of data from twins and siblings allows for the decomposition of the variance of a trait (V) into genetic and environmental components. Often, three variance components representing additive genetic (A), common environmental (C) and

unique environmental (E) influences are considered (Boomsma et al., 2002; Posthuma et al., 2000; Neale and Cardon, 1992). Heritability is the standardized additive genetic component: $h^2 = \frac{A}{V} = \frac{A}{A+C+E}$. The unique environmental influences are confounded by measurement error (M) that can be separated from the "true" unique environmental (E') influences by the measurement model: $E = E' + M$. The variance of the reliable trait (i.e. the reliable component in the measurement model) becomes $V' = V - M$ or $V' = A + C + E'$. Heritability of the reliable trait is estimated as the standardized additive genetic component after excluding measurement error: $h^2 = \frac{A}{V'} = \frac{A}{A+C+E'}$. The heritability of the reliable component of functional connectivity was estimated for a measurement model on the half-score measures of functional connectivity (**Supplementary Figure S3**). Heritability estimates of the reliable component of functional connectivity are compared to heritability estimates from the full (i.e. uncorrected) measure of functional connectivity (**Supplementary Figure S3**). Full details on the heritability analysis are provided in the Supplementary Materials.

2.3.3. Evaluation of the measurement model at different samples sizes and scan durations

We performed a parameter sweep to empirically determine the efficiency of the measurement model on improving the reliability of functional connectivity and the strength of the association between traits and functional connectivity for various *sample sizes* and *total scan duration* (**Table 2**). We repeated the analysis for four sessions across two different days, two sessions on the same day, and single session scan data.

Each combination of *total scan duration* and *sample size* was sampled 100 times per functional connection, with a random set of the desired number of participants drawn from the full sample of participants at each iteration. To reduce the computational burden of evaluating the performance of the measurement model for all connections, the same 25 connections identified through the sparse sampling scheme previously described were used for each of the seven exemplar traits.

BOLD time series for the desired scan durations were extracted from the original full-length BOLD time series data by distributing four time blocks of equal length across all four scan sessions, with each block starting at the first volume of each scan session. The blocks were then concatenated to construct the time series of the full-, and half–score measures of functional connectivity (**Supplementary Figure S4**). This distributed approach was adopted for multi-session scan data to prevent half-score measures from crossing scan boundaries; e.g. for a total time series length of 1600 vol (20 min), the second half-score measure would otherwise have been computed across data from both scan #1 (volumes 801:1200) and #2 (volumes 1:400), which has a detrimental effect on the reliability of the second half-score. This distributed approach was also used when the full time series could have fit in the length of a sin-
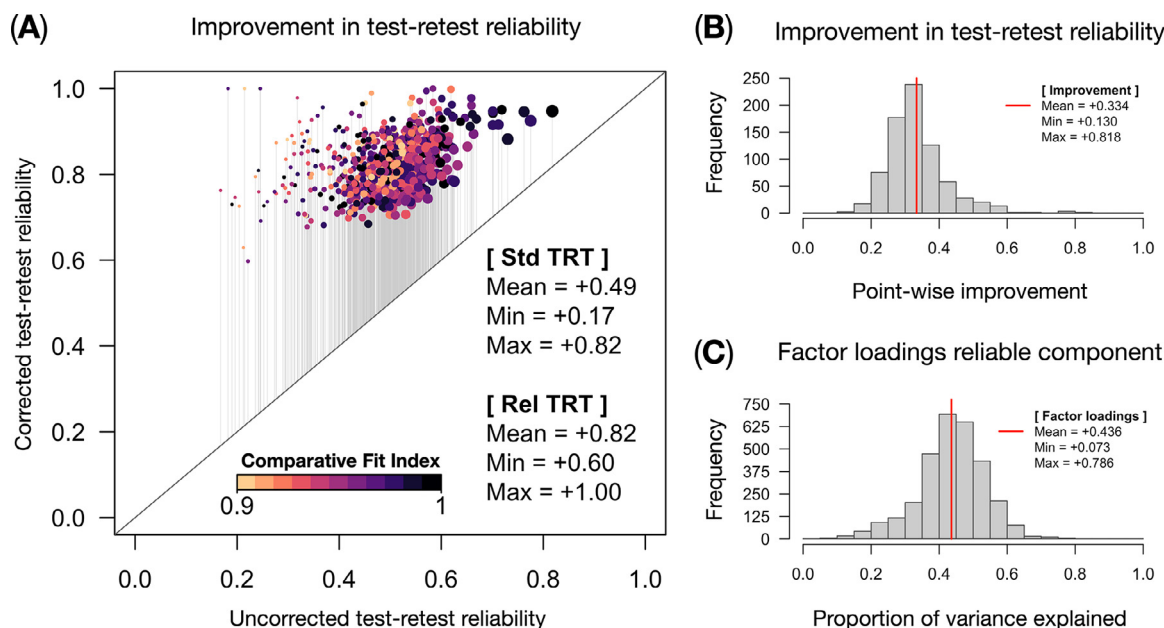
Fig. 2. Test-retest reliability of functional connectivity estimates. (**A**) Improvement in test-retest reliability between the standard association model (*x*-axis) and the measurement model (*y*-axis); data points are scaled by the average proportion of variance explained by the reliable component, thereby emphasizing the more reliable and stable connections, and color-coded by the Comparative Fit Index. (**B**) Point-wise improvement in test-retest reliability between the standard association model and the measurement model. (**C**) Proportion of variance of the quarter-score measures explained by the reliable components of functional connectivity. For all panels, *uncorrected* test-retest reliability was estimated as the association between half-score measures of functional connectivity. The *corrected* test-retest reliability from the measurement model was estimated as the association between the two reliable components of functional connectivity based on the quarter-score measures of functional connectivity. The red lines indicate the mean of the distributions.

gle session (i.e. total scan duration <15 min) to ensure time series are based on the same data. For single session data, contiguous blocks were selected (**Supplementary Figure S4**).

2.5. Minimal requirements on the input dataset

The requirements, in terms of *sample size* and *total scan duration*, for a dataset suitable for reliability modelling was evaluated empirically to determine the threshold where the goodness of fit indices for the measurement model started to deteriorate. Since no universal absolute threshold exists, the proportion of bad fits over the hundred iterations for each combination of *sample size* and *total scan duration* is provided with a lower and upper boundary marked at 25% and 50% quantiles. The same goodness of fit indices Comparative Fit Index (CFI) and root-mean-squared-error of approximation (RMSEA) and their judgement criteria were used as described before.

3. Results

3.1. High consistency of group-level mean functional connectivity

Group-level mean functional connectivity was highly consistent across the two half-score measures of functional connectivity (*rho* = +0.996; intraclass correlation ICC = 0.995, see supplementary information for details), with functional connectivity estimates ranging from –0.52 to +0.66 (mean FC = 0.003) (**Supplementary Figure S5**), and absolute differences between the two half-score measure of less than 0.06 for individual connections. However, there is high variation within individuals.

3.2. Reliability modelling improves test-retest reliability between scan sessions

At the level of individual connection, test-retest reliability of functional connectivity between scan sessions acquired on different days improved substantially for 760 connections (62% of all 1225 connec-

tions) with an acceptable or good fit of the measurement model (**Fig. 2**; **Supplementary Figure S6**). On average, the uncorrected test-retest reliability estimate of functional connectivity was +0.49 (range = +0.17 to +0.82; **Fig. 2A**) and the test-retest reliability estimates of the reliable component of functional connectivity was +0.83 (range = +0.60 to +1.00; **Fig. 2A**). The test-retest reliability estimates improved on average with +0.33 points (range = +0.13 to +0.82; **Fig. 2B**). Connections with lower uncorrected test-retest reliability improved more than connections with already high test-retest reliability due to the ceiling effect of the upper bound of +1.00 on test-retest reliability estimates (**Fig. 2A**). On average, 44% of the variance of the quarter-score measures of functional connectivity was explained by the reliable components of functional connectivity (range = 7% to 79%; **Fig. 2C**).

3.3. Reliability modelling improves the association between functional connectivity and traits

The improvement in the strength of the associations between functional connectivity and the seven extensively tested traits ranged from +17.0% to +23.7% (**Fig. 3A**; **Supplementary Figure S7**). On average, the improvement in association strength for all 110 traits was +20% (range = +12 to +30%; **Fig. 3B**; **Supplementary Data File F3**) using a sparse sampling scheme to approximate the improvement factor when the full set of connections would have been used (**Supplementary Table S2**). On average, 1035 connections (84% of all connections) passed the goodness of fit criteria, with highly similar distributions for the seven traits (**Supplementary Figure S6**). On average, 50% of the variance of the half-score measures of functional connectivity is explained by the reliable component, with a range from 17% to 81%.

3.4. Reliability modelling increases heritability estimates of functional connectivity

The heritability estimates of full-score functional connectivity were on average 39% (range = 0% to 75%; **Supplementary Figure S8**). The
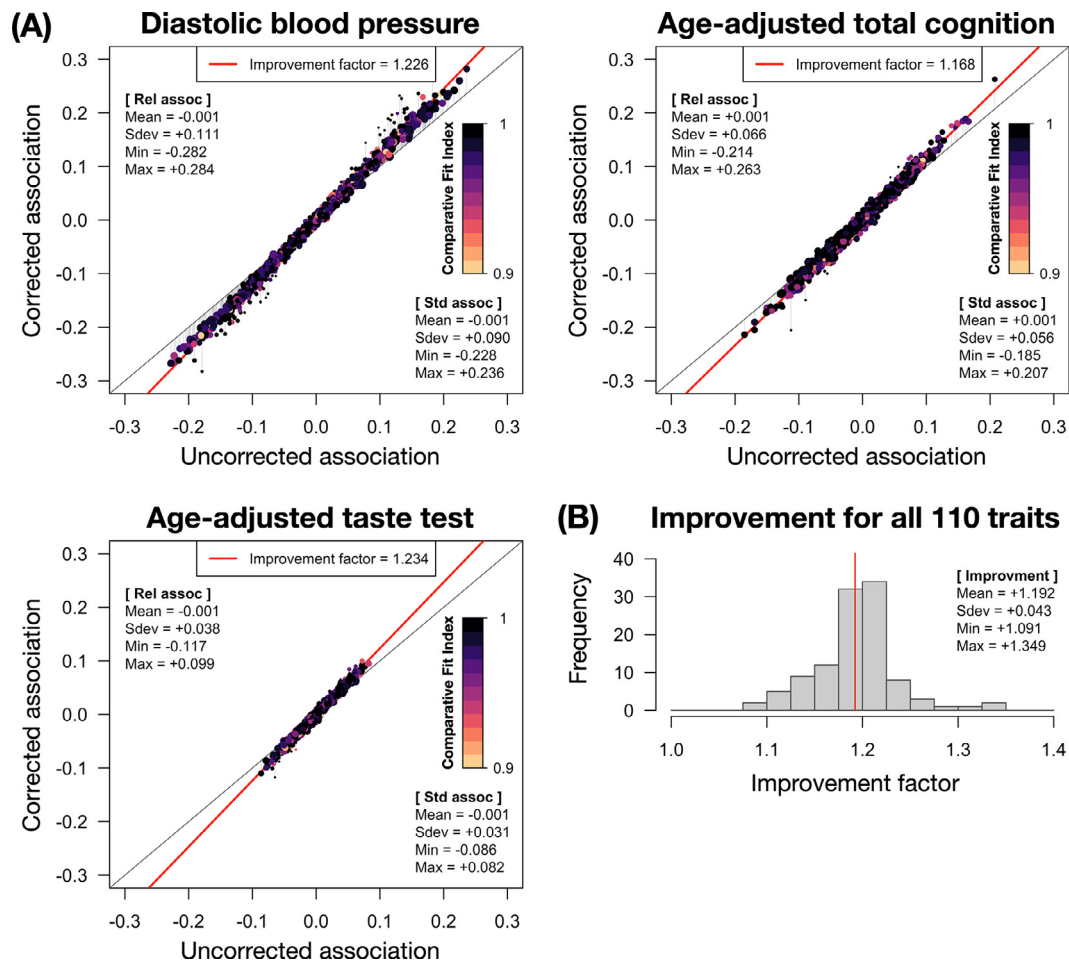
Fig. 3. Association between functional connectivity and traits. (**A**) Improvement in association strength between functional connectivity and the traits diastolic blood pressure level (*BPDiastolic*), total composite score for cognition adjusted for age (*CogTotalComp_AgeAdj*), and taste intensity test adjusted for age (*Taste_AgeAdj*); with association strength from the uncorrected association model on the *x-axis* and the corrected association strength from the measurement model on the *y-axis*. The remaining four extensively tested traits show comparable results and are presented in the Supplement Materials (**Supplementary Figure S7**). The red line represents the slope coefficient of the regression (i.e. the improvement factor). (**B**) Improvement factor for all 110 continuous and normally distributed traits (**Supplementary Data File F3**). The red line represents the mean of the distribution. For all panels, a standard association model was used to estimate the *uncorrected* association between the full-score measure of functional connectivity and the trait. A measurement model applied to the half-score measures of functional connectivity was used to estimate the *corrected* association between the reliable component of functional connectivity and the trait.

heritability estimates of the reliable component of functional connectivity were on average 59% (range = 0% to 93%; **Supplementary Figure S8**). On average, the heritability estimates increased with +20% points (range = −3% to +54% points; **Supplementary Figure S8**).

3.5. Efficiency and minimal requirements for using a measurement model

Test-retest reliability of functional connectivity depends on scan duration and the number of scan sessions (**Fig. 4**). With the standard association model, the average *uncorrected* test-retest reliability across connections in the brain for a single session out-performed the *uncorrected* test-retest reliability of multi-session scan data (dashed lines; **Fig. 4**). However, given the same number of scan sessions, the *corrected* test-retest reliability from the measurement model (solid lines; **Fig. 4**) exceeded the *uncorrected* test-retest reliability for all variants and across all scan durations (dashed lines; **Fig. 4**). Of special note is the dataset variant with two scan sessions. The current measurement model (**Fig. 1B**) estimates the corrected test-retest reliability based on the reliable components of functional connectivity that are corrected for intra-session variation with each session only. Despite improvement in the test-retest reliability, there is still substantial variation between the two sessions. If the quarter scores Q2 and Q4 are interchanged, such that the reliable

components of functional connectivity now account for inter-session variation instead of intra-session variation, the corrected test-retest reliability approaches near perfect scores. Similar results were observed for single-session and four-session scan data after interchanging quarter scores Q2 and Q4. The sample size had little effect on the average test-retest reliability apart from increased variation due to sampling bias for the smaller sample sizes (**Supplementary Figure S9**). Note that for improbable connections (e.g. connection 47–49) model fits started to degrade (i.e. increasing number of bad fits out of the 100 random samples drawn).

All seven traits exhibit the same general pattern of improvement in the association strength, with greater benefit from reliability modelling for shorter scan durations and a slight increase for the smaller sample sizes up to 1.8-fold increase averaged over the 100 iterations per combination of total scan duration and sample size (**Fig. 5**; **Supplementary Figure S10**). Note that for the combinations of very low sample size (25 subjects) and short scan durations (≤1600 vol, or ≤20 min), model fits started to degrade (i.e. increasing number of bad fits out of the 100 random samples drawn for a specific combination of sample size and total scan duration for each of the seven traits). Results for two scan sessions on the same day are nearly identical (**Supplementary Figure S11**). However, the improvement in association strengths are slightly di-
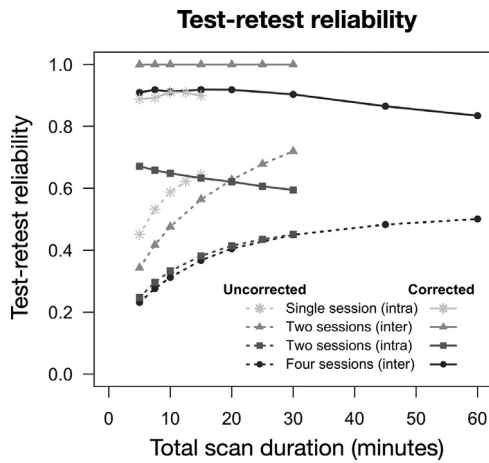
Fig. 4. Test-retest reliability depends on scan duration and number of sessions. The average uncorrected test-retest reliability across all connections (*y*-axis) was estimated for the full-score estimate of functional connectivity for four variants of the full dataset (symbol- and colour-coded curves): a single scan session where the measurement model accounts for intra-session variation between half scores of the same session (*), two scan sessions on the same day accounting for inter-session variation between half-scores of different sessions (▲), two scan sessions on the same day accounting for intra-session variation between half-scores of the same session (■), and four scan sessions across two days accounting for inter-session variation between half-scores of sessions on the same day (●). The total scan duration was varied from 5 min up to the maximum allowed by the dataset of 15 min per available scan session.

vealed a consistent pattern for all seven measures; for brevity we present the results for the total composite score on cognition adjusted for age trait (*CogTotalComp_AgeAdj*) (**Supplementary Figure S12**). Keeping the total scan duration the same for a single session and two sessions on the same day, both datasets produce similar uncorrected associations (**Supplementary Figure S12A**). However, the multi-session variant produces stronger corrected associations after applications of a measurement model (**Supplementary Figure S12A**). There is no additional benefit to using four sessions across two days compared to two sessions on the same day (**Supplementary Figure S12B**). However, it should be noted that the current model only accounts for measurement error between the two half-scores (i.e. Day 1 versus Day 2); more elaborate models that account for both inter- and intra-session measurement error might still benefit from multi-session data across different days (Brandmaier et al., 2018). Similarly, there is no benefit to splitting single session data into more than two half-score measures of functional connectivity (**Supplementary Figure S12C**).

The goodness of fit assessment from the parameter sweep was used to evaluate the minimal requirements on the input dataset in terms of *sample size* and *total scan duration* for reliability modelling. Although no clear boundary can be defined when a measurement model is no longer suitable or practical to use, the chance that the measurement model does not describe the data well for a random sample of participants starts to increase with lower sample size or shorter scan durations (Fig. 6), with similar profiles for all seven measures (**Supplementary Figure S13**).

4. Discussion

We have shown that reliability modelling of functional connectivity using a measurement model on split-session half-score estimates of functional connectivity is able to extract a reliable component of functional connectivity with improved test-retest reliability between scan sessions acquired on separate days. Secondly, we found that the reliable component of functional connectivity is more strongly associated with traits

minished for single session scan data; up to 1.5-fold increase for smaller sample sizes and short scan durations (**Supplementary Figure S11**).
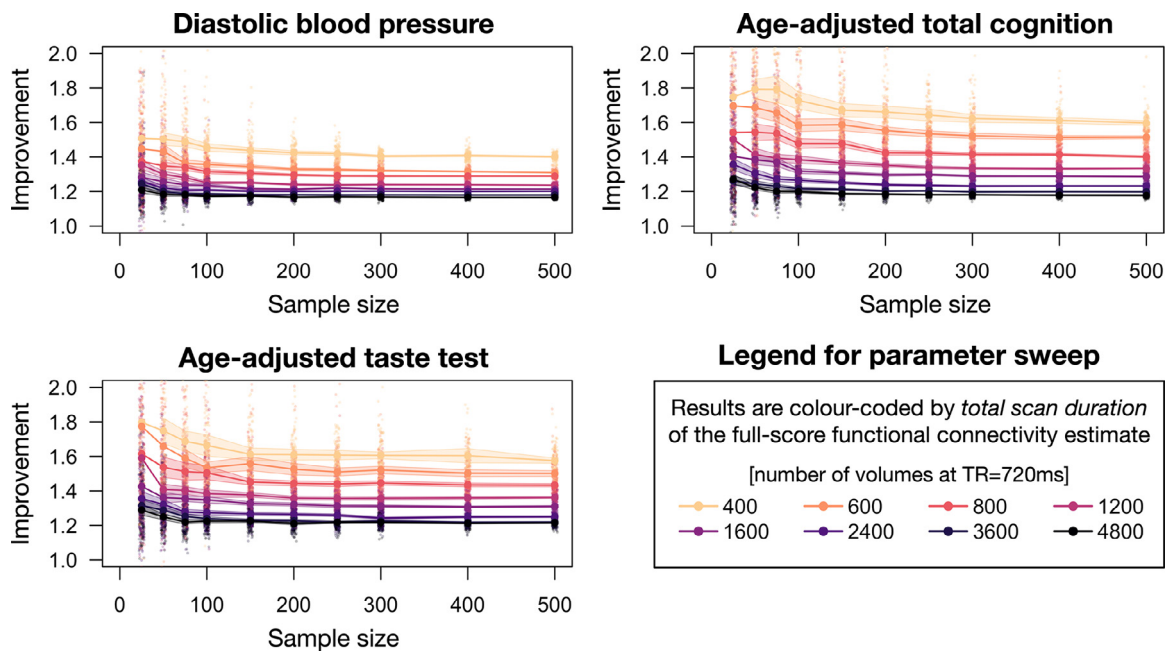
Comparing the different variants of the dataset (i.e. single session, two sessions on the same day, and four sessions across two days) re-

Fig. 5. Improvement in association strength between functional connectivity and the traits diastolic blood pressure levels (*BPDiastolic*), age-adjusted total cognitive component (*CogCompTotal_AgeAdj*), and age-adjusted taste test score (*Taste_AgeAdj*) for various *sample sizes* (*x*-axis) and *total scan duration* (color-coding). Improvement factor (*y*-axis) is defined by the slope coefficient from the linear regression of the corrected association strength onto the uncorrected association strength over all sparsely sampled connections. color-shaded bands represent the 95% confidence interval of the means. For all panels, a standard association model was used to estimate the *uncorrected* association between the full-score measure of functional connectivity and the traits. A measurement model applied to the half-score measures of functional connectivity was used to estimate the *corrected* association between the reliable component of functional connectivity and the traits. The remaining four extensively tested measures showed similar patterns (**Supplementary Figure S10**).

Grand average percentage of bad fits
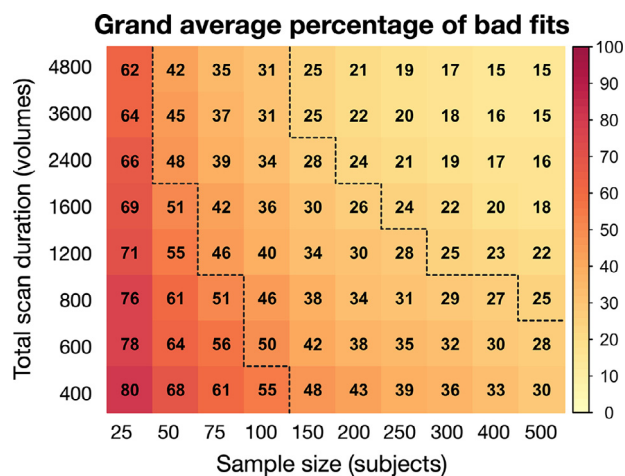
Fig. 6. Percentage of sampled connections for each combination of *sample size* and *total scan duration* for which the goodness of fit for the behavioral association measurement models deteriorated below acceptable levels (CFI < 0.90 or RMSEA > 0.08), averaged across all seven measures (see **Supplementary Figure S13** for the profiles of the individual measures). Total scan duration is reported in minutes. Dotted lines mark the boundary where on average more than 25% and 50% of the model fits are considered bad.

than the full-score estimate of functional connectivity. Finally, we have empirically evaluated the minimal requirements of the dataset for reliability modelling of functional connectivity in terms of scan duration and sample size. We have previously reported increased heritability estimates for the stable and reliable component of functional connectivity in a longitudinal adolescent twin cohort (Teeuw et al., 2019). Here, we have replicated this finding in the Human Connectome Project.

4.1. Reliability modelling is able to extract a stable and more reliable component of functional connectivity

The moderate uncorrected test-retest reliability of functional connectivity (average *rho* = +0.49) that we found is comparable to other studies using the HCP Young Adult dataset (Shah et al., 2016; Ge et al., 2017; S. Noble et al., 2017a; Mejia et al., 2018; Elliot et al., 2019). Many factors influence test-retest reliability of functional connectivity, and as such, the test-retest reliability can vary substantially between datasets, ranging from poor (mean ICC ~ 0.15) to moderate (mean ICC ~ 0.65) (Noble et al., 2019; S. Noble et al., 2017b; Andoh et al., 2017). Despite the high quality of the HPC Young Adult dataset, we found a substantial improvement in test-retest reliability using a measurement model (average increase = +0.33), in some cases resulting in good to excellent scores (average *rho* = +0.82). Longer scan duration has a positive effect on the uncorrected test-retest reliability, as has previously been reported (Birn et al., 2013; Laumann et al., 2015; S. Noble et al., 2017a; Meija et al., 2018; Elliot et al., 2019). However, there is a penalty to inter-session test-retest reliability for multi-session scan data that should be taken into consideration when setting up a new study design. Other methods that improve test-retest reliability, such as disattenuation (i.e. scaling a measure by its reliability to obtain a true estimate), shrinkage (i.e. gravitating unreliable measures towards a group-mean estimate), or combining multiple modalities (e.g. resting-state and task-based functional connectivity), are able to increase the reliability of functional connectivity from +25% upto two-fold improvement (Mueller et al., 2015; Shou et al., 2014; Mejia et al., 2018; Elliot et al., 2019). Our results show that the split-session measurement model is able to extract a component of functional connectivity separately from independent scan sessions acquired on separate days that is more reliable than the individual half-score estimate of functional connectivity. This component represents a stable and reliable component of functional connectivity

for a single link connection that explains on average about half (44%) of the variance in the split-session measurements. The remaining half is attributed to random variation between split-session measurements and is considered "measurement error". In a 7-Tesla resting-state functional MRI study, a similar proportion (50%) of the variance could be attributed to spontaneous neural activity, and the other half to non-thermal physiological noise (Bianciardi et al., 2009). Functional connectivity shows a strong state-like nature where it is influenced by intrinsic and extrinsic factors such as caffeine consumption, heart rate variability, circadian rhythm, daily mood, or attention (Wu et al., 2014; Choe et al., 2015; Hodkinson et al., 2014; Facer-Childs et al., 2019; Ismaylova et al., 2018; Geerligs et al., 2017). These short-term fluctuations in connectivity strength have been the topic of investigation for dynamic functional connectivity (Chang and Glover, 2010; Handwerker et al., 2012; Hutchison et al., 2013; Abrol et al., 2017). The residual variance that is specific to the individual half-score estimate of functional connectivity is considered "measurement error" in the measurement model, but is likely to represent both different sources of measurement error (Brandmaier et al., 2018) and relevant biological transients (Ge et al., 2017). For example, the increased variation between sessions (either same day or several days apart) that resulted in lower test-retest reliability will be treated as measurement-specific variation by the measurement model, but may reflect the "mental state" of the participants at the time of the scan. Instead, the stable component of functional connectivity would be of particular interest for researchers in the pursuit of a reliable and state-independent (or "trait-like") endophenotype that could serve as a biomarker (Beauchaine and Constantino, 2017).

4.2. Reliability modelling improves association of functional connectivity with various traits by revealing the true association in absence of measurement error

The stable and reliable component of functional connectivity obtained by the measurement model on half-score estimates of functional connectivity is more strongly associated with all 110 traits compared to the association between full-score functional connectivity and the traits (average improvement factor 1.2) in the full-sized dataset of the Human Connectome Project (*N* = 1003 participants; ~1 hour of resting-state fMRI data). Overall, measures were only weakly associated with full-score measure of functional connectivity (maximum absolute *rho* < 0.25) when measured by the connection with the strongest association. These low associations are typical for resting-state functional connectivity (Vaidya and Gordon, 2013; Kruschwitz et al., 2018; Geerligs et al., 2017; S. Noble et al., 2017a; Basten et al., 2015; Toschi et al., 2018; Siegel et al., 2017). Even at connectome level, functional connectivity is only moderately associated with behavioural measures (Smith et al., 2015; Finn et al., 2015; Rosenberg et al., 2016). It was recently shown that individual variation in the spatial distribution of functional brain network organization are stable trait-like features that may be associated with behaviour (Seitzman et al., 2019; Kong et al., 2018), and that general cognitive ability is associated with the stability of the dynamic functional connectome (Hilger et al., 2020). This could indicate that although the purpose of a measurement model is to provide a more reliable measure from parallel test scores, it could be the *temporally stable* (i.e. "trait-like") component extracted from the split-session half-score measures of functional connectivity that provide the improved associations with behavioural measures. The short-term stability of within-session functional connectivity likely reflects the "mental state" of the participant at time of the scan. Changes in the mental state of the participant between scan sessions contributes to the inter-session variations and negatively impacts the test-retest reliability. Using multi-session scan data will allow the measurement model to account for these session-specific contributions of the mental states and extract the stable "trait-like" component. Splitting single session scan data into more sections (e.g. using quarter-scores instead of half-scores to represent the latent component of functional connectivity) revealed almost identical,

but with overall slightly lower performance in improvement of the association strength between the reliable component of functional connectivity and behaviour for the model based on the quarter-scores compared to the model based on half-scores. In addition, the increased complexity of the quarter-score model and the associated increase in computational runtime suggest there is little benefit to gain by splitting the scan data of a single session into more than two half-scores. Splitting multi-session scan data into more than two half-score may be beneficial in combination with a model that accounts for both intra- and inter-session variation (Brandmaier et al., 2018). It must be noted however, that in some cases applying the measurement model decreases (and should decrease) the observed association. For example, when a true association is absent but may arise by chance for the full-score measure, when the connection studied is not between two functionally connected regions, or when the connection is unstable or highly dynamic, we expect the corrected association to go down.

4.3. Studies with short scan duration and small samples size will experience greater benefits from reliability modelling

The parameter sweep reveals that the improvement in association strength is dependant on scan duration and, to lesser extent, on sample size. Datasets with a shorter total scan duration (down to five minutes of resting-state fMRI data) show much greater benefit from reliability modelling (on average up to 1.8-fold increase in association strength, with a similar profile for all seven extensively tested traits). A similar pattern was observed for variants of the dataset with a single session or two sessions on the same day, but with diminished results for single session scan data. The diminished results are most likely due to the higher test-retest reliability of single-session scan data. Multi-session scan data provided better results for the corrected associations after application of a measurement model, but no additional benefit was seen for four sessions across different days. However, it should be noted that the current model only accounts for measurement error between the two half-scores (i.e. Day 1 versus Day 2); more elaborate models that account for both inter- and intra-session measurement error might still benefit from multi-session data across different days (Brandmaier et al., 2018). Previously, scan duration has been reported to influence reliability and reproducibility of functional connectivity estimates (Birn et al., 2013; Laumann et al., 2015; S. Noble et al., 2017a; Meija et al., 2018; Elliot et al., 2019) but estimates on the recommended scan duration for maximal reliability vary from 5 to 90 min. With typical scan duration for resting-state fMRI studies anywhere between the minimum recommended 5 min (Birn et al., 2013) to 8 min (Waheed et al., 2016), equivalent to about 400 to 667 vol in the Human Connectome Project dataset, suggests that most dataset can expect a decent boost in association strength with reliability modelling, despite the fact that the corrected associations remain modest. Datasets with smaller sample sizes show a slightly greater benefit of reliability modelling. However, smaller sample sizes are accompanied by increased variation in the improvement factor that is due to sampling bias, which a measurement model cannot account for. It is therefore possible that the larger improvements we observe for smaller sample sizes are directly caused by the fact that the initial estimates of the uncorrected association were worse. For datasets with small sample sizes ($\leq$50 participants), the utility of the measurement model starts to drop, with 50% to 80% of the sampled connections resulting in a bad fit. Note that there is a baseline rejection rate of on average 15% of the connections with a bad fit of the measurement model in the full-sized dataset. This baseline is also present in the parameter sweep because we did not exclude connections with a bad fit prior to selecting the connections for sparse sampling. Previous studies have examined the sample size required for structural equation modelling, with initial estimates suggested that at least 200 participants are needed (Boomsma, 1985), or as a rule of thumb ten times the number of estimated parameters (Bentler and Chou, 1987; Wolf et al., 2013). Sample sizes as low as 50 participants might be enough to obtain satis-

factory fits for task-based fMRI (Sideridis et al., 2014). Our parameter sweep shows that variation due to sampling bias goes down with a sample size around 100 to 150 participants, suggesting that, combined with a typical scan length of around 8 min, reliability modelling would be feasible for most contemporary resting-state fMRI studies.

4.4. Stronger genetic signal for the reliable component of functional connectivity

Measurement models have previously been used in the context of twin studies to obtain heritability estimates for the reliable portion of the variation (van Baal et al., 1998; van Beijsterveldt et al., 2001; Ge et al., 2017). We have previously applied a measurement model to functional connectivity in a longitudinal adolescent twin cohort (Teeuw et al., 2019). Here, we replicated our earlier finding that heritability estimates of functional connectivity can be increased substantially (from average $h^2 = 39\%$ to $h^2 = 59\%$) by using a measurement model on data from split-half scan sessions. Previous studies on the heritability of functional connectivity using the Human Connectome Project dataset have typically reported low heritability for single link connections (Ge et al., 2017; Colclough et al., 2017; Adhikari et al., 2018). One earlier study applied a custom linear mixed effects model to repeated measures of functional connectivity from the Human Connectome Project dataset (i.e. considered scans on Day 1 and Day 2 as repeated measures, similar to the two half-score measures for the full-sized dataset used in this study) and reported similar improvements for heritability estimates of around +20% points when averaged across connections for the major functional networks (Ge et al., 2017).

4.5. Methodological considerations for the application of a measurement model

There are a few methodological considerations of measurement models in general that should be mentioned (Muchinsky, 1996). First, it is important to note that the ultimate purpose of the measurement model is to obtain estimates that are closer to the true value in the absence of error in the measurements. The measurement model will provide more accurate estimates that can guide future studies, but will not change the quality of the data. Secondly, while we discussed our results in terms of improvement factors, correlations obtained with the measurement model are associations between the stable components (i.e. reliable variation) rather than between the full traits (i.e. full variation) and those two are not directly comparable. In addition, there is no ground-truth available for resting-state functional connectivity of the human brain that could verify the correctness of the measurement model outputs. We would like to emphasise again that an improvement of association is not informative for the existence of "true" correlated neuronal activity. It is therefore good practice to always report the uncorrected results from uncorrected association in addition to the corrected results from the measurement model, and important to assess the goodness of fit of the measurement model.

Given our finding that multi-session scan data seems to outperform a single scan session of the same length, we would recommend researchers to adopt an acquisition protocol that splits their fMRI resting-state scans into two parts; for example, 5 min at the beginning and 5 min at the end of a scan session or day, instead of a single 10 min scan. This approach would maximize the (unwanted) variation in measurement error and mental state of the participant between half scores to better capture the stable core (or trait-like) component of functional connectivity with a measurement model. For the researcher that has single session resting-state fMRI data already acquired, we would recommend splitting the single session into two parts. Splitting single-session scans into more than two parts has shown not to improve the results in our data. We have put a practical example and a step-by-step "walkthrough" of the model on GitHub (Teeuw, 2020). The output of the measurement model and its goodness of fit should be interpreted within the context of what

the latent component of functional connectivity represents; specifically, how much variation is explained by the latent component, as determined by the factor loadings of the measurement model, and is roughly proportional to the test-retest reliability between the half-score measures of functional connectivity. Models with a low proportion of variance explained are likely nonsense connections or not functionally connected regions, or highly unstable or dynamic connections, and may result in a poor goodness of fit. For practical studies, we would recommend targeting specific connections that are known to be useful or relevant to the trait under investigation and that the connection is between functionally connected regions based on prior knowledge. Additionally, extending the model may improve the fit to the data; e.g. including correlated error covariance structure may resolve the bad fit (e.g. Brandmaier et al., 2018).

4.6. Limitations to the current study

There are some limitations specific to the current study. First, there is quite some variation in quality between resting-state fMRI datasets (Noble et al., 2019). Other datasets might see a shift in the performance and requirements curves based on the quality of the dataset. Secondly, reliability modelling was not applied to the traits and the reported corrected association might still be limited by the reliability of the trait. If multiple or repeated measures of the trait or measurements from different modalities are available, a measurement model can be applied to both the brain measure (e.g. functional connectivity) and the trait to obtain a more accurate estimate of the association between the two in the absence of measurement error (Beaty et al., 2015; Cooper et al., 2019; Köhncke et al., 2020).

4.7. Conclusion

In conclusion, reliability modelling of functional connectivity using a measurement model on split-half session resting-state fMRI data is an effective method to compensate for attenuation of the temporal correlation coefficient due to noise in the BOLD signal. The measurement model is able to extract a stable and reliable component of functional connectivity that can reveal the true associations with traits and increased heritability estimates compared to the analysis with full-score estimate of functional connectivity. The benefit of a measurement model is greater for studies with short scan duration or a limited number of participants.

Declaration of Competing Interest

The authors declare no conflicting interests.

Data availability

The original data used in this analysis is publicly available after registration at https://db.humanconnectome.org. The definitions of the OpenMx models are available on https://github.com/jalmar/openmx-models. Requests for access to the scripts used in this analysis should be directed to the corresponding author.

Funding

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.117842.

References

Abrol, A., Damaraju, E., Miller, R., Stephen, J., Claus, E., Mayer, A., Calhoun, V., 2017. Replicability of time-varying connectivity patterns in large resting state fMRI samples. Neuroimage 163, 160–176. doi:10.1016/j.neuroimage.2017.09.020.

Andoh, J., Ferreira, M., Leppert, I.R., Matsushita, R., Pike, B., Zatorre, R.J., 2017. How restful is it with all that noise? Comparison of Interleaved silent steady state (ISSS) and conventional imaging in resting-state fMRI. Neuroimage 147, 726–735. doi:10.1016/j.neuroimage.2016.11.065.

Adhikari, B., Jahanshad, N., Shukla, D., Glahn, D., Blangero, J., Fox, P., Reynolds, R., Cox, R., Fieremans, E., Veraart, J., Novikov, D., Nichols, T., Hong, L., Thompson, P., Kochunov, P., 2018. Comparison of heritability estimates on resting state fMRI connectivity phenotypes using the ENIGMA analysis pipeline. Hum. Brain Mapp. 39 (12), 4893–4902. doi:10.1002/hbm.24331.

Basten, U., Hilger, K., Fiebach, C., 2015. Where smart brains are different: a quantitative meta-analysis of functional and structural brain imaging studies on intelligence. Intelligence 51, 10–27. doi:10.1016/j.intell.2015.04.009.

Beaty, Roger E., Kaufman, Scott B., Benedek, Mathias, Jung, Rex E., Kenett, Yoed N., Jauk, Emanuel, Neubauer, Aljoscha C., Silvia, Paul J., 2015. Personality and complex brain networks: The role of openness to experience in default network efficiency. Human Brain Mapping 37 (2), 773–779. doi:10.1002/hbm.23065.

Beauchaine, T., Constantino, J., 2017. Redefining the endophenotype concept to accommodate transdiagnostic vulnerabilities and etiological complexity. Biomark. Med. 11 (9), 769–780. doi:10.2217/bmm-2017-0002.

Bentler, P., Chou, C., 1987. Practical issues in structural modeling. Sociol. Methods Res 16 (1), 78–117. doi:10.1177/0049124187016001004.

Bianciardi, M., Fukunaga, M., van Gelderen, P., Horovitz, S., de Zwart, J., Shmueli, K., Duyn, J., 2009. Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. Magn. Reson. Imaging 27 (8), 1019–1029. doi:10.1016/j.mri.2009.02.004.

Birn, R., Cornejo, M., Molloy, E., Patriat, R., Meier, T., Kirk, G., Nair, V., Meyerand, M., Prabhakaran, V., 2014. The influence of physiological noise correction on test–retest reliability of resting-state functional connectivity. Brain Connect. 4 (7), 511–522. doi:10.1089/brain.2014.0284.

Birn, R., Molloy, E., Patriat, R., Parker, T., Meier, T., Kirk, G., Nair, V., Meyerand, M., Prabhakaran, V., 2013. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. Neuroimage 83, 550–558. doi:10.1016/j.neuroimage.2013.05.099.

Biswal, B., Yetkin, F., Haughton, V., Hyde, J., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. 34 (4), 537–541. doi:10.1002/mrm.1910340409.

Boker, S., et al., 2018. OpenMx R software Package Version 2.11.5. https://openmx.ssri.psu.edu.

Boomsma, A., 1985. Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. Psychometrika 50 (2), 229–242. doi:10.1007/bf02294248.

Boomsma, D., Busjahn, A., Peltonen, L., 2002. Classical twin studies and beyond. Nat. Rev. Genet. 3 (11), 872–882. doi:10.1038/nrg932.

Brandmaier, A., Wenger, E., Bodammer, N., Kühn, S., Raz, N., Lindenberger, U., 2018. Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). Elife 7, e35718. doi:10.7554/elife.35718.

Bright, M., Murphy, K., 2015. Is fMRI "noise" really noise? Resting state nuisance regressors remove variance with network structure. Neuroimage 114, 158–169. doi:10.1016/j.neuroimage.2015.03.070.

Browne, M., Cudeck, R., 1992. Alternative ways of assessing model fit. Sociol. Methods Res. 21 (2), 230–258. doi:10.1177/0049124192021002005.

Burgess, G., Kandala, S., Nolan, D., Laumann, T., Power, J., Adeyemo, B., Harms, M., Petersen, S., Barch, D., 2016. Evaluation of denoising strategies to address motion-correlated artifacts in resting-state functional magnetic resonance imaging data from the human connectome project. Brain Connect. 6 (9), 669–680. doi:10.1089/brain.2016.0435.

Caballero-Gaudes, C., Reynolds, R., 2017. Methods for cleaning the BOLD fMRI signal. Neuroimage 154, 128–149. doi:10.1016/j.neuroimage.2016.12.018.

Chang, C., Glover, G., 2010. Time–frequency dynamics of resting-state brain connectivity measured with fMRI. Neuroimage 50 (1), 81–98. doi:10.1016/j.neuroimage.2009.12.011.

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., Dong, H., Yang, Z., Zang, Y., Zuo, X., Weng, X., 2015. Individual variability and test-retest reliability revealed by

ten repeated resting-state brain scans over one month. PLoS One 10 (12), e0144963. doi:10.1371/journal.pone.0144963.

Choe, A., Jones, C., Joel, S., Muschelli, J., Belegu, V., Caffo, B., Lindquist, M., van Zijl, P., Pekar, J., 2015. Reproducibility and temporal structure in weekly resting-state fMRI over a Period of 3.5 Years. PLoS One 10 (10), e0140134. doi:10.1371/journal.pone.0140134.

Colclough, G., Smith, S., Nichols, T., Winkler, A., Sotiropoulos, S., Glasser, M., van Essen, D., Woolrich, M., 2017. The heritability of multi-modal connectivity in human brain activity. Elife 6, e20178. doi:10.7554/elife.20178.

Cooper, S., Jackson, J., Barch, D., Braver, T., 2019. Neuroimaging of individual differences: a latent variable modeling perspective. Neurosci. Biobehav. Rev. 98, 29–46. doi:10.1016/j.neubiorev.2018.12.022.

Elliott, M., Knodt, A., Cooke, M., Kim, M., Melzer, T., Keenan, R., Ireland, D., Ramrakha, S., Poulton, R., Caspi, A., Moffitt, T., Hariri, A., 2019. General functional connectivity: shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. Neuroimage 189, 516–532. doi:10.1016/j.neuroimage.2019.01.068.

Facer-Childs, E., Campos, B., Middleton, B., Skene, D., Bagshaw, A., 2019. Circadian phenotype impacts the brain's resting state functional connectivity, attentional performance and sleepiness. Sleep 42 (5), zsz033. doi:10.1093/sleep/zsz033.

Finn, E., Shen, X., Scheinost, D., Rosenberg, M., Huang, J., Chun, M., Papademetris, X., Constable, R., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18 (11), 1664–1671. doi:10.1038/nn.4135.

Fox, M., Buckner, R., Liu, H., Chakravarty, M., Lozano, A., Pascual-Leone, A., 2014. Resting-state networks link invasive and noninvasive brain stimulation across diverse psychiatric and neurological diseases. Proc. Natl. Acad. Sci. 111 (41), E4367–E4375. doi:10.1073/pnas.1405003111.

Ge, T., Holmes, A., Buckner, R., Smoller, J., Sabuncu, M., 2017. Heritability analysis with repeat measurements and its application to resting-state functional connectivity. Proc. Natl. Acad. Sci. 114 (21), 5521–5526. doi:10.1073/pnas.1700765114.

Geerligs, L., Tsvetanov, K., Cam-CAN, Henson, R., 2017. Challenges in measuring individual differences in functional connectivity using fMRI: the case of healthy aging. Hum. Brain Mapp. 38 (8), 4125–4156. doi:10.1002/hbm.23653.

Glasser, M., Sotiropoulos, S., Wilson, J., Coalson, T., Fischl, B., Andersson, J., Xu, J., Jbabdi, S., Webster, M., Polimeni, J., van Essen, D., Jenkinson, M., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124. doi:10.1016/j.neuroimage.2013.04.127.

Golestani, A., Goodyear, B., 2011. A resting-state connectivity metric independent of temporal signal-to-noise ratio and signal amplitude. Brain Connect. 1 (2), 159–167. doi:10.1089/brain.2011.0003.

Gottesman, I., Gould, J., 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. Am. J. Psychiatry 160 (4), 636–645. doi:10.1176/appi.ajp.160.4.636.

Greicius, M., Krasnow, B., Reiss, A., Menon, V., 2002. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc. Natl. Acad. Sci. 100 (1), 253–258. doi:10.1073/pnas.0135058100.

Hager, B., Keshavan, M., 2015. Neuroimaging biomarkers for psychosis. Curr. Behav. Neurosci. Rep. 2 (2), 102–111. doi:10.1007/s40473-015-0035-4.

Handwerker, D., Roopchansingh, V., Gonzalez-Castillo, J., Bandettini, P., 2012. Periodic changes in fMRI connectivity. Neuroimage 63 (3), 1712–1719. doi:10.1016/j.neuroimage.2012.06.078.

van Baal, G., de Geus, E., Boomsma, D., 1998. Longitudinal study of genetic influences on ERP-P3 during childhood. Dev. Neuropsychol. 14 (1), 19–45. doi:10.1080/87565649809540699.

van Beijsterveldt, C., van Baal, G., Molenaar, P., Boomsma, D., de Geus, E., 2001. Stability of genetic and environmental influences on P300 amplitude: a longitudinal study in adolescent twins. Behav. Genet. 31 (6), 533–543. doi:10.1023/a:1013389226795.

van den Heuvel, M., Hulshoff Pol, H., 2010. Exploring the brain network: a review on resting-state fMRI functional connectivity. Eur. Neuropsychopharmacol. 20 (8), 519–534. doi:10.1016/j.euroneuro.2010.03.008.

Hilger, K., Fukushima, M., Sporns, O., Fiebach, C., 2020. Temporal stability of functional brain modules associated with human intelligence. Hum. Brain Mapp. 41 (2), 362–372. doi:10.1002/hbm.24807.

Hodkinson, D., O'Daly, O., Zunszain, P., Pariante, C., Lazurenko, V., Zelaya, F., Howard, M., Williams, S., 2014. Circadian and homeostatic modulation of functional connectivity and regional cerebral blood flow in humans under normal entrained conditions. J. Cereb. Blood Flow Metab. 34 (9), 1493–1499. doi:10.1038/jcbfm.2014.109.

Hohenfeld, C., Werner, C., Reetz, K., 2018. Resting-state connectivity in neurodegenerative disorders: is there potential for an imaging biomarker? NeuroImage: Clinical 18, 849–870. doi:10.1016/j.nicl.2018.03.013.

Hu, L., Bentler, P., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model.: Multidiscip. J. 6 (1), 1–55. doi:10.1080/10705519909540118.

Hutchison, R., Womelsdorf, T., Allen, E., Bandettini, P., Calhoun, V., Corbetta, M., Penna, S., Duyn, J., Glover, G., Gonzalez-Castillo, J., Handwerker, D., Keilholz, S., Kiviniemi, V., Leopold, D., Pasquale, F., Sporns, O., Walter, M., Chang, C., 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage 80, 360–378. doi:10.1016/j.neuroimage.2013.05.079.

Ismaylova, E., Sante, J., Gouin, J., Pomares, F., Vitaro, F., Tremblay, R., Booij, L., 2018. Associations between daily mood states and brain gray matter volume, resting-state functional connectivity and task-based activity in healthy adults. Front. Hum. Neurosci. 12, 168. doi:10.3389/fnhum.2018.00168.

Kong, R., Li, J., Orban, C., Sabuncu, M., Liu, H., Schaefer, A., Sun, N., Zuo, X., Holmes, A., Eickhoff, S., Yeo, B., 2018. Spatial topography of individual-specific cortical networks

predicts human cognition, personality, and emotion. Cereb. Cortex 29 (6), 2533–2551. doi:10.1093/cercor/bhy123.

Krishnamurthy, V., Krishnamurthy, L., Schwam, D., Ealey, A., Shin, J., Greenberg, D., Morris, R., 2018. Retrospective correction of physiological noise: impact on sensitivity, specificity, and reproducibility of resting-state functional connectivity in a reading network model. Brain Connect. 8 (2), 94–105. doi:10.1089/brain.2017.0513.

Kruschwitz, J., Waller, L., Daedelow, L., Walter, H., Veer, I., 2018. General, crystallized and fluid intelligence are not associated with functional global network efficiency: a replication study with the human connectome project 1200 data set. Neuroimage 171, 323–331. doi:10.1016/j.neuroimage.2018.01.018.

Köhncke, Y., Düzel, S., Sander, M.C., Lindenberger, U., Kühn, S., Brandmaier, A.M., 2020. Hippocampal and Parahippocampal gray matter structural integrity assessed by multimodal imaging is associated with episodic memory in old age. Cereb. Cortex doi:10.1093/cercor/bhaa287.

Laumann, T., Gordon, E., Adeyemo, B., Snyder, A., Joo, S., Chen, M., Gilmore, A., McDermott, K., Nelson, S., Dosenbach, N., Schlaggar, B., Mumford, J., Poldrack, R., Petersen, S., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87 (3), 657–670. doi:10.1016/j.neuron.2015.06.037.

Lindquist, M., 2020. Neuroimaging results altered by varying analysis pipelines. Nature 582 (7810), 36–37. doi:10.1038/d41586-020-01282-z.

Liu, T., 2016. Noise contributions to the fMRI signal: an overview. Neuroimage 143, 141–151. doi:10.1016/j.neuroimage.2016.09.008.

Logothetis, N., Wandell, B., 2004. Interpreting the BOLD Signal. Ann. Rev. Physiol. 66 (1), 735–769. doi:10.1146/annurev.physiol.66.082602.092845.

Mejia, A., Nebel, M., Barber, A., Choe, A., Pekar, J., Caffo, B., Lindquist, M., 2018. Improved estimation of subject-level functional connectivity with full and partial correlation with empirical Bayes shrinkage. Neuroimage 172, 478–491. doi:10.1016/j.neuroimage.2018.01.029.

Miller, M., 1995. Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modeling. Struct. Eq. Model.: Multidiscip. J. 2 (3), 255–273. doi:10.1080/10705519509540013.

Muchinsky, P., 1996. The correction for attenuation. Educ. Psychol. Meas. 56 (1), 63–75. doi:10.1177/0013164496056001004.

Mueller, S., Wang, D., Fox, M., Pan, R., Lu, J., Li, K., Sun, W., Buckner, R., Liu, H., 2015. Reliability correction for functional connectivity: theory and implementation. Hum. Brain Mapp. 36 (11), 4664–4680. doi:10.1002/hbm.22947.

Murphy, K., Birn, R., Bandettini, P., 2013. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349–359. doi:10.1016/j.neuroimage.2013.04.001.

Neale, M., Hunter, M., Pritikin, J., Zahery, M., Brick, T., Kirkpatrick, R., Estabrook, R., Bates, T., Maes, H., Boker, S., 2016. OpenMx 2.0: extended structural equation and statistical modeling. Psychometrika 81 (2), 535–549. doi:10.1007/s11336-014-9435-8.

Neale, M., Cardon, L., 1992. Methodology For Genetic Studies of Twins and Families. Springer Netherlands, Dordrecht. doi:10.1007/978-94-015-8018-2.

Noble, S., Spann, M., Tokoglu, F., Shen, X., Constable, R., Scheinost, D., 2017a. Influences on the test–retest reliability of functional connectivity MRI and its relationship with behavioral utility. Cereb. Cortex 27 (11), 5415–5429. doi:10.1093/cercor/bhx230.

Noble, S., Scheinost, D., Finn, E.S., Shen, X., Papademetris, X., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D.M., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., Erp, T.G.M.van, Walker, E.F., Hamann, S., Woods, S.W., Cannon, T.D., Constable, R.T., 2017b. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959–970. doi:10.1016/j.neuroimage.2016.10.020.

Noble, S., Scheinost, D., Constable, R., 2019. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. Neuroimage 203, 116157. doi:10.1016/j.neuroimage.2019.116157.

Parkes, L., Fulcher, B., Yücel, M., Fornito, A., 2018. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. Neuroimage 171, 415–436. doi:10.1016/j.neuroimage.2017.12.073.

Posthuma, D., Geus, E., Neale, M., Hulshoff Pol, H., Baaré, W., Kahn, R., Boomsma, D., 2000. Multivariate genetic analysis of brain structure in an extended twin design. Behav. Genet. 30 (4), 311–319. doi:10.1023/a:1026501501434.

Power, J., Barnes, K., Snyder, A., Schlaggar, B., Petersen, S., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59 (3), 2142–2154. doi:10.1016/j.neuroimage.2011.10.018.

Power, J., Schlaggar, B., Petersen, S., 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage 105, 536–551. doi:10.1016/j.neuroimage.2014.10.044.

Rosenberg, M., Finn, E., Scheinost, D., Papademetris, X., Shen, X., Constable, R., Chun, M., 2016. A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. 19 (1), 165–171. doi:10.1038/nn.4179.

Seitzman, B., Gratton, C., Laumann, T., Gordon, E., Adeyemo, B., Dworetsky, A., Kraus, B., Gilmore, A., Berg, J., Ortega, M., Nguyen, A., Greene, D., McDermott, K., Nelson, S., Lessov-Schlaggar, C., Schlaggar, B., Dosenbach, N., Petersen, S., 2019. Trait-like variants in human functional brain networks. Proc. Natl. Acad. Sci. 116 (45), 22851–22861. doi:10.1073/pnas.1902932116.

Shah, L., Cramer, J., Ferguson, M., Birn, R., Anderson, J., 2016. Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. Brain Behav. 6 (5). doi:10.1002/brb3.456.

Shen, X., Cox, S., Adams, M., Howard, D., Lawrie, S., Ritchie, S., Bastin, M., Deary, I., McIntosh, A., Whalley, H., 2018. Resting-state connectivity and its association with cognitive performance, educational attainment, and household income in the UK Biobank. Biol. Psychiatry: Cogn. Neurosci. Neuroimaging 3 (10), 878–886. doi:10.1016/j.bpsc.2018.06.007.

Shou, H., Eloyan, A., Nebel, M., Mejia, A., Pekar, J., Mostofsky, S., Caffo, B., Lindquist, M., Crainiceanu, C., 2014. Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI. Neuroimage 102, 938–944. doi:10.1016/j.neuroimage.2014.05.043.

Sideridis, G., Simos, P., Papanicolaou, A., Fletcher, J., 2014. Using structural equation modeling to assess functional connectivity in the brain. Educ. Psychol. Meas. 74 (5), 733–758. doi:10.1177/0013164414525397.

Siegel, J., Mitra, A., Laumann, T., Seitzman, B., Raichle, M., Corbetta, M., Snyder, A., 2017. Data quality influences observed links between functional connectivity and behavior. Cereb. Cortex 27 (9), 4492–4502. doi:10.1093/cercor/bhw253.

Smith, S., Beckmann, C., Andersson, J., Auerbach, E., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D., Griffanti, L., Harms, M., Kelly, M., Laumann, T., Miller, K., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A., Vu, A., Woolrich, M., Xu, J., Yacoub, E., Uğurbil, K., van Essen, D., Glasser, M., 2013. Resting-state fMRI in the Human Connectome Project. Neuroimage 80, 144–168. doi:10.1016/j.neuroimage.2013.05.039.

Smith, S., Nichols, T., Vidaurre, D., Winkler, A., Behrens, T., Glasser, M., Ugurbil, K., Barch, D., van Essen, D., Miller, K., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat. Neurosci. 18 (11), 1565–1567. doi:10.1038/nn.4125.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72. doi:10.2307/1412159.

Streiner, D., 2003. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J. Pers. Assess. 80 (1), 99–103. doi:10.1207/s15327752jpa8001_18.

R Core Team, 2018. The R Project For Statistical Computing https://www.r-project.org.

Teeuw, J., 2020. A Repository of OpenMx structural Equation models. GitHub, Release 2020-08-25 https://github.com/jalmar/openmx-models doi:10.5281/zenodo.3999406.

Teeuw, J., Brouwer, R., Guimarães, J., Brandner, P., Koenis, M., Swagerman, S., Verwoert, M., Boomsma, D., Hulshoff Pol, H., 2019. Genetic and environmental influences on functional connectivity within and between canonical cortical resting-state networks throughout adolescent development in boys and girls. Neuroimage 202, 116073. doi:10.1016/j.neuroimage.2019.116073.

Toschi, N., Riccelli, R., Indovina, I., Terracciano, A., Passamonti, L., 2018. Functional connectome of the five-factor model of personality. Personal. Neurosci. 1, e2. doi:10.1017/pen.2017.2.

Vaidya, C., Gordon, E., 2013. Phenotypic variability in resting-state functional connectivity: current status. Brain Connect. 3 (2), 99–120. doi:10.1089/brain.2012.0110.

van Essen, D., Smith, S., Barch, D., Behrens, T., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human Connectome project: an overview. Neuroimage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.

Varikuti, D., Hoffstaedter, F., Genon, S., Schwender, H., Reid, A., Eickhoff, S., 2017. Resting-state test–retest reliability of a priori defined canonical networks over different preprocessing steps. Brain Struct. Funct. 222 (3), 1447–1468. doi:10.1007/s00429-016-1286-x.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 4 (3), 274–290. doi:10.1111/j.1745-6924.2009.01125.x.

Waheed, S., Mirbagheri, S., Agarwal, S., Kamali, A., Yahyavi-Firouz-Abadi, N., Chaudhry, A., DiGianvittorio, M., Gujar, S., Pillai, J., Sair, H., 2016. Reporting of resting-state functional magnetic resonance imaging preprocessing methodologies. Brain Connect. 6 (9), 663–668. doi:10.1089/brain.2016.0446.

Wang, L., 2010. Disattenuation of Correlations due to fallible measurement. Newborn Infant Nurs. Reviews 10 (1), 60–65. doi:10.1053/j.nainr.2009.12.013.

Whitfield-Gabrieli, S., Ford, J., 2012. Default mode network activity and connectivity in psychopathology. Ann. Rev. Clin. Psychol. 8 (1), 49–76. doi:10.1146/annurev–clinpsy-032511-143049.

Wolf, E., Harrington, K., Clark, S., Miller, M., 2013. Sample size requirements for structural equation models. Educ. Psychol. Meas. 73 (6), 913–934. doi:10.1177/0013164413495237.

Wu, W., Lien, S., Chang, J., Yang, S., 2014. Caffeine alters resting-state functional connectivity measured by blood oxygenation level-dependent MRI. NMR Biomed. 27 (4), 444–452. doi:10.1002/nbm.3080.

Zhang, D., Raichle, M., 2010. Disease and the brain's dark energy. Nat. Rev. Neurol. 6 (1), 15–28. doi:10.1038/nrneurol.2009.198.