PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2017), 24–33. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2016 Society for Psychophysiological Research
DOI: 10.1111/psyp.12769

A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control

W. BOEKEL, B. U. FORSTMANN, AND M. C. KEUKEN

Amsterdam Brain & Cognition Centre, University of Amsterdam, Amsterdam, The Netherlands, and Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands

Abstract

Recent efforts to replicate structural brain-behavior correlations have called into question the replicability of structural brain measures used in cognitive neuroscience. Here, we report an evaluation of test-retest reliability of diffusion tensor imaging (DTI) measures, including fractional anisotropy, mean diffusivity, axial diffusivity, and radial diffusivity, in several white matter tracts previously shown to be involved in cognitive control. In a data set consisting of 34 healthy participants scanned twice on a single day, we observe overall stability of DTI measures. This stability remained in a subset of participants who were also scanned a third time on the same day as well as in a 2-week follow-up session. We conclude that DTI measures in these tracts show relative stability, and that alternative explanations for the recent failures of replication must be considered.

Descriptors: Test-retest reliability, Diffusion tensor imaging, DTI, Cognitive control

Many studies in the cognitive neurosciences aim to investigate the link between brain and behavior. Recently, researchers have exploited significant advances in diffusion weighted imaging (DWI) to detect subtle differences in brain structure associated with differences in behavioral measures (e.g., Kanai & Rees, 2011) including the stop-signal task (Aron, Behrens, Smith, Frank, & Poldrack, 2007; Forstmann et al., 2012; Rae, Hughes, Anderson, & Rowe, 2015), conflict tasks such as the Simon task (Forstmann et al., 2008), and strategic decision-making tasks (Coxon, van Impe, Wenderoth, & Swimmen, 2012; Forstmann et al., 2010; Mulder, Boekel, Ratcliff, & Forstmann, 2014). In a recent study from our group using a preregistered confirmatory framework (Boekel, Forstmann, & Wagenmakers, 2016), we attempted to replicate studies that adopt this structural brain-behavior (SBB) correlational approach. Confirmatory Bayesian statistical tests suggested that eight out of 17 SBB effects were reliably absent in our independent replication data set. This apparent instability of effects calls into question the test-retest reliability of DWI-derived measures (for a comprehensive discussion of our replication results, see Boekel, Wagenmakers et al., 2015; Kanai, 2015; Muhlert & Ridgway, 2016).

This work was supported by a Vidi grant from the Dutch Organization for Scientific Research (BUF) and an ERC starter grant (BUF). We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Computer Cluster. We also thank Martijn Mulder for providing the vmPFC mask and Dora Matzke for assisting with the analysis of the SSRT behavioral data.

Address correspondence to: Wouter Boekel, University of Amsterdam, Valckenierstraat 651, 1018 XE Amsterdam, The Netherlands. E-mail: W.E.Boekel@uva.nl

Previous investigations into the test-retest reliability of DWI have generally suggested stability (Buchanan, Pernet, Gorgolewski, Storkey, & Bastin, 2014; Fox et al., 2012; Heiervang, Behrens, Mackay, Robson, & Johansen-Berg, 2006; Jansen, Kooi, Kessels, Nicolay, & Backers, 2007; Jovicich et al., 2014; Madhyastha et al., 2014; Owen et al., 2013; Pfefferbaum, Adalsteinsson, & Sullivan, 2003; Vollmar et al., 2010; Wang, Abdi, Bakhadirov, Diaz-Arrastia, & Devous, 2012). These studies have mostly used whole-brain methods to calculate an overall estimate of the reliability of the DWI-derived measures. Some have also specifically tested major white matter tracts to investigate the possibility that areas of low reliability selectively impede robust measurements of DWI measures in white matter tracts such as the corpus callosum (Heiervang et al., 2006) and the inferior frontooccipital fasciculus (IFOF; Wang et al., 2012). Yet, another class of tracts is often investigated by researchers in the field of cognitive neuroscience, particularly those adopting the SBB approach. Informed by functional findings, researchers use probabilistic tractography to identify white matter pathways between areas found to be involved in the performance of a task. After the delineation of such a tract, DTI measures can be extracted and correlated to individual differences in behavior.

For example, Mulder, Wagenmakers, Ratcliff, Boekel, and Forstmann (2012) found that providing participants with prior information about the reward balance of a two-alternative forced choice perceptual decision-making task elicits activation in the right ventromedial prefrontal cortex (vmPFC). Subsequently, in Mulder, et al. (2014), a white matter pathway between the vmPFC and the subthalamic nucleus (STN) was estimated using probabilistic tractography. The tract strength between the vmPFC and STN was then shown to be quantitatively predictive of individual differences

in value-based choice bias. This SBB finding requires the assumption that diffusion tensor fitting on data from a single DWI scanning session provides robust diffusion measures. However, this assumption is challenged by a study from Jansen et al. (2007). In their study, an area of decreased reliability across sessions in the basal ganglia (BG) was found. The authors argued that this decrease in reliability was possibly due to increased iron content in the BG causing susceptibility artifacts in the DWI data (Drayer et al., 1986). It is possible that tracts originating from the BG are negatively affected in terms of their test-retest reliability. This example suggests that it is not sufficient to investigate whole-brain DWI robustness; specific white matter pathways should be tested for test-retest reliability to exclude the influence of local islands of increased variability.

