Test-retest and repositioning effects of white matter microstructure measurements in selected white matter tracts

Chaitali Anand [a,1], Andreas M. Brandmaier [b,c,d,1], Jonathan Lynn [e], Muzamil Arshad [f], Jeffrey A. Stanley [g], Naftali Raz [b,e,h,*]

[a] Department of Radiology, University of California, San Francisco, CA, USA
[b] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany
[c] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany
[d] Department of Psychology, MSB Medical School Berlin, Berlin, Germany
[e] Institute of Gerontology, Wayne State University, MI, USA
[f] Department of Cellular and Radiation Oncology, University of Chicago Hospital, Chicago, IL, USA
[g] Department of Psychiatry and Behavioral Neurosciences, Wayne State University, Detroit, MI, USA
[h] Department of Psychology, Wayne State University, Detroit, MI, USA

ARTICLE INFO

ABSTRACT

We used intra-class effect decomposition (ICED) to evaluate the reliability of myelin water fraction (MWF) and geometric mean T_2 relaxation time (geomT$_{2IEW}$) estimated from a multi-echo MRI sequence. Our evaluation addressed test-retest reliability, with and without participant re-positioning, for seven commonly assessed white matter tracts: anterior and posterior limbs of the internal capsule, dorsal and ventral branches of the cingulum, the inferior fronto-occipital fasciculus, the superior longitudinal fasciculus, and the fornix in 20 healthy adults. We acquired two back-to-back scans in a single session, and a third after a break and repositioning the participant in the scanner. For both indices and for all white matter tracts assessed, reliability for an immediate retest, and after the participant's repositioning in the scanner was high. Variance partitioning revealed that in addition to measurement noise, which was significant in all regions, repositioning contributed to unreliability mainly in longer association fibers. Hemispheric location did not significantly contribute to unreliability in any region of interest (ROI). Thus, despite non-negligible error of measurement, for all ROIs, MWF and geomT$_{2IEW}$ have good test–retest reliability, regardless of the hemispheric location and are, therefore, suitable for longitudinal investigations in healthy adults.

1. Introduction

In vivo magnetic resonance imaging (MRI) holds great promise for investigating brain changes over the course of lifespan development and unfolding pathological processes. To meet these expectations, MRI methods must be suitable for longitudinal studies. They must show high test-retest reliability and be insensitive to changes in participants' position in the scanner, which is inevitable in repeated acquisitions. Multi-occasion assessment of brain properties is most important in studies of developmental and pathological changes and of such key properties is the myelin content of the cerebral white matter and its distribution across white matter tracts. Myelination of axons is a key step in brain development (Barkovich et al., 1988), myelin loss and degradation have been proposed as substrates of age-related cognitive decline (Bartzokis, 2004; Peters, 2009), and progressive demyelination is a hallmark of neuropathology associated with multiple sclerosis (Laule et al., 2006; Whittall et al., 2002).

Multi-echo spin–spin relaxation (ME-T_2) imaging, introduced almost three decades ago, remains the gold standard of assessing fundamental properties of the white-matter microstructure - myelin content and axon diameter/packing density *in vivo* (Alonso-Ortiz et al., 2015; MacKay et al., 1994). This approach relies on estimating myelin content via a proxy - myelin water fraction (MWF) - the fraction of the water signal attributed to the short-T_2 relaxation time constant (<40 ms at 3 T) relative to the total observed water signal (MacKay et al., 1994). In

---

* Corresponding author. Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany.
E-mail addresses: chaitali.anand@ucsf.edu (C. Anand), brandmaier@mpib-berlin.mpg.de (A.M. Brandmaier), jonathan.lynn@wayne.edu (J. Lynn), arshadmuzamil92@gmail.com (M. Arshad), jeffrey.stanley@wayne.edu (J.A. Stanley), nraz@wayne.edu (N. Raz).
[1] These authors contributed equally to this work.

Acronyms

MRI	Magnetic Resonance Imaging
ME-T_2	Multi-echo T_2 relaxation imaging
MWF	Myelin Water Fraction
GeomT$_{2IEW}$	Geometric mean of intra-/extracellular water T_2 relaxation time
ROI	Region of Interest
ALIC	Anterior limb of the Internal Capsule (projection fibers)
PLIC	Posterior limb of the Internal Capsule (projection fibers)
DCG	Dorsal Cingulum (part of the cingulum that runs superior to the corpus callosum) (projection fibers)
VCG	Ventral Cingulum (part of cingulum that extends to the temporal lobes) (projection fibers)
IFOF	Inferior Fronto-Occipital fasciculus (association fibers)
SLF	Superior Longitudinal Fasciculus (association fibers)
FNX	Fornix (projection fibers)
ICC	Intra-class Correlation
ICED	Intra-class Effect Decomposition

addition to MWF estimates of myelin content, the ME-T_2 approach provides, by assessing intra- and extra-cellular water relaxation, geometric T_2 (geomT$_{2IEW}$) - an index of axon size and packing density that reflects the increased mobility of water molecules within the intra- and extra-cellular space (T_2 of ~50–80 ms at 3 T) (Arshad et al., 2017). Both indices have been histologically validated for the major white matter tracts (Laule et al., 2006; Webb et al., 2003) and have been used to study cross-sectional lifespan age-related differences in the above-mentioned microstructural properties in white matter tracts (Arshad et al., 2017; Lynn et al., 2021). At the time of this writing, we know of no longitudinal studies of age-related change in MWF and geomT$_{2IEW}$ and therefore, assessment of their test-retest reliability and sensitivity to participant repositioning are very much in need.

Using Intraclass Correlation (ICC), Arshad and colleagues have already established both MWF and geomT$_{2IEW}$ as reliable indices that are relatively insensitive to changes in participants' repositioning in the scanner during repeated testing (Arshad et al., 2017). In that study, however, different sources of unreliability were conflated within a single ICC value. A more nuanced method of reliability evaluation that allowed separation of sources of variance/unreliability attributable to repetition (back-to-back) scans versus participant repositioning between scanning sessions was developed by Brandmaier and colleagues (Brandmaier et al., 2018). In their approach, separation of different sources of variance is attained by applying a structural equation modelling approach that partitioned the observed variability into different latent sources that are identifiable based on the design of a given reliability study. This approach is termed intra-class effect decomposition (ICED) (Brandmaier et al., 2018). To date, the analyses of sources of unreliability using ICED has been applied only to commissural fibers (Anand et al., 2019). Here, using the data collected in the previous study (Arshad et al., 2017), we extend the ICED-based analyses to evaluate the test–retest reliability and the effect of repositioning on MWF and geomT$_{2IEW}$ in the seven representative white matter tracts, in addition to commissural fibers examined by Anand et al. (2019). The selected tracts covered association and projection fiber groups and are frequently used in studies of aging and development (Catani et al., 2002). In addition, instead of averaging the hemispheres as in the Arshad et al. (2017) study, we extended the original ICED model (Brandmaier et al., 2018) to include an additional source of variability – the left and right brain hemispheres, as a within-subject factor. Thus, we assessed reliability of two white matter indices – MWF and geomT$_{2IEW}$, over simple repetition and retest-with-repositioning acquisition as described in detail in the previous publications (Anand et al., 2019; Arshad et al., 2017).

2. Methods

2.1. Participants

The sample consisted of 20 healthy adults (mean age $\pm$ SD = 45.9 $\pm$ 17.1 years, range 24.4–69.5 years; 10 of each sex) from the Metro Detroit area. The participants provided informed consent in accordance with Wayne State University Institutional Review Board guidelines, were screened via a telephone interview, and completed a detailed mail-in health questionnaire. Inclusion and exclusion criteria are available in the previous publications (Anand et al., 2019; Arshad et al., 2017). Briefly, the exclusion criteria were history of a neurological or major psychiatric disorder, cardiovascular disease (other than medically controlled hypertension), cerebrovascular disease, endocrine and metabolic disorders, cancer, or head trauma (with loss of consciousness for more than 5 min). Persons taking anxiolytics, antidepressants, or anti-seizure medication were excluded from the study, and so were pregnant and lactating women. All participants completed a Center for Epidemiologic Studies-Depression Scale (CES-D) to rule out current depression (CES-D; cut-off = 15) (Radloff, 1977). The Mini Mental State Examination (MMSE; cutoff = 26) (Folstein et al., 1975) served as a screening device for cognitive impairment. The participants underwent thorough MRI safety screening to exclude those with any metal in or on their bodies. Participants were requested to avoid intake of caffeine and/or alcohol before the MRI scan. To rule out potential time-of-day effects (Karch et al., 2019), the time of the scan was kept consistent across participants by collecting all data in the morning around 9:00 a. m. In addition, the intervals between the test, retest, and repositioning sessions were kept constant among participants.

2.2. MRI acquisition

The data were collected on a 3T Siemens MAGNETOM Verio MRI system with a 12-channel receive-only volume head coil. The T_1-weighted images were acquired in the axial plane with the following parameters: repetition time (TR) = 2400 ms, echo time (TE) = 2.63 ms, flip angle (FA) = 8°, inversion time (TI) = 1100 ms, matrix size = 256 × 256, number of slices = 160; GRAPPA factor = 2; voxel size = 1.0 × 1.0 × 1.0 mm³. The ME-T_2 images were acquired in the axial plane using the 3D gradient and spin echo (GRASE) sequence: TR = 1100 ms, number of echoes = 32, first echo = 11 ms, inter-echo spacing = 11 ms, FOV = 190 × 220 mm², matrix size = 165 × 192, slice thickness = 5 mm, number of slices = 24, slice oversampling = 0, in-plane resolution = 1.1 × 1.1, acquisition time 16:09 min. These slices were further resampled to a 1 × 1 × 1 mm³ resolution by interpolating each data set to 2.5 mm thickness, and co-registered to the structural T_1-weighted image. The study protocol included three acquisitions in a single visit. In the first scanning session, T_1-weighted MRI images were acquired, followed by two back-to-back scans of ME-T_2 images. At the end of this first session, participants were removed from the scanner, given a 5-min break, and placed back in the scanner for the second scanning session, which included a T_1-weighted and ME-T_2 scan.

Further details about the study design, image acquisition, and pre-processing can be obtained from the previous publication (Arshad et al., 2017). The two outcome measures investigated for the seven white matter regions of interest (ROIs) mentioned above are, MWF expressed as a percent of the total water signal and geomT$_{2IEW}$ expressed in milliseconds (ms).

2.3. Region of interest post-processing

Regions of interest (ROIs) were defined in FSL (Jenkinson et al., 2012) using the following two atlases: the ICBM-DTI-81 white matter atlas and the JHU white matter tractography atlas (Hua et al., 2008; Mori et al., 2005). The ROIs included two left and right projection tracts [anterior (ALIC) and posterior (PLIC) limbs of the internal capsule], two

left and right association tracts [superior longitudinal fasciculus (SLF) and the inferior fronto-occipital fasciculus (IFOF)], two aspects of the cingulum bundle [dorsal (DCG) and ventral (VCG)], and the body of the fornix (FNX).

For each scan, the ROIs in the Montreal Neurological Institute (MNI-152) template space were co-registered to the participant's GRASE images. First, the T_1-weighted images were co-registered to the MNI template space using the FSL non-linear registration tool, FNIRT. Second, the volume of the first echo of the GRASE images was co-registered to the T_1-weighted images using the linear registration tool, FLIRT, with 6 degrees of freedom. Third, the inverse of the warp-field transformation from step one was applied to each ROI (in MNI template space) followed by the application of the inverse transformation from step two. The T_1-weighted images were segmented to generate tissue-probability maps of white matter, gray matter, and cerebrospinal fluid using the FSL segmentation tool FAST (Zhang et al., 2001), which assigns each voxel a probability of belonging to white matter, gray matter, or cerebrospinal fluid. The obtained white-matter probability maps were then co-registered to the same space as the GRASE images. To ensure that each ROI consisted of primarily white matter, the white-matter probability maps were thresholded and binarized to generate a mask reflecting probability values of 95% or greater, which was then multiplied by each ROI for each subject. For the two cingulum ROIs (VCG and DCG), an additional 75% threshold mask was also created. Similarly, for the fornix, additional 75% and 5% masks were created. For the final analysis we used the 95%-thresholded ALIC, PLIC, SLF, IFOF, and DCG; and 75%-thresholded VCG and FNX. The choice of the ROI mask thresholds was based on the size of the individual white matter tracts and the extent to which the chosen threshold proved to be an accurate representation of the white matter tract in question. The ROI masks in subject space were applied to the ME-T_2 data. For each ROI in each subject, a voxel-wise ME-T_2 relaxation analysis was conducted using regularized non-negative least square algorithm (Arshad et al., 2017; Whittall et al., 2002). The seven white matter tracts are depicted in Fig. 1.

2.4. Statistical analyses

Data were assessed for normality using the Shapiro-Wilk test. The analysis of reliability was based on the ICED approach (Brandmaier et al., 2018) that partitions the observed between-person variance into orthogonal error variance components attributable to different measurement characteristics. The variance attributed to the effect of repositioning (session-specific variance, SSV) was separated from the true-score variance (TSV), which represents individual differences in the measure of interest, and is defined as the shared variance over all three scanning sessions. We modeled the hierarchical structure of three acquisitions nested in two sessions to identify the SSV component. The residual error variance (REV) was the third source of variance. Means were estimated as free parameters and ignored in all further analyses, which thus only pertained to the partition of variances and covariances. Model specification and estimation were performed in Ωnyx (ver. 1.0–1026) (von Oertzen et al., 2015). and *lavaan*, an SEM package for R (Rosseel, 2012). The path diagram in Fig. 2 illustrates the ICED model for estimating the individual variance components of the total observed variance in MWF values in the left ALIC. This model estimated the three sources of variability: total, session-specific, and residual (TSV, SSV, REV). The three ME-T_2 scans (scan 1, scan 2, and scan 3) of the MWF for the left ALIC (as an example) are labeled in the path diagram as "ALIC_L.1", "ALIC_L.2", and "ALIC_L.3". The same labeling convention was applied to all regions. In the SEM framework, observed (i.e., measured) variables are depicted as rectangles and latent variables as circles. Single-headed arrows are regressions whereas double-headed arrows are variances. Numbers represent maximum likelihood estimates of parameters. This convention will be followed in all figures below that illustrate the SEM-based model of sources of variance.

An unconstrained SEM model freely estimated all three variance parameters related to the three sources of variance (TSV, SSV, and REV). Three additional null models were created in which each source's (true, session-related, and error) variance was set to zero, one at a time. To assess the significance of the magnitudes of these separate sources of
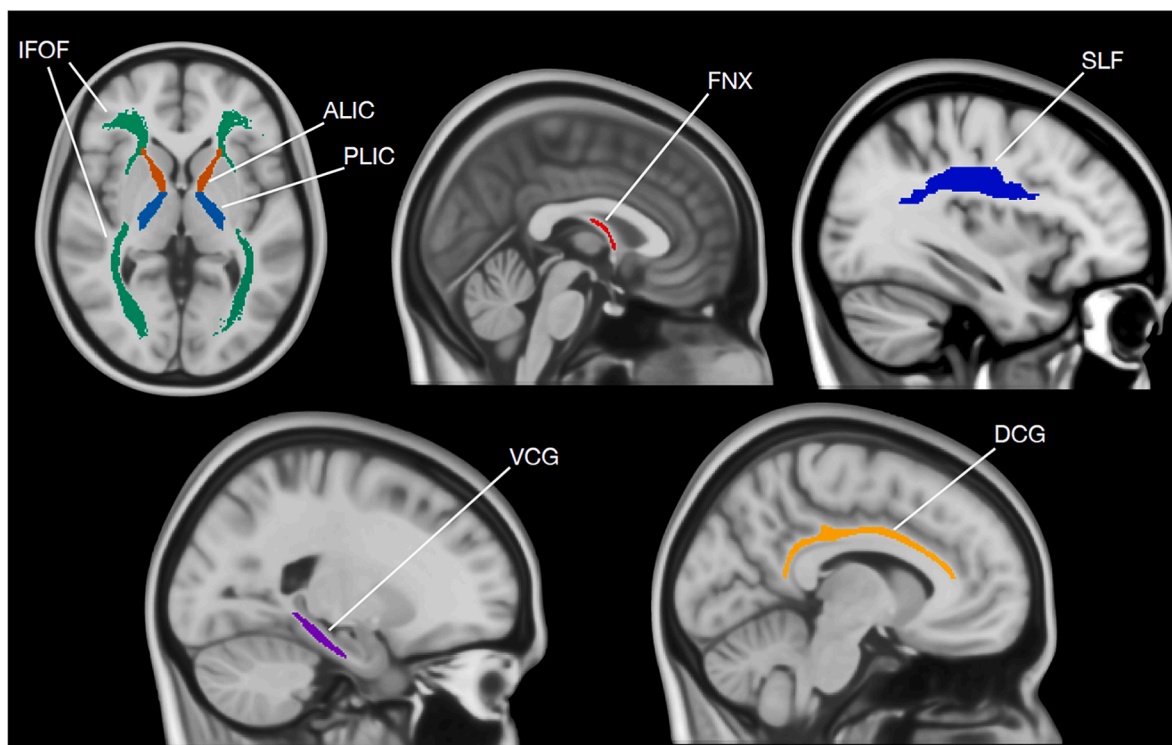

Fig. 1. Mapping of the white matter ROIs on the MNI standard brain using FSLeyes. ALIC: anterior limb of the internal capsule, PLIC: posterior limb of the internal capsule, IFOF: inferior fronto-occipital fasciculus, SLF: superior longitudinal fasciculus, DCG: dorsal cingulum, VCG: ventral cingulum, FNX: fornix.
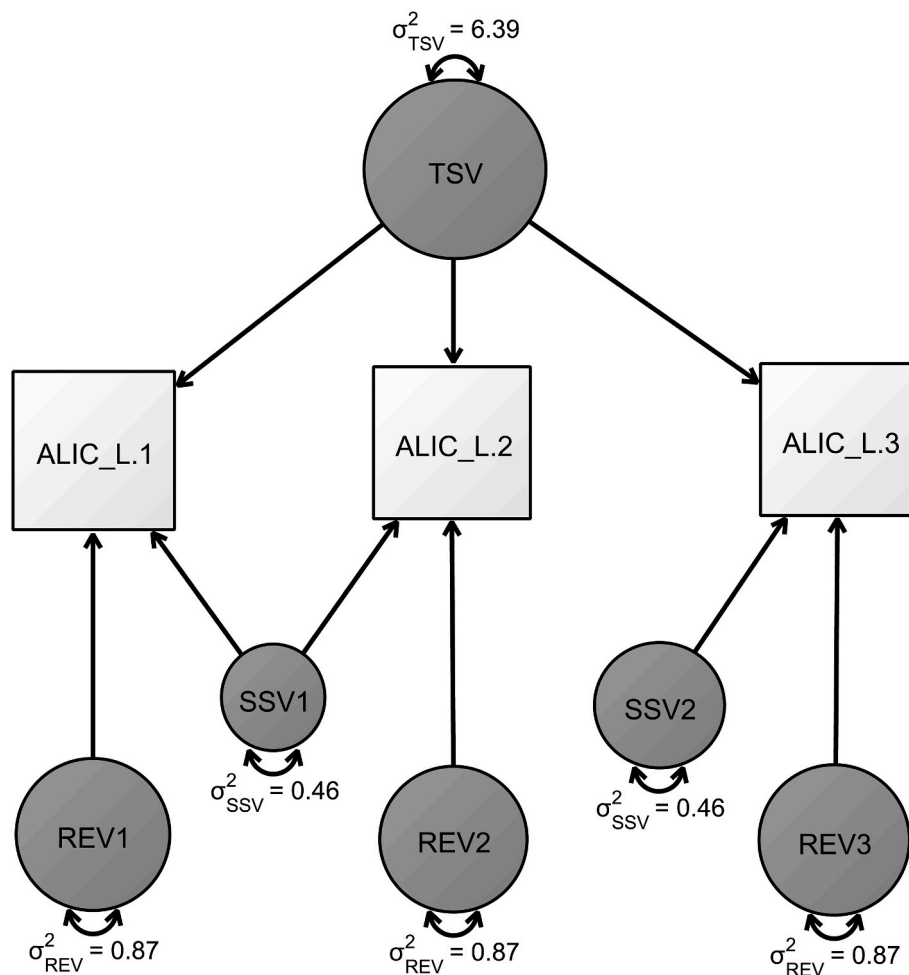
Fig. 2. Path diagram of a structural equation model. In a repeated-measures design, each participant is scanned three times, yielding three MWF values for the left ALIC (ALIC_L) at occasions 1–3. Variance component estimates TSV, SSV, and REV are represented by parameters σ^2_{TSV}, σ^2_{SSV}, and σ^2_{REV}, respectively.

error, likelihood ratio tests were used to compare the unconstrained models against the respective null models. The covariance matrix of variances shared between two scans was set up as in our previous study (Anand et al., 2019).

The variance components were used to calculate ICC as an index of reliability of a single measurement (Brandmaier et al., 2018). In addition, ICC2 was calculated to assess the reliability of the latent estimate based on the entire reliability study including all three measurement occasions (test, retest, and repositioning). Bootstrapped 95% confidence intervals, using 1000 samples, were generated for ICC and ICC2 values of the MWF and geomT$_{2IEW}$ as well as for each variance component. Acceptable target reliability was set at ICC and ICC2 ≥ 0.80 (Shrout and Fleiss, 1979). These analyses were conducted on each region (tract). All statistical analyses were conducted in R *lavaan* (Rosseel, 2012) or OpenMx (Neale et al., 2016) for the SEM, and applied *boot.ci* (Canty and Brian, 2019) for bootstrapping. The bootstrapped confidence intervals were calculated using the adjusted bootstrap percentile (bca) method in boot.ci. Bar graphs illustrating the distribution of the MWF and geomT$_{2IEW}$ variance in the selected ROIs for each variance source (TSV, SSV, REV) were generated using the ggplot2 package (Wickham, 2016) in R.

Hemisphere as a source of variance: Test-retest reliability studies on various brain metrics either treat left and right hemispheres as separate ROIs, sum the values from both hemispheres, or average them (e.g. (Homayouni et al., 2021; Jing et al., 2018):). ICED, however, allows inclusion of the hemisphere as a within-subject observed variable and additional latent sources of hemisphere-specific variability. This allows for testing how well the left and right hemispheres measure a

hypothesized underlying region factor and how much unique variance there is in each hemisphere.

The ICED model (Fig. 2) was extended by the inclusion of region-specific variance (RSV) as well residual hemisphere-error variance (HEV) for the bilateral ROIs. These latent variables capture the variance that is not hemisphere-specific but is shared over both hemispheres of a region. Thus, for each ROI (except the fornix), RSV and HEV were estimated with a latent variable at each scanning occasion (test, retest, and repositioning) with observations from the left and right hemispheres as measured within-subject variables. Parameters in the structural equation model that underlies ICED correspond to estimates of TSV, RSV, SSV, and HEV components. The path diagram in Fig. 3 illustrates the expanded ICED model for estimating the variance components of the total observed variance in MWF values in the left and right ALIC. The three ME-T$_2$ scans (scan 1, scan 2, and scan 3) of the MWF for the left ALIC are labeled in the path diagram as "ALIC_L.1", "ALIC_L.2", and "ALIC_L.3", and those for the right ALIC are labeled as "ALIC_R.1", "ALIC_R.2", and "ALIC_R.3". The same labeling convention was applied to all regions. Residual hemisphere error variances (HEVs), termed σ^2_{HEVL} (HEVs 1, 3, and 5) and σ^2_{HEVR} (HEVs 2, 4, and 6) for the left and right hemispheres, respectively, were allowed to covary within a hemisphere between the three scanning sessions (a so-called methods factor) whereas the covariances within a hemisphere were constrained to be equal.

We then tested whether the left and right hemispheric-error variances (HEVs) are different over time and whether there is differential reliability across hemispheres. To do this, we conducted a likelihood
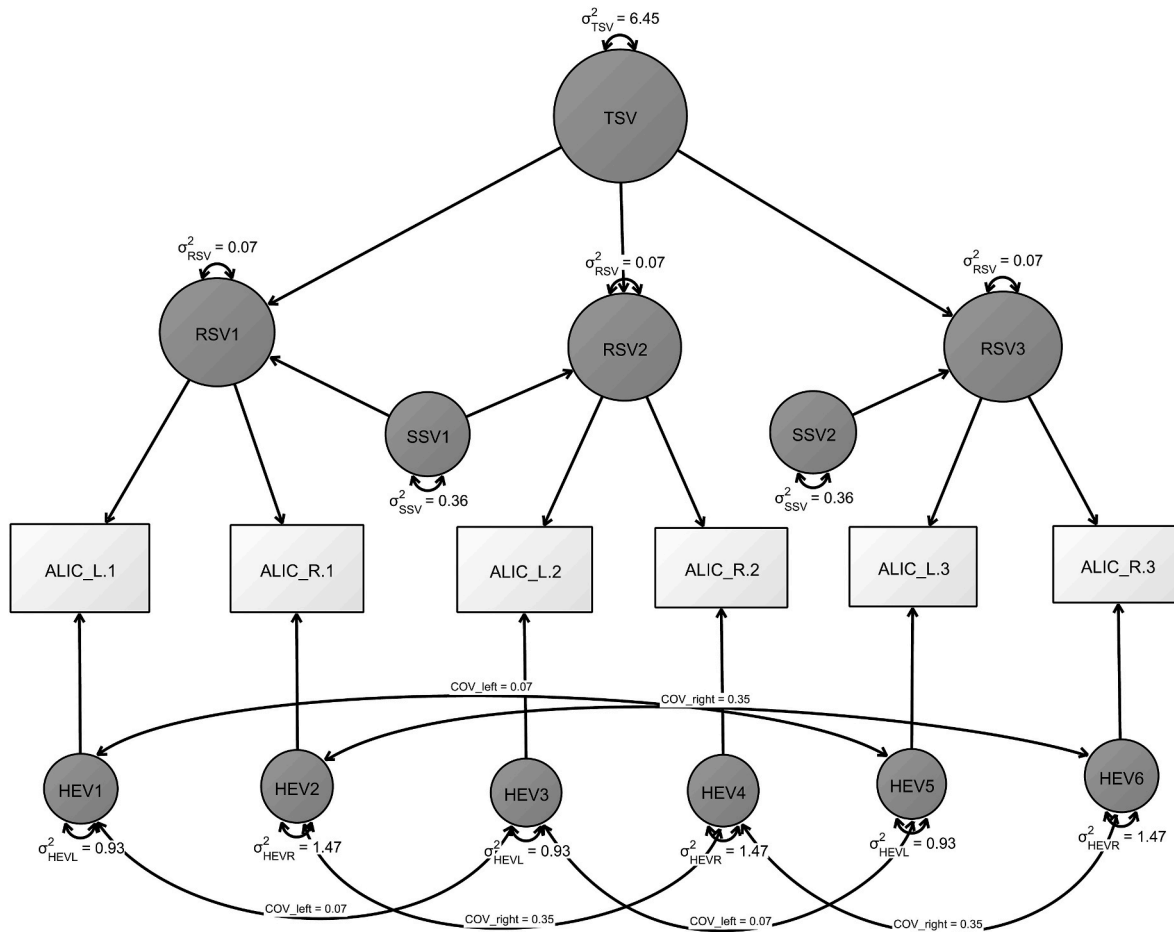
Fig. 3. Path diagram of a structural equation model. In this repeated-measures design, each participant is scanned three times, producing three MWF values for the left and right ROI (e.g., ALIC_L and ALIC_R for the ALIC) at occasions 1–3. The variances for latent variables TSV, RSV, SSV, and HEV are represented by σ^2_{TSV}, σ^2_{RSV}, σ^2_{SSV}, and σ^2_{HEV}, respectively. Hemispheric error variances (HEV) within a hemisphere (σ^2_{HEVL} for left and σ^2_{HEVR} for right) between the three sessions are allowed to covary. The covariances between the three latent variables of the left hemisphere residuals (HEV1, HEV3, HEV5) are labeled COV_left, and those of the right hemisphere residuals (HEV2, HEV4, HEV6) are labeled COV_right.

ratio (LR) test using the lavTestLRT function in R *lavaan* package against a model in which the left and right HEVs were constrained to be equal (Fig. 4). This approach also freed one degree of freedom, leading to a more parsimonious model. It also enabled testing whether the measurement error was identical across both hemispheres. A significant LR test indicates a difference in the HEVs between the two hemispheres. The LR test was conducted separately on each of the bilateral regions and a decision to include left and right HEVs separately or constrained to be the same was made based on the results of the LR test. A non-significant LR test means that there is no evidence that the HEVs for left and right hemispheres were different. In this case, we report a single joint estimate for the respective variance component. In such a case, the model with the left and right HEVs constrained to be equal was chosen. For all ROIs for MWF and geomT$_{2IEW}$, the LR test was non-significant, and hence the more parsimonious models with left and right HEVs constrained to be equal were selected (i.e., the model in Fig. 4 was applied to all ROIs). In addition, some variance component estimates attained negative values. However, model fitting appeared to be optimal as assessed by referring to the comparative fit indices. Hence, we applied non-negativity constraints to the variance components of the above model for all ROIs and compared the models with and without the non-negativity constraints using the LR test to see if the negative values were significantly different from zero. Thus, for the final analysis of the model including RSV and HEV, we proceeded with the version containing HEVs between left and right hemispheres constrained to be equal (Fig. 4) and

non-negativity constraints applied.

The ICC for this model was calculated as described previously (Anand et al., 2019; Brandmaier et al., 2018), by dividing the TSV by the sum of all variance components. For calculating the ICC2, along with the new variance components (region-specific variance and hemisphere error variance) the covariance estimates between the left and right hemisphere measurements were also included.

3. Results

One outlier was identified for geomT$_{2IEW}$ for three regions - ALIC, IFOF, and SLF. Thus, all results presented here are for 20 participants for MWF in all ROIs and for geomT$_{2IEW}$ in PLIC, DCG, and VCG; but 19 participants for geomT$_{2IEW}$ in the ALIC, IFOF, and SLF.

3.1. Mean MWF and geomT$_{2IEW}$ values

The means and the coefficient of variation (CV) for MWF (percent of the total) and geomT$_{2IEW}$ (ms) are presented in Table 1. The PLIC had the highest mean MWF and the longest geomT$_{2IEW}$.

3.2. Reliability and variance partitioning for the ROIs

The ICC and ICC2 estimates for MWF and geomT$_{2IEW}$ for all ROIs are presented in Table 2. The ICC estimates the reliability of a single
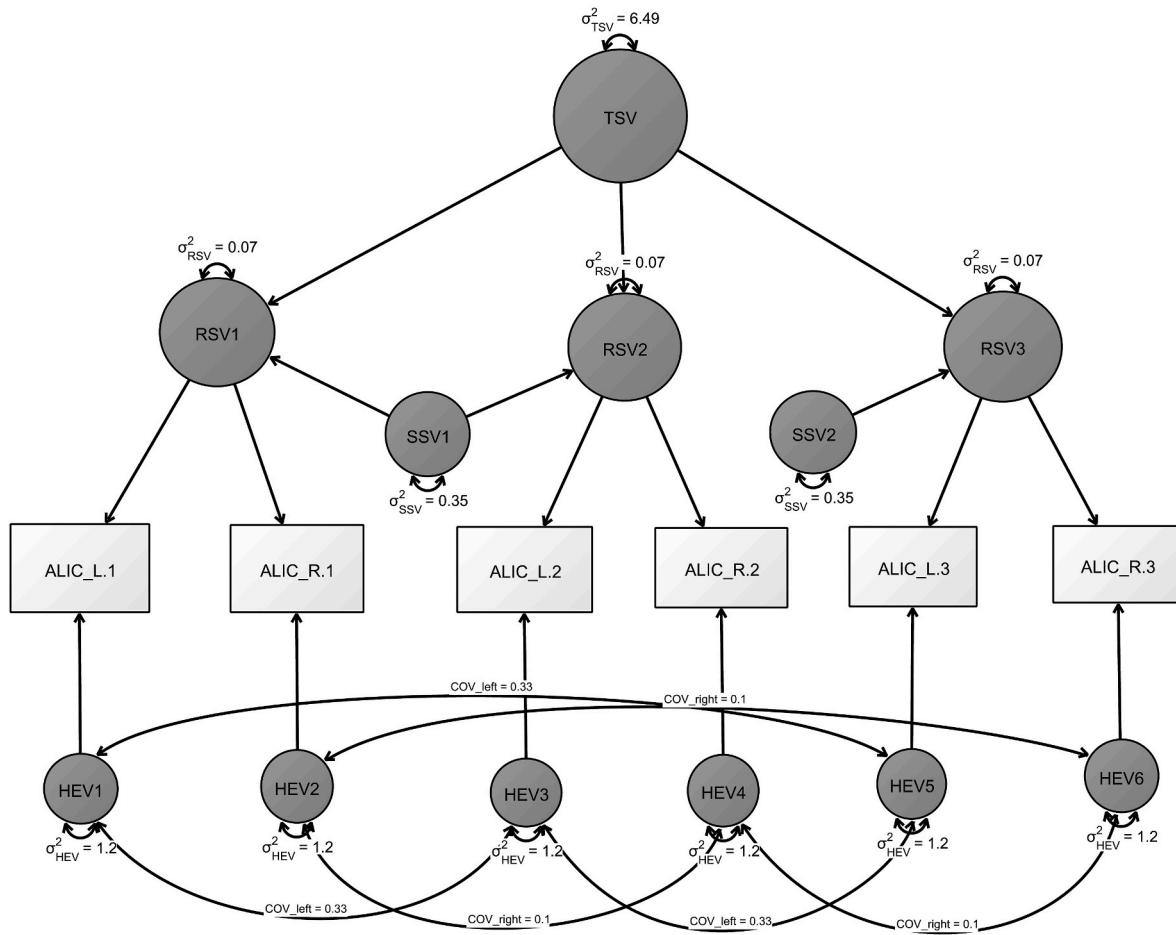
Fig. 4. Path diagram of a structural equation model, comparable to that shown in Fig. 3. The difference here is that the HEVs for the left and right hemisphere are constrained to be equal and are labeled σ^2_{HEV}.

Table 1

Mean and coefficient of variation (CV) for MWF and geomT$_{2IEW}$.

Region of Interest	MWF (CV)	Geometric T$_{2IEW}$ (CV)
ALIC_left	14.8 (0.19)	61.0 (0.01)
ALIC_right	14.9 (0.19)	61.5 (0.01)
PLIC_left	21.0 (0.15)	71.7 (0.03)
PLIC_right	20.4 (0.14)	72.1 (0.03)
DCG_left	7.1 (0.21)	62.7 (0.02)
DCG_right	6.5 (0.24)	62.7 (0.02)
VCG_left	14.7 (0.47)	67.3 (0.03)
VCG_right	13.7 (0.47)	66.7 (0.03
IFOF_left	14.7 (0.25)	66.5 (0.01)
IFOF_right	14.1 (0.28)	66.4 (0.01)
SLF_left	14.1 (0.19)	67.7 (0.01)
SLF_right	11.9 (0.18)	65.7 (0.01)
FNX	11.6 (0.25)	69.1 (0.04)

Table 2

Point estimates and 95% confidence intervals for ICC and ICC2 values for MWF and geomT$_{2IEW}$ for the regions of interest. These are based on the classic ICED model and were computed for each hemisphere separately.

Region of interest	ICC for MWF	ICC for GeomT$_2$	ICC2 for MWF	ICC2 for GeomT$_2$
ALIC_left	0.83 [0.71–0.95]	0.88 [0.72–1.05]	0.92 [0.86–0.99]	0.95 [0.86–1.04]
ALIC_right	0.82 [0.69–0.95]	0.88 [0.76–1.00]	0.92 [0.85–1.01]	0.94 [0.87–1.01]
PLIC_left	0.84 [0.75–0.94]	0.91 [0.82–0.99]	0.94 [0.90–0.99]	0.96 [0.92–1.00]
PLIC_right	0.81 [0.71–0.91]	0.92 [0.78–1.05]	0.91 [0.84–0.97]	0.97 [0.89–1.04]
DCG_left	0.79 [0.67–0.91]	0.95 [0.90–0.99]	0.91 [0.83–0.98]	0.98 [0.96–1.00]
DCG_right	0.74 [0.62–0.87]	0.91 [0.85–0.98]	0.87 [0.79–0.96]	0.96 [0.93–0.99]
VCG_left	0.79 [0.62–0.97]	0.74 [0.53–0.96]	0.89 [0.78–1.00]	0.86 [0.69–1.03]
VCG_right	0.82 [0.65–0.99]	0.83 [0.72–0.95]	0.90 [0.80–1.01]	0.92 [0.85–0.99]
IFOF_left	0.82 [0.72–0.91]	0.85 [0.72–0.98]	0.91 [0.84–0.97]	0.93 [0.85–1.01]
IFOF_right	0.84 [0.68–0.99]	0.90 [0.82–0.98]	0.92 [0.83–1.01]	0.95 [0.91–0.99]
SLF_left	0.82 [0.71–0.94]	0.92 [0.82–1.02]	0.92 [0.84–0.99]	0.97 [0.92–1.02]
SLF_right	0.79 [0.65–0.94]	0.88 [0.71–1.06]	0.89 [0.80–0.99]	0.94 [0.85–1.04]
FNX	0.77 [0.63–0.91]	0.88 [0.78–0.98]	0.89 [0.80–0.99]	0.95 [0.91–0.99]

measurement, whereas the ICC2 is the reliability of an entire experimental session in a nested design with three scanning sessions considered simultaneously. The 'repositioning' reflects the effect of a different scanning session, thus adding a third source of variability. ICC2, the construct-level reliability is higher than ICC because it conceptually reflects the average score of three measurements in our given design. Although the ICC values for MWF in some ROIs (bilateral DCG, left VCG, right SLF, FNX) and that for the geomT$_{2IEW}$ in the left VCG were less than 0.80, the ICC2 values for MWF and geomT$_{2IEW}$ of all ROIs ranged between 0.86 and 0.98, indicating high reliability of the entire experimental session (Table 2).

In Tables 3 and 4, we present the variance component estimates

Table 3
Point estimates and 95% confidence intervals for true score, session-specific, and residual variances for MWF. These values are based on the classic ICED model and were computed for each hemisphere separately.

Region of interest	True score variance (TSV)	Session-specific variance (SSV)	Residual error variance (REV)
ALIC_left	6.39 [1.94–10.84]	0.46 [-0.33 – 1.25]	0.87 [0.34–1.39]
ALIC_right	6.98 [2.16–11.81]	0.37 [-0.60 – 1.33]	1.16 [0.41–1.91]
PLIC_left	8.20 [2.65–13.75]	0	1.54 [0.85–2.23]
PLIC_right	6.43 [1.89–10.97]	0.89 [-0.01 – 1.80]	0.61 [0.22–0.99]
DCG_left	1.74 [0.51–2.97]	0.10 [-0.18 – 0.38]	0.36 [0.14–0.59]
DCG_right	1.74 [0.46–3.01]	0.27 [-0.09 – 0.63]	0.33 [0.12–0.54]
VCG_left	39.03 [11.56–66.50]	7.20 [2.11–12.29]	1.02 [0.37–1.67]
VCG_right	30.72 [8.70–52.73]	6.59 [1.89–11.31]	1.04 [0.38–1.69]
IFOF_left	11.26 [3.31–19.20]	1.96 [0.44–3.49]	0.54 [0.19–0.87]
IFOF_right	12.88 [3.92–21.84]	1.69 [0.19–3.20]	0.81 [0.30–1.31]
SLF_left	6.02 [1.81–10.23]	0.67 [-0.09 – 1.44]	0.62 [0.23–0.99]
SLF_right	3.91 [1.12–6.69]	0.63 [-0.01 – 1.27]	0.42 [0.14–0.69]
FNX	6.32 [1.75–10.89]	0.62 [-0.53 – 1.76]	1.30 [0.51–2.09]

Table 4
Point estimates and 95% confidence intervals for true score, session-specific, and residual variances for $geomT_{2IEW}$. These values are based on the classic ICED model and were computed for each hemisphere separately.

Region of interest	True score variance (TSV)	Session-specific variance (SSV)	Residual error variance (REV)
ALIC_left	2.74 [0.86–4.62]	0.08 [-0.16 – 0.32]	0.29 [0.09–0.48]
ALIC_right	3.6 [1.09–6.11]	0.38 [0.08–0.69]	0.11 [0.04–0.18]
PLIC_left	6.08 [2.05–10.11]	0.24 [-0.14 – 0.61]	0.37 [0.13–0.61]
PLIC_right	5.93 [2.02–9.83]	0.09 [-0.25 – 0.43]	0.45 [0.17–0.73]
DCG_left	2.64 [0.92–4.35]	0.01 [-0.08 – 0.10]	0.13 [0.05–0.21]
DCG_right	2.50 [0.84–4.16]	0.06 [-0.09 – 0.21]	0.18 [0.06–0.29]
VCG_left	3.45 [0.88–6.02]	0.78 [0.05–1.52]	0.43 [0.16–0.70]
VCG_right	2.65 [0.81–4.49]	0.29 [-0.03 – 0.61]	0.24 [0.09–0.39]
IFOF_left	3.64 [1.07–6.21]	0.29 [-0.11 – 0.69]	0.36 [0.13–0.56]
IFOF_right	4.68 [1.46–7.89]	0.37 [0.06–0.68]	0.13 [0.05–0.22]
SLF_left	3.29 [1.06–5.51]	0.09 [-0.09 – 0.28]	0.20 [0.07–0.34]
SLF_right	3.02 [0.93–5.12]	0.25 [0.01–0.49]	0.14 [0.05–0.24]
FNX	7.78 [2.56–13.00]	0.05 [-0.64 – 0.74]	1.04 [0.41–1.67]

explained by the three sources (TSV, SSV, and REV) for the MWF and $geomT_{2IEW}$ of the chosen ROIs.

3.3. Variance in myelin water fraction

The 95% confidence intervals around the raw scores for the three sources of variance in MWF are presented in Table 3. Session-specific variability was essentially nil for all ROIs except the bilateral VCG and IFOF, but residual non-zero variance was noted for all examined regions.

3.4. Variance in $geomT_{2IEW}$

The variances in $geomT_{2IEW}$ (with 95% confidence intervals for all three sources) are presented in Table 4. Session-specific variance was significant in the right ALIC, left VCG, right IFOF, and right SLF, as the 95% confidence intervals for those ROIs did not include 0.

Fig. 5 (a and b) illustrates the relative contribution of each of the three sources of variance in MWF and $geomT_{2IEW}$, respectively, across the examined ROIs.

3.5. Variance in MWF and $geomT_{2IEW}$ explained by adding hemisphere-error variance (HEV)

An additional source of variance - the hemisphere - was included in a model for the bilateral ROIs. The ICC and ICC2 values for MWF and $geomT_{2IEW}$ for all ROIs after the inclusion of HEV are presented in Table 5. The ICC values for MWF in some ROIs (VCG and IFOF) slightly decreased after the inclusion of HEV (compare Tables 2 and 5), but still met or exceeded the customary level of reliability (Shrout and Fleiss, 1979), except that for the SLF. Interestingly, after including the HEV in the model, the ICC values for $geomT_{2IEW}$ in all ROIs decreased and in the VCG, even dropped below the desirable reliability value. Inclusion of the hemisphere decreased the ICC2 values for MWF in the VCG, IFOF, and SLF, but still left them at or above the acceptable reliability level. Inclusion of HEV decreased the ICC2 values for $geomT_{2IEW}$ in all ROIs, with the VCG evidencing ICC2 values below the target reliability value.

Tables 6 and 7 present the values for the four sources of variance (TSV, RSV, SSV, HEV) for MWF and $geomT_{2IEW}$. Negative variances for both metrics have been forced to be zero by applying non-negativity constraints (RSV in Tables 6 and 7). Session-specific variance is significant in MWF for VCG, IFOF, and SLF (Table 6). For $geomT_{2IEW}$
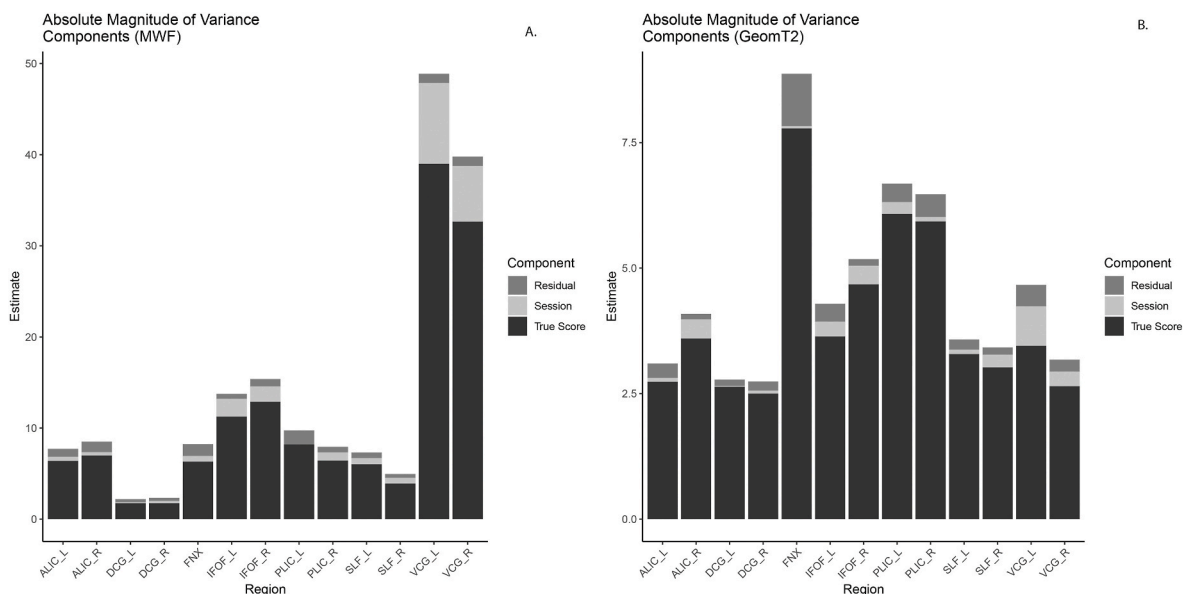

Fig. 5. Distribution of absolute magnitudes of the three sources of variances for MWF (a) and $geomT_{2IEW}$ (b) in the ROIs. Note: Instances of non-significant negative variances were constrained to zero. Further note the different scaling of the y-axes in a and b.

Table 5

Point estimates and 95% confidence intervals for ICC and ICC2 values for MWF and geomT$_{2IEW}$ for the bilateral ROIs after inclusion of region-specific latent variables (i.e., region-specific variance and hemisphere-error variance, RSV and HEV) in the model.

Region of interest	ICC for MWF	ICC for GeomT$_2$	ICC2 for MWF	ICC2 for GeomT$_2$
ALIC	0.86 [0.73–0.99]	0.86 [0.67–1.05]	0.93 [0.86–1.00]	0.91 [0.82–1.01]
PLIC	0.85 [0.78–0.92]	0.89 [0.78–0.99]	0.92 [0.87–0.98]	0.92 [0.82–1.02]
DCG	0.80 [0.67–0.93]	0.88 [0.79–0.97]	0.89 [0.83–0.96]	0.91 [0.83–0.99]
VCG	0.80 [0.63–0.96]	0.67 [0.42–0.92]	0.88 [0.77–0.99]	0.74 [0.50–1.02]
IFOF	0.82 [0.66–0.98]	0.79 [0.59–0.99]	0.89 [0.80–0.98]	0.84 [0.65–1.03]
SLF	0.72 [0.48–0.95]	0.86 [0.65–1.06]	0.79 [0.55–1.02]	0.90 [0.70–1.09]

Table 6

Point estimates and 95% confidence intervals for true score, hemisphere-specific, session-specific, and residual variances for MWF.

Region of interest	True score variance (TSV)	Region-specific variance (RSV)	Session-specific variance (SSV)	Hemisphere-error variance (HEV)
ALIC	6.49 [2.04–10.95]	0.07 [-0.34 – 0.48]	0.35 [-0.19 – 0.90]	1.20 [0.75–1.66]
PLIC	7.19 [2.21–12.17]	0.28 [-0.19 – 0.76]	0.34 [-0.42 – 1.09]	1.26 [0.75–1.77]
DCG	1.68 [0.48–2.87]	0.11 [-0.05 – 0.28]	0.15 [-0.09 – 0.39]	0.32 [0.19–0.45]
VCG	33.87 [9.17–58.57]	0.63 [-0.09 – 1.36]	6.68 [1.84–11.52]	2.87 [1.32–4.42]
IFOF	11.36 [3.19–19.53]	0.01 [-0.28 – 0.30]	1.66 [0.41–2.91]	1.66 [0.89–2.42]
SLF	3.85 [0.69–7.01]	0	0.60 [0.10–1.11]	1.83 [0.89–2.77]

Table 7

Point estimates and 95% confidence intervals for true score, hemisphere-specific, session-specific, and residual variances for geomT$_{2IEW}$.

Region of interest	True score variance (TSV)	Region-specific variance (RSV)	Session-specific variance (SSV)	Hemisphere-error variance (HEV)
ALIC	2.83 [0.79–4.87]	0	0.19 [-0.02 – 0.49]	0.53 [0.28–0.77]
PLIC	5.33 [1.63–9.03]	0	0.13 [-0.01 – 0.33]	1.09 [0.56–1.62]
DCG	2.22 [0.67–3.77]	0.05 [-0.01 – 0.11]	0.04 [-0.04 – 0.12]	0.44 [0.20–0.68]
VCG	2.14 [0.24–4.04]	0	0.31 [0.02–0.60]	1.52 [0.76–2.28]
IFOF	3.21 [0.68–5.73]	0	0.28 [0.03–0.54]	1.17 [0.53–1.80]
SLF	2.74 [0.73–4.75]	0.01 [-0.08 – 0.09]	0.12 [-0.02 – 0.27]	0.65 [0.30–0.99]

(Table 7), the VCG and IFOF experience significant session-specific variance. For both, MWF and geomT$_{2IEW}$, all ROIs evidenced significant contribution of the hemisphere-error variance (HEV) but no region-specific variance (RSV).

Fig. 6 (a and b) illustrates the relative contribution of each of the four sources of variance in MWF and geomT$_{2IEW}$, respectively, across the examined bilateral ROIs.

Table 8 shows the covariances between all repeated measures within one hemisphere.

4. Discussion

We evaluated the test–retest reliability of MRI-derived indices of the regional microstructural characteristics in seven white matter tracts and found good to excellent reliability of MWF and geomT$_{2IEW}$ - across all examined regions. The sources of unreliability varied. Measurement noise (residual variance) significantly contributed to unreliability for both indices in all regions. However, the effect of repositioning was significant only in relatively long association tracts that run along the posterior-anterior aspect of the brain: VCG, SLF, and IFOF. Although, there was no significant contribution of the hemisphere to the region-specific variance in MWF and geomT$_{2IEW}$ in any ROI (Tables 6 and 7), the ICC2 for the MWF in the SLF and the geomT$_{2IEW}$ in the VCG seemed to suffer after including the hemisphere in the model. This could indicate that these ROIs have larger and less stable hemispheric differences compared to the others.

The first reliability model (illustrated in Fig. 2) was designed to separate three different sources of variance in the observed scores, that is, true-score variance, session-specific variance, and residual variance. The second model (illustrated in Figs. 3 and 4) extended this idea by adding hemisphere-specific indicators. This allowed us to gauge how well common variance across the hemispheres can be measured (akin to a sum score of left and right measurements). This model is similar to a multi-method model, in which a given construct is measured by multiple methods, where methods here correspond to hemispheres. For each pair of measurements within one hemisphere, we added covariances that capture common method variance within the hemisphere. That is, the covariances capture what is common to the hemisphere but not to the common factor of interest. In other words, they represent stable individual differences that are specific to each hemisphere but not shared across hemispheres. A statistically equivalent reparameterization of this model includes two explicit latent hemispheric "methods" factors that each load onto all measurements from one hemisphere. Instead of estimating two covariance parameters (between each pair of indicators of one hemisphere), one would estimate the variances of the two method factors. In our current implementation, the residual error variances represent the sum of the stable hemisphere-specific variances and residual variance. The re-parametrization would separate the hemisphere-specific variances from the residual variance component and may be better suited if an investigation of stable individual differences unique to hemisphere are of interest.

Thus, in extension of the previous report (Anand et al., 2019), we found that, overall, both MWF and geomT$_{2IEW}$ are reliable indices suitable for longitudinal investigation. Repositioning, albeit a threat to reliability, did not contribute sufficiently to suppress overall reliability of measurements. Importantly, we found no evidence of differential unreliability as a function of hemispheric location, thus alleviating a threat to validity of investigations that focus on hemispheric asymmetry in developmental or pathological change.

5. Limitations

The sample used in this study was small, age-heterogeneous, and restricted to healthy adults. Thus, it is unclear how well the reliability of the MWF and geomT$_{2IEW}$ will hold up in more homogeneous samples and in patients with significant burden of white-matter lesions. The statistical power in this sample was rather low, and the confidence intervals around the reliability estimates are wide. Therefore, it is possible that investigation in a larger sample (approximately 200 participants, for the observed effect sizes) would detect small differences in reliability across the white matter tracts. The small sample size also precludes analyses of age and sex differences in the reliability of MWF and geomT$_{2IEW}$ measurements. Both effects need to be examined in studies with greater statistical power.
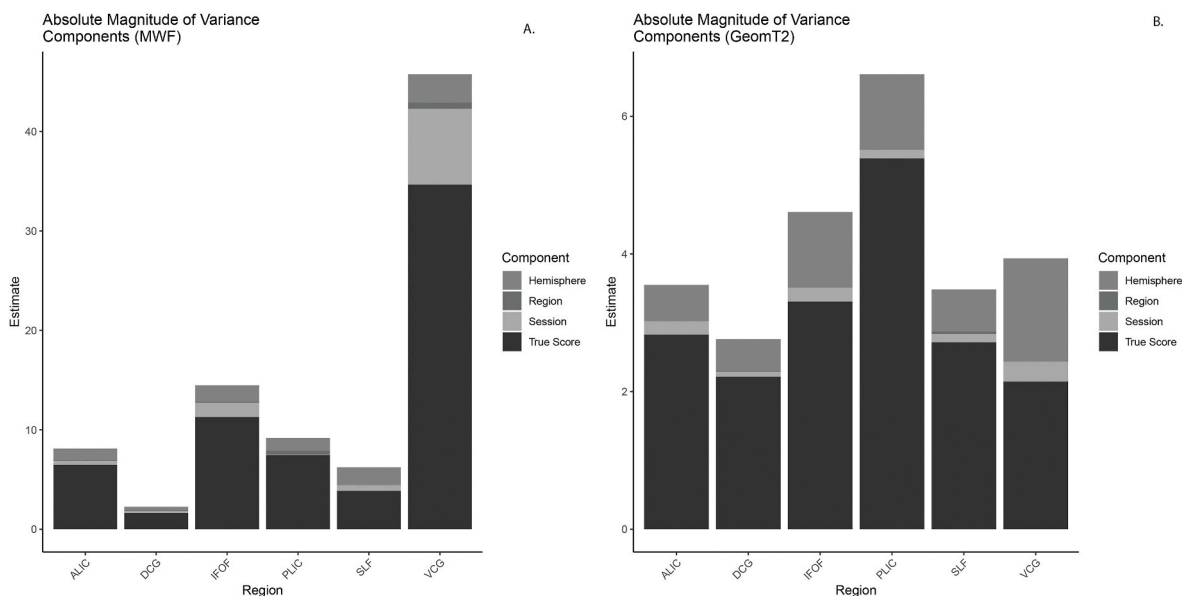
Fig. 6. Distribution of absolute magnitudes of the four sources of variance for MWF (a) and geomT$_{2IEW}$ (b) in the selected bilateral ROIs. Note: Instances of non-significant negative variances were constrained to zero. Further note the different scaling of the y-axes in a and b.

Table 8
Covariances among sessions within a hemisphere (test, retest, repositioning). Significant covariances are indicated by *: p < 0.05.

Region of interest	Hemisphere	Covariance for MWF	Covariance for GeomT$_{2IEW}$
ALIC	Left	0.33 [-0.20 – 0.86]	0.23 [-0.04 – 0.50]
	Right	0.10 [-0.43 – 0.63]	0.35* [0.08–0.62]
PLIC	Left	0.12 [-0.51 – 0.75]	0.66* [0.13–1.19]
	Right	0.62 [-0.05 – 1.29]	0.65* [0.12–1.18]
DCG	Left	0.09 [-0.07 – 0.25]	0.37* [0.13–0.61]
	Right	0.05 [-0.11 – 0.21]	0.30* [0.06–0.54]
VCG	Left	1.92* [0.23–3.61]	0.85* [0.04–1.65]
	Right	2.47* [0.76–4.18]	1.26* [0.48–2.04]
IFOF	Left	1.01* [0.19–1.83]	0.72* [0.05–1.39]
	Right	0.77 [-0.05 – 1.59]	1.01* [0.36–1.66]
SLF	Left	1.19* [0.23–2.15]	0.45* [0.09–0.80]
	Right	1.38* [0.40–2.40]	0.45* [0.09–0.80]

6. Summary and conclusions

Two white matter microstructure indices, MWF and geomT$_{2IEW}$, showed high reliability in seven major white matter tracts, for single measurements as well as for averages over all three measurements. The contribution of repositioning to unreliability was, however, significant and more widespread in longer association fibers. Hemispheric location per se was not a significant contributor to unreliability.

Key points

1.Myelin water fraction and non-myelin T$_2$ times estimated by MRI.
2.Test-retest reliability with participant repositioning design was employed.
3.High reliability found in seven commonly studied white matter tracts, irrespective of hemisphere.

Data availability statement

Multiecho T$_2$ images, data files with individual MWF and geometric T$_2$ values, as well as R code files for model setup and bar graph generation will be shared via the Open Science Framework platform repository at https://osf.io/8dsh3/.

Ethics approval

This study contains an extended analysis of the MRI data that were previously collected by our group. Non-overlapping analyses have been published in Arshad et al. (2017), Brandmaier et al. (2018), and Anand et al. (2019). The data collection was conducted following the protocol approved by the Wayne State University Institutional Review Board, in accordance with the Declaration of Helsinki (1964).

Informed consent

All participants provided informed consent.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alonso-Ortiz, E., Levesque, I.R., Pike, G.B., 2015. MRI-based myelin water imaging: a technical review. Magn. Reson. Med. 73 (1), 70–81. https://doi.org/10.1002/mrm.25198.

Anand, C., Brandmaier, A.M., Arshad, M., Lynn, J., Stanley, J.A., Raz, N., 2019. White-matter microstructural properties of the corpus callosum: test–retest and repositioning effects in two parcellation schemes. Brain Struct. Funct. 224 (9), 13.

Arshad, M., Stanley, J.A., Raz, N., 2017. Test-retest reliability and concurrent validity of in vivo myelin content indices: myelin water fraction and calibrated T1 w/T2 w image ratio. Hum. Brain Mapp. 38 (4), 1780–1790. https://doi.org/10.1002/hbm.23481.

Barkovich, A.J., Kjos, B.O., Jackson Jr., D.E., Norman, D., 1988. Normal maturation of the neonatal and infant brain: MR imaging at 1.5 T. Radiology 166 (1 Pt 1), 173–180. https://doi.org/10.1148/radiology.166.1.3336675.

Bartzokis, G., 2004. Age-related myelin breakdown: a developmental model of cognitive decline and Alzheimer's disease. Neurobiol. Aging 25 (1), 5–18 author reply 49-62.

Brandmaier, A.M., Wenger, E., Bodammer, N.C., Kuhn, S., Raz, N., Lindenberger, U., 2018. Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). Elife 7. https://doi.org/10.7554/eLife.35718.

Canty, A., Brian, R., 2019. Boot: Bootstrap R (S-Plus) Functions. R package version 1, pp. 3–24.

Catani, M., Howard, R.J., Pajevic, S., Jones, D.K., 2002. Virtual in vivo interactive dissection of white matter fasciculi in the human brain. Neuroimage 17 (1), 77–94. https://doi.org/10.1006/nimg.2002.1136.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 10.

Homayouni, R., Yu, Q., Ramesh, S., Tang, L., Daugherty, A.M., Ofen, N., 2021. Test–retest reliability of hippocampal subfield volumes in a developmental sample: implications for longitudinal developmental studies. J. Neurosci. Res. 13. https://doi.org/10.1002/jnr.24831.

Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D.S., Mori, S., 2008. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. Neuroimage 39 (1), 336–347. https://doi.org/10.1016/j.neuroimage.2007.07.053.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. *Neuroimage* 62 (2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015.

Jing, B., Liu, B., Li, H., Lei, J., Wang, Z., Yang, Y., Xu, Z.D., 2018. Within-subject test-retest reliability of the atlas-based cortical volume measurement in the rat brain: a voxel-based morphometry study. J. Neurosci. Methods 307, 46–52. https://doi.org/10.1016/j.jneumeth.2018.06.022.

Karch, J.D., Filevich, E., Wenger, E., Lisofsky, N., Becker, M., Butler, O., Kühn, S., 2019. Identifying Predictors of Within-Person Variance in MRI-Based Brain Volume Estimates. Neuroimage (in press).

Laule, C., Leung, E., Lis, D.K., Traboulsee, A.L., Paty, D.W., MacKay, A.L., Moore, G.R., 2006. Myelin water imaging in multiple sclerosis: quantitative correlations with histopathology. Mult. Scler. 12 (6), 747–753. https://doi.org/10.1177/1352458506070928.

Lynn, J.D.a.A., C, Arshad, M., Homayouni, R., Rosenberg, D.R., Ofen, N., Raz, N., Stanley, J.A., 2021. Microstructure of human corpus callosum across the lifespan: regional variations in axon caliber, density, and myelin content. Cerebr. Cortex 31 (2), 1032–1045. https://doi.org/10.1093/cercor/bhaa272.

MacKay, A., Whittall, K., Adler, J., Li, D., Paty, D., Graeb, D., 1994. In vivo visualization of myelin water in brain by magnetic resonance. Magn. Reson. Med. 31 (6), 673–677.

Mori, S., Wakana, S., van Zijl, P.C.M., Nagae-Poetscher, L.M., 2005. MRI Atlas of Human White Matter, first ed. Elsevier Science.

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

Peters, A., 2009. The effects of normal aging on myelinated nerve fibers in monkey central nervous system. Front. Neuroanat. 3, 11. https://doi.org/10.3389/neuro.05.011.2009.

Radloff, L.S., 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl. Psychol. Meas. 1 (3), 7.

Rosseel, 2012. Lavaan : an R package for structural equation modeling. J. Stat. Software 48, 36.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428.

von Oertzen, T., Brandmaier, A.M., Tsang, S., 2015. Structural equation modeling with Onyx. Struct. Equ. Model.: A Multidiscip. J. 22, 12.

Webb, S., Munro, C.A., Midha, R., Stanisz, G.J., 2003. Is multicomponent T2 a good measure of myelin content in peripheral nerve? Magn. Reson. Med. 49, 8.

Whittall, K.P., MacKay, A.L., Li, D.K., Vavasour, I.M., Jones, C.K., Paty, D.W., 2002. Normal-appearing white matter in multiple sclerosis has heterogeneous, diffusely prolonged T(2). Magn. Reson. Med. 47 (2), 403–408.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imag. 20, 45–57.