Here, we report an evaluation of test-retest reliability of diffusion measures in white matter tracts of the cognitive control network delineated by probabilistic tractography. We specifically inspect the DTI measures fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD), radial diffusivity (RD), tract strength, and tract volume. FA is a measure of the degree of anisotropic diffusion of molecules, where low FA values correspond to equidirectional (un)restricted diffusion (i.e., Brownian motion), and high FA values reflect restricted linear diffusion. MD is a measure of the total diffusion. AD is defined by the principal eigenvalue of the tensor model, which represents the degree of diffusion in the main diffusion direction. RD is defined by the average of the second and third eigenvalue of the tensor model, representing the degree of diffusion perpendicular to the main diffusion direction. Tract strength is derived from the tractography procedure (see Method, "Probabilistic tractography"), and tract volume is given in mm^3.

We investigate these measures in four white matter pathways: (1) the tract between the subthalamic nucleus (STN) and the inferior frontal cortex (IFC) shown to be involved in stopping behavior (Aron et al., 2007; Aron, Robbins, & Poldrack, 2014); (2) the IFOF involved in the Simon task (Forstmann et al., 2008); (3) the tract between the striatum (in our analysis comprising putamen and caudate, excluding the nucleus accumbens) and presupplementary motor area (pre-SMA), which has been implicated in the speed-accuracy tradeoff (Forstmann et al., 2010); and (4) the tract between STN and vmPFC, involved in value-driven choice bias (Mulder et al., 2014). All but the IFOF were identified using probabilistic tractography in our dataset (the IFOF was extracted from the JHU white matter tractography atlas implemented in FSL, Hua et al., 2008; and registered to individual space).

Here, we investigate a structural DWI data set including 34 participants who were scanned in two DWI sessions 1 h apart. Of these 34 participants, 15 were additionally scanned 1 h after the second scan session, as well as in a 2-week follow-up scan session. In comparison to the majority of previous DWI reliability studies, this data set provides (a) a somewhat larger sample size, and (b) relatively more data per scan. We have relatively more data per scan because the scan sessions include four repetitions of a DWI sequence, each of which is merged within-session, prior to testing reliability between sessions.

The reliability assessment of DWI measures in the abovementioned four tracts is particularly interesting to researchers in the field of cognitive control. More generally, it allows the investigation of potential sources of variance, which is important when investigating relationships between brain structure and behavior.

Method

Participants

This data set is an extension of an unpublished pilot study set up as an exploration into the possibility that practice effects on a stop-signal task elicit short-term structural changes in a network previously labeled the *cognitive control network* (Aron et al., 2007, 2014). This pilot study initially comprised data from 15 participants, who were divided into a stop group (nine participants performing the stop-signal task between scans), and a passive control group (six participants who did not perform the stop-signal task between scans). Later, this data set was expanded to include an active control group of seven participants, who performed a go task (i.e., the stop-signal task without stop signals) between DWI scans. Moreover, nine and three additional participants were respectively assigned to the stop and passive control group. As such, our final data set consists of structural brain scans of 34 healthy young participants (18 females) with a mean age of 22.76 ($SD = 3.12$, range 19.17–35.67). The study was approved by the local ethics committee at the University of Amsterdam. All participants gave their written consent prior to scanning and received a monetary compensation.

Experimental Design

Figure 1 displays the design. Participants in the stop-signal group and the active control group started with a go task to familiarize them with the left/right decision component of our task. This practice session was followed by the first DWI scan. Subsequently, the stop-signal group performed behavioral stop-signal tasks between scans, and the active control group performed go tasks between scans. The passive control group did not perform any task in between the scans. Participants in this group were asked to remain in the waiting room of the scanning center while they waited for the next scan. Our stop task was a computerized perceptual two-alternative forced choice directional discrimination task using arrows pointing left or right, with the inclusion of auditory tones prompting the participant to inhibit their response. Stop-signal delay started at 190 ms and was updated after every stop trial by an addition or subtraction of 50 ms, depending on the subjects' stop-respond rate so far, leading to an eventual average stop-respond rate of 50%. The go task was a copy of the stop task using only go trials. The stop-signal reaction time (SSRT) was estimated per session using the BEEST (Bayesian ex-Gaussian estimation of stop-signal reaction time [RT] distributions) software (version 2.0; Matzke et al., 2013). The MCMC (Markov chain Monte Carlo) sampling settings were number of chains: 3; number of samples: 20,000; number of burn-in: 5,000; and amount of thinning: 5. All incorrect RTs and RTs shorter than 200 ms were excluded for the estimation of the SSRT.

The final data set consisted of a group of 34 participants scanned in Session 1 and 2, of which a smaller subset of 15 participants were also scanned in a third session on the same day, as well as in a 2-week follow-up session.

DWI imaging acquisition. Imaging data were acquired on a 3T Philips Achieva XT scanner (80 mT/m maximum amplitude gradient strength and a maximum slew rate of 200 mT/m/ms) using a 32-channel head coil. For each participant, a T_1 anatomical scan was acquired (T_1 turbo field echo, 220 coronal slices with an isotropic voxel resolution of 1 mm, field of view = 240 × 188 ×

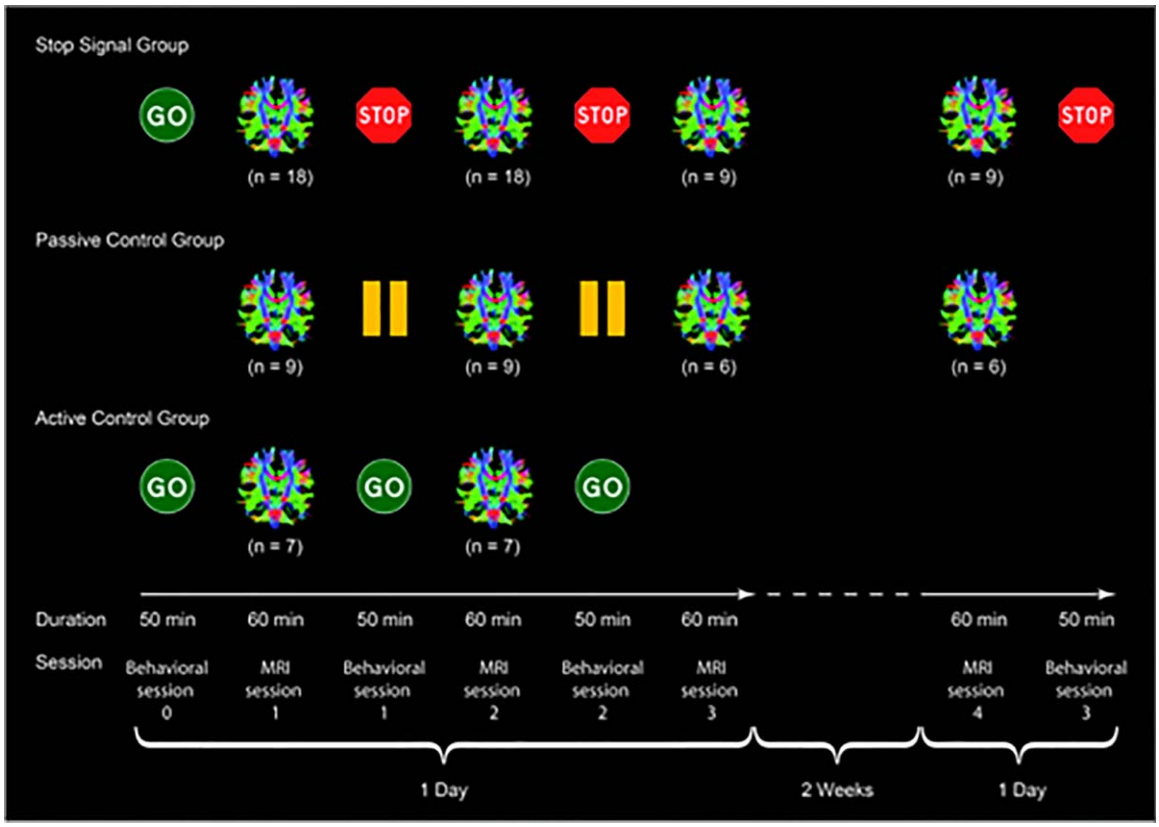

Figure 1. Experimental design. Three groups underwent scanning interleaved with a stop task (stop signal group), a go task (active control group), or an equal amount of time to be spent in the waiting room of the scanning center (passive control group). The stop signal and active control groups performed on a go practice task prior to the first scanning session to familiarize them with the left/right discrimination aspect of the task.

220 mm, flip angle = 8°, TR = 8.4 ms, TE = 3.9 ms, SENSE factor (RL) = 2.5, SENSE factor (FH) = 2, bandwidth 191.4 Hz/Px, acquisition time 3.06 min).

In each DWI scanning session, four repetitions of a multislice spin echo (MS-SE), single-shot DWI scans were acquired on a 3T MRI (60 transverse slices with an isotropic voxel resolution of 2 mm, field of view = 224 × 224, TR = 7,545 ms, TE = 86 ms, SENSE factor (AP) = 2, bandwidth = 32.1 Hz/Px, acquisition time 5.30 min each). Diffusion weighting was isotropically distributed along 32 directions (b value = 1,000 s/mm^2). For each repetition, six images with no diffusion weighting (b0; b value = 0 s/mm^2) were acquired and averaged by the scanner before adding to the raw data. All DWI data are made freely available on http://www.nitrc.org/projects/dwi_test-retest/.

DWI preprocessing. All DWI data (pre)processing and analyses were carried out using FMRIB's Software Library (FSL, version 5.0.8; www.fmrib.ox.ac.uk/fsl; Smith et al., 2004). For each participant and session, all four DWI repetitions were concatenated and corrected for eddy currents. Affine registration was used to register each volume to a reference volume (Jenkinson & Smith, 2001). A single image without diffusion weighting (b0; b value = 0 s/mm^2) was extracted from the concatenated data, and nonbrain tissue was removed using FMRIB's brain extraction tool (Smith, 2002) to create a brain mask that was used in subsequent analyses. DtiFit (Behrens et al., 2003) was applied to fit a tensor model at each voxel of the data (Smith et al., 2004) to derive FA, MD, AD, and RD measures for further analyses.

ROI definition. We extracted the striatum (STR) and STN regions of interest (ROIs) from the probabilistic atlas from Keuken et al. (2014). The pre-SMA ROI was drawn in MNI (Montreal Neurological Institute) space by using the coordinates reported by Johansen-Berg et al. (2004). The IFC ROI was extracted from the Harvard/Oxford atlas included in FSL (Desikan et al., 2006). The vmPFC ROI was provided by Mulder et al. (2014). Finally, the IFOF was extracted from the JHU white matter tractography atlas included in FSL (Hua et al., 2008). Bilateral ROIs were extracted and separated by hemisphere. As the right vmPFC was functionally defined, we generated the left hemisphere version of this ROI based on its mirrored x coordinate. All probabilistic ROIs were thresholded at 10%. In order to bring these ROIs into individual space, we first registered the standard MNI template to the participant's whole-brain MPRAGE (magnetization-prepared rapid gradient echo) using FLIRT (12 degrees of freedom, correlation ratio, trilinear interpolation). The registered MNI template was then registered to individual b0 images. We subsequently nonlinearly optimized these transformations using the symmetric image normalization method, which is part of the advance normalization tools (Avants et al., 2008). Using the resulting transformation matrices and warp fields, the ROIs were then transformed into individual space. Figure 2 provides an overview of the resulting ROIs that were subsequently used in probabilistic tractography.

Probabilistic tractography. BedpostX (Behrens et al., 2003) was applied to the preprocessed DWI data to estimate voxelwise diffusion parameter distributions. Estimation of tract strengths was conducted using probabilistic tractography (Behrens et al., 2003). Fifty

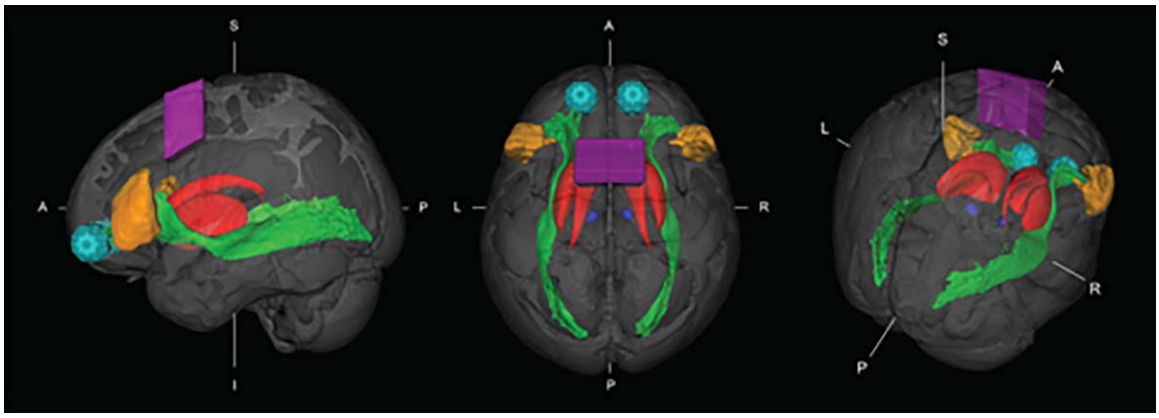

Figure 2. Masks used in probabilistic tractography. Cyan = ventromedial prefrontal cortex (vmPFC); green = inferior frontooccipital fasciculus (IFOF); ochre = inferior frontal cortex (IFC); purple = presupplementary motor area (pre-SMA); red = striatum (STR); blue = subthalamic nucleus (STN).

thousand tracts were sampled from each voxel in the seed masks at a curvature threshold of 0.2. We used two separate tractography analyses per tract: a seed-to-classification analysis, which we used to extract tract strength measures; and a seed-to-termination analysis, which we used to generate the tract images for the assessment of tract-average FA/MD/AD/RD test-retest reliability.

First, in the seed-to-classification analysis, we used a seed mask from which to start tracking, an individually drawn midline mask to prevent fibers from crossing over to the other hemisphere, and a classification mask serving as target for the tractography. This analysis returns an image containing, for each voxel in the seed mask, the number of samples reaching the classification mask. To remove any spurious connections, this image was thresholded at 10% of robust range using the fslmaths -thrP command. Subsequently, the number of nonzero voxels was divided by the total number of voxels in the seed mask, resulting in a value that represents the proportion of the seed mask that was probabilistically connected to the classification mask. A similar procedure was applied in the opposite direction (where the seed and classification masks were switched). Tract strength was defined as the average of the two proportions that resulted from the seed-to-classification and classification-to-seed analyses.

Second, in the seed-to-termination analysis, we used a seed mask from which to start tracking, an individually drawn midline mask to prevent fibers from crossing over to the other hemisphere, a waypoint mask (the inclusion of which effectively discards any tracts not reaching it), and a termination mask that terminates but keeps the tracts reaching this mask. In these analyses, the waypoint mask and termination mask were always the same. The image resulting from this analysis has probabilistic information only in voxels where tracts passed through that (a) originated from the seed-mask, (b) did not cross over to the contralateral hemisphere, and (c) reached the termination/waypoint mask.

Tract-based spatial statistics. The tracts delineated by the previously described tractography procedure were used in a reliability assessment of DTI measures. After having visually inspected these tracts, the question emerged whether overlap between the tracts and nonwhite matter regions could have introduced noise in our average DTI measures (for a visual example, see Figure 3). To answer this question, we performed tract-based spatial statistics (TBSS) in FSL (Smith et al., 2006). First, FA images were slightly eroded and end slices were zeroed in order to remove likely outliers from the diffusion tensor fitting. Second, all FA images were aligned to 1 mm standard space using nonlinear registration to the FMRIB58_FA standard-space image. Affine registrations were then used to align images into $1 \times 1 \times 1$ mm MNI 152 space, and finally skeletonized. Subsequently, the mean skeletonized FA image was thresholded at FA of 0.2. (In the online supporting information Table S1–S4, we include an additional analysis using an

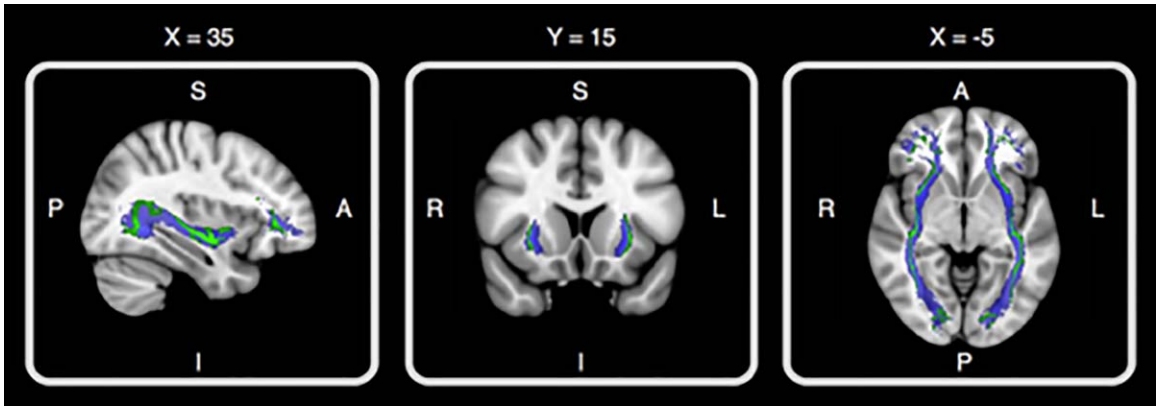

Figure 3. Bilateral IFOF in MNI space. The blue IFOF tract is wide and still shows overlap with nonwhite matter regions. The green mask represents the IFOF after applying the shrinkage operator.

Table 1. *Suggested Categories for Interpreting Bayes Factors*

Bayes factor BF_{01}			Interpretation
		100	Extreme evidence for H_0
30	–	100	Very strong evidence for H_0
10	–	30	Strong evidence for H_0
3	–	10	Moderate evidence for H_0
1	–	3	Anecdotal evidence for H_0
	1		No evidence
1/3	–	1	Anecdotal evidence for H_1
1/10	–	1/3	Moderate evidence for H_1
1/30	–	1/10	Strong evidence for H_1
1/100	–	1/30	Very strong evidence for H_1
	<	1/100	Extreme evidence for H_1

FA threshold of 0.4, which is an even more conservative threshold to only include voxels that have a relatively high FA value). Participants' FA data were then projected onto the mean skeletonized FA image and concatenated. For each participant, tracts from Session 1 were then additionally masked with corresponding tracts from the other sessions, resulting in tracts that only included voxels shared by all sessions, in addition to being skeletonized. This was done to further shrink and equalize our masks. We subsequently extracted average DWI measures from these tracts and performed an intraclass correlation (ICC) analysis augmented by Bayesian statistical tests to assess stability. We will henceforth refer to this multistep process as our shrinking operator.

ICC. The consistency between the different scan sessions was estimated using the ICC correlation as implemented in the R Package irr (version 0.84, Garner, Fellows, Lemon, & Singh, 2012). The ICC is a descriptive statistic that describes the similarity between measurements. We will adopt the labels provided by Cicchetti (1994) where a value between 0.4 and 0.59 is fair, 0.6 and 0.74 is good, and an ICC between 0.75 and 1.0 is excellent similarity between measurements.

Bayesian statistics. We performed Bayesian repeated measures analyses of variance (ANOVAs) with subject as a random factor and paired *t* tests using the BayesFactor toolbox (version 0.9.12-2; Morey, Rouder, & Jamil, 2015) in R (version 3.0.2; R Foundation for Statistical Computing, http://www.R-project.org). *T* tests were run between Session 1 and 2, and 1 and 4. ANOVAs were run over all four sessions. In all tests, the null hypothesis represented stability (i.e., no difference/change). These Bayesian *t* tests and ANOVAs are arbitrarily similar to their frequentist counterparts. In terms of their interpretation, they differ mostly in their outcome measure. The outcome of these Bayesian hypothesis tests is a single number known as the Bayes factor (Dienes, 2008; Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2013; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The Bayes factor (BF) quantifies the support that the data provide for the null hypothesis H_0 (no change) vis-à-vis the composite alternative hypothesis H_1 (change). For instance, $BF_{10} = 3$ indicates that the observed data are three times as likely to have occurred under H_1 than under H_0, and $BF_{10} = 0.2$ (or $BF_{01} = 1/0.2 = 5$) indicates that the data are five times as likely to have occurred under H_0 than under H_1. The evidential support that the BF_{01} gives to the null hypothesis can be categorized based on a set of labels proposed by Jeffreys (1961). Table 1 shows this suggested evidence categorization for the BF_{01}, edited by and taken from Boekel, Forstmann, and Wagenmakers (2016; Table 1, p. 119). We will adopt these labels to facilitate the interpretation of our Bayes factors. Nevertheless, the labels should not be zealously adhered to.

Results

We start by presenting behavioral findings that suggest we can merge the conditions in our data to investigate test-retest reliability irrespective of which task participants performed between scans. We then report the results of the reliability assessment of four tracts derived from probabilistic tractography, after which we present results from an additional analysis in which we investigate the reliability of conservative versions of our tracts in an attempt to exclude nonwhite matter sources of noise using a shrinkage operator. An additional, more conservative shrinkage operator was also applied, the results of which can be found in supporting information Table S1–S4.

Behavior

One participant did not complete the second stop-signal block and was omitted from behavioral analysis. See Table 2 for an overview of descriptive behavioral results. Below, we describe the behavioral results in detail.

Go-trial response times. To assess the between-groups behavioral differences in go-trial response time, we performed a Bayesian *t* test on the go response times of the go-task practice session (Session 0), between the stop and go active control groups. We found anecdotal evidence in favor of the null hypothesis of no difference ($BF_{01} = 2.52$).

To investigate behavioral changes over time, we performed several Bayesian ANOVAs testing for main effects of session. A Bayesian ANOVA of the go-trial response times of the go active control group in Session 0 (practice), 1, and 2 showed very strong

Table 2. *Behavioral Results Split by Groups and Sessions*

Group	Session	Go RT	Accuracy	SSRT	P(StopFail)
Stop signal group	0	364.26 (19.16)	0.96 (0.04)	–	–
	1	471.63 (65.06)	0.96 (0.04)	235.61(89.92)	0.51 (0.03)
	2	437.42(66.32)	0.94 (0.07)	215.39(66.48)	0.50 (0.03)
	3	432.90 (82.74)	0.97 (0.03)	225.11(123.89)	0.52 (0.04)
Active control group	0	364.72(30.56)	0.96 (0.04)	–	–
	1	353.40 (30.77)	0.95 (0.03)	–	–
	2	354.72 (31.95)	0.93 (0.04)	–	–

Note. Session 0 is the go practice session performed by participants to familiarize them with the directional decision component of the task. Standard deviations appear in parentheses. P(StopFail) represents the probability of stop trials in which subjects responded.

evidence in favor of the presence of a main effect of session ($BF_{01} = 0.03$). A similar ANOVA on go-trial response times from Session 1, 2, and 3 of the stop group showed anecdotal evidence in favor of the absence of a main effect of session ($BF_{01} = 2.64$). Because the latter analysis only included the nine stop participants who completed the 2-week follow-up session, we performed an additional Bayesian *t* test (including all 18 stop participants) on the difference in go response time between Session 1 and 2. The resulting Bayes factor shows anecdotal evidence in favor of the absence of a difference in go response times between Session 1 and 2 for the stop group ($BF_{01} = 1.40$). This result reflects absence of evidence rather than evidence for absence, and as such precludes a definite conclusion in favor of either hypothesis.

Accuracy. We performed comparable analyses to the go-trial response times for the accuracy data (i.e., the proportion of responses directionally congruent with the stimulus). The directional discrimination in our stop task was intentionally made trivial, and accordingly accuracy was generally high (see Table 2).

A Bayesian *t* test on the accuracy of the practice session (go task) between the stop and go active control groups showed anecdotal evidence in favor of the null hypothesis of no difference between groups ($BF_{01} = 2.52$).

A Bayesian ANOVA of the accuracy of the go active control group in Session 0, 1, and 2 showed anecdotal evidence in favor of the null hypothesis ($BF_{01} = 1.08$). For the stop participants who completed the 2-week follow-up session, a Bayesian ANOVA on accuracy in Session 1, 2, and 3 showed moderate evidence in favor of the absence of a main effect of session ($BF_{01} = 3.61$). An additional Bayesian *t* test between the accuracy in Session 1 and 2 for the complete stop group provided anecdotal evidence in favor of the absence of a change in accuracy ($BF_{01} = 1.20$).

SSRT. Two tests were performed to investigate behavioral changes in SSRT over time and between groups. A Bayesian ANOVA of SSRT in Session 1, 2, and 3 of the nine participants who completed all three stop sessions showed moderate evidence in favor of the absence of a main effect of session ($BF_{01} = 3.37$). An additional Bayesian *t* test including the entire stop group on the difference in SSRT between Session 1 and 2 showed anecdotal evidence in favor of the absence of a difference ($BF_{01} = 1.66$).

Beyond the simple go-RT session effect, there appear to be no practice effects in our data. As such, we continue with DTI test-retest reliability analyses.

DTI

Next, we report test-retest reliability of the following DTI measures: mean FA/MD/AD/RD, tract strength, and tract volume. We computed ICCs between the first two sessions for all 34 participants, between the first and last session of a subset of 15 subjects, as well as over all sessions for this subset. We augment this ICC analysis using Bayesian *t* tests and Bayesian ANOVAs to facilitate statistical inference.

STN-IFC. We delineated a tract between STN and IFC and extracted average DTI measures for each session and participant. An overview of the results of this tract can be seen in Table 3. For the consistency between the first and second session, in our bigger sample of 34 subjects, we observe high ICCs for all DWI measures and tract strength (ICCs > 0.86). For tract volume, we find slightly lower ICCs (left: 0.77; right: 0.61). However, all but two of the Bayes factors for the associated *t* test are higher than 3 (only left FA and right tract strength show anecdotal evidence, albeit in favor of the null), suggesting that the evidence is moderately in favor of an absence of a difference between these sessions. For the consistency between the first and last session, in our smaller sample of 15 subjects, we observe reasonably high ICCs for all DTI measures and tract strength (ICCs > 0.72). For tract volume, we find a lower ICC of 0.44 in the left hemisphere (although the ICC for the right hemisphere is 0.88). Our Bayes factors for this test are generally in favor of the absence of a difference, although they suggest this only anecdotally. We suspect that this is due to the smaller sample size of this group ($n = 15$). Finally, for the overall consistency, we observe reasonably high ICC values for all measures in this tract (ICCs > 0.71). All but two (left FA: $BF_{01} = 0.55$; left RD: $BF_{01} = 2.14$) Bayes factors of the associated ANOVAs are higher than 3, suggesting that the evidence is moderately in favor of the absence of a difference over the four sessions included in this test. These Bayes factors seem high in comparison to those coming from the comparison of Session 1 and Session 4. We suspect that this is due to the ANOVA taking more data into account, since it is run over all four sessions as opposed to only two.

Table 3. *Reliability of DTI Measures in the STN-IFC Tract Thresholded at 10% of Robust Range*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	2.40	0.90	0.68	0.84	0.55	0.95
	R	3.92	0.89	3.46	0.90	9.80	0.93
MD	L	4.98	0.87	1.85	0.72	4.47	0.90
	R	4.78	0.95	1.87	0.91	6.72	0.96
AD	L	4.26	0.87	3.68	0.84	9.77	0.89
	R	5.44	0.94	2.08	0.87	5.92	0.95
RD	L	3.97	0.89	1.31	0.77	2.14	0.93
	R	4.67	0.94	2.15	0.92	7.89	0.95
VOL	L	5.17	0.77	2.80	0.44	4.44	0.71
	R	4.76	0.61	2.94	0.88	8.68	0.77
TS	L	4.89	0.89	3.70	0.86	8.05	0.87
	R	2.65	0.86	3.72	0.82	9.35	0.89

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; VOL = volume; TS = tract strength; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

STN-vmPFC. We delineated a tract between STN and vmPFC and extracted average DTI measures for each session and participant. An overview of the results of this tract can be seen in Table 4. For the consistency between the first and second session, in our bigger sample of 34 subjects, we observe reasonably high ICCs for all but one measure (ICCs > 0.69). For AD in the right hemisphere version of this tract, we find a rather low ICC of 0.33. However, all but two (left FA: $BF_{01} = 1.33$; left RD: $BF_{01} = 2.70$) Bayes factors for the associated *t* test are higher than 3, suggesting that the evidence is moderately in favor of an absence of a difference between these sessions. For the consistency between the first and last session, in our smaller sample of 15 subjects, we observe reasonably high ICCs for all DTI measures and tract strength (ICCs > 0.77). For tract volume, we find slightly lower ICCs (left: 0.60; right: 0.57). Our Bayes factors for this test are in favor of the absence of a difference, although they suggest this only anecdotally. Finally, for the overall consistency, we observe reasonably high ICC values for all measures in this tract (ICCs > 0.80), with only one exception

Table 4. *Reliability of DTI Measures in the STN-vmPFC Tract Thresholded at 10% of Robust Range*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	1.33	0.88	0.79	0.87	0.87	0.95
	R	3.42	0.85	2.86	0.90	6.07	0.88
MD	L	3.89	0.91	1.86	0.78	3.69	0.91
	R	5.01	0.69	2.05	0.86	8.24	0.86
AD	L	4.72	0.92	3.52	0.78	8.43	0.90
	R	5.43	0.33	0.81	0.77	6.55	0.63
RD	L	2.70	0.90	1.38	0.82	2.32	0.93
	R	4.49	0.80	2.97	0.88	8.84	0.90
VOL	L	4.80	0.81	3.78	0.60	1.64	0.83
	R	5.26	0.75	3.72	0.57	7.44	0.80
TS	L	5.40	0.90	3.48	0.85	10.39	0.87
	R	3.47	0.91	3.73	0.81	6.40	0.89

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; VOL = volume; TS = tract strength; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

of AD in the left hemisphere showing a slightly lowered ICC of 0.63. Bayes factors of the associated ANOVAs are all higher than 3, suggesting that the evidence is moderately (or in some cases strongly; $BF_{01} > 10$) in favor of the absence of a difference over the four sessions included in this test.

STR-pre-SMA. We delineated a tract between STR and pre-SMA and extracted average DTI measures for each session and participant. An overview of the results of this tract can be seen in Table 5. For the consistency between the first and second session, in our bigger sample of 34 subjects, we observe reasonably high ICCs for all but one measure (ICCs > 0.72), with a single slightly lower ICC of 0.65 in the AD of the left hemisphere version of this tract. All but one Bayes factor for the associated *t* test are higher than 3, suggesting that the evidence is moderately in favor of an absence of a difference between these sessions. The Bayesian *t* test for tract strength in the left hemisphere version of this tract resulted in a lower Bayes factor of 1.58, although this was still in favor of the absence of a difference. For the consistency between the first and last session, in our smaller sample of 15 subjects, we observe high ICCs for all DTI measures (ICCs > 0.78). Lower ICCs were found in the tract strength measure (left: 0.55; right: 0.53), and in terms of volume (left: 0.64; right: 0.72). Our Bayes factors for this test are in favor of the absence of a difference, although they suggest this only anecdotally. Finally, for the overall consistency, we observe reasonably high ICC values for all measures in this tract (ICCs > 0.78). All but three (right FA: $BF_{01} = 2.83$; right MD: $BF_{01} = 2.70$; right RD: $BF_{01} = 1.19$) Bayes factors of the associated ANOVAs are higher than 3, suggesting that the evidence is moderately in favor of the absence of a difference over the four sessions included in this test.

Table 5. *Reliability of DTI Measures in the STR-PreSMA Tract Thresholded at 10% of Robust Range*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	5.44	0.93	1.71	0.98	8.48	0.94
	R	5.42	0.79	3.81	0.95	2.83	0.96
MD	L	4.92	0.81	3.81	0.89	9.02	0.90
	R	5.32	0.73	3.59	0.92	2.70	0.96
AD	L	4.50	0.65	3.18	0.78	9.76	0.80
	R	5.14	0.72	3.60	0.82	9.96	0.93
RD	L	5.15	0.86	3.36	0.94	7.63	0.91
	R	5.38	0.73	3.65	0.95	1.19	0.96
VOL	L	5.44	0.88	3.69	0.64	7.16	0.81
	R	4.37	0.83	3.5	0.72	4.61	0.87
TS	L	1.58	0.75	2.52	0.55	3.88	0.78
	R	5.02	0.77	3.72	0.53	8.18	0.82

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; VOL = volume; TS: tract strength; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

Table 6. *Reliability of DTI Measures in the IFOF Thresholded at 10% of Robust Range*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	5.01	0.85	3.81	0.84	4.06	0.95
	R	2.75	0.89	2.67	0.95	6.89	0.96
MD	L	4.47	0.97	3.78	0.81	10.18	0.96
	R	4.73	0.95	0.58	0.98	2.73	0.98
AD	L	2.71	0.98	3.74	0.87	3.20	0.96
	R	5.32	0.95	1.60	0.98	2.03	0.98
RD	L	5.28	0.96	3.79	0.80	9.37	0.96
	R	3.78	0.94	0.63	0.97	3.96	0.97

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

IFOF. Finally, we investigated the reliability of DTI measures in the IFOF. We delineated the IFOF based on a registration method (see Method) to mimic the analyses of Forstmann et al. (2010). No tractography was run for this particular tract, and therefore no tract strength measures or informative tract volumes were derived. Bayesian reliability analyses were computed only including mean FA, MD, AD, and RD. An overview of the results of this tract can be seen in Table 6. For the consistency between the first and second session, in our bigger sample of 34 subjects, we observe high ICCs for all measures (ICCs > 0.85). All Bayes factors for the associated *t* test are higher than 3, suggesting that the evidence is moderately in favor of an absence of a difference between these sessions. For the consistency between the first and last session, in our smaller sample of 15 subjects, we observe high ICCs for all DTI measures (ICCs > 0.80). Our Bayes factors for this test are in favor of the absence of a difference, although they suggest this only anecdotally. Finally, for the overall consistency, we observe reasonably high ICC values for all measures in this tract (ICCs > 0.95). Bayes factors of the associated ANOVAs are all higher than 3, suggesting that the evidence is moderately (or in some cases strongly, $BF_{01} > 10$) in favor of the absence of a difference over the four sessions included in this test.

We were uncertain about our construction of the IFOF, considering the deviation in its methods compared to our construction of the other tracts. Whereas other tracts were obtained using probabilistic tractography, individual IFOF ROIs were generated by registering the template IFOF from Hua et al. (2008) in MNI space to each participant's DWI data in individual space. We hypothesized that, despite our initial 10% thresholding procedure, misregistrations could have resulted in the IFOF ROI overlapping with non-white matter tissue. To illustrate, Figure 3 depicts the bilateral

Table 7. *Reliability of DTI Measures in the STN-IFC Tract After the Shrinking Procedure*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	3.52	0.96	1.95	0.91	2.59	0.97
	R	5.26	0.95	2.66	0.92	6.97	0.96
MD	L	3.62	0.94	1.47	0.91	1.25	0.97
	R	5.35	0.94	3.33	0.93	9.45	0.94
AD	L	4.87	0.96	1.46	0.97	1.77	0.98
	R	5.19	0.95	1.49	0.95	4.90	0.94
RD	L	3.53	0.95	1.78	0.89	1.94	0.97
	R	5.43	0.94	3.81	0.92	11.06	0.95

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

IFOF masks in standard space showing, in some places, overlap with nonwhite matter. Visual inspection of individual left IFOF ROIs confirmed that this overlap was also the case at an individual level.

We decided that, more generally, additional shrinkage of all our tracts might mitigate the unwanted influence of nonwhite matter voxels, thereby further increasing reliability. We performed TBSS (see Method, "Tract-based spatial statistics"), which yields group-averaged white matter skeletons representing only the core white matter fibers. We transformed our individual tracts to the common TBSS space and skeletonized them (i.e., we masked individual tracts with the group white matter skeleton; see Figure 3 for a visualization of a skeletonized IFOF in green). In addition, for each participant, tracts from Session 1 were masked with corresponding tracts from the other sessions, resulting in tract masks that only included voxels shared by all sessions. This latter step was done to further shrink our masks and ensure comparison between only spatially overlapping voxels. The skeletonization procedure alongside the between-sessions masking represents our shrinkage operator. We extracted average DTI measures from these tracts and report on their test-retest reliability below. Furthermore, we include results from a more conservative shrinkage operator in Table S1–S4.

Reliability of DTI measures in tracts after shrinkage. Here, we report the reliability assessment after applying our shrinkage operator to tracts used in our experiment. All results can be viewed in Table 7 through (8–10).

We expected to see higher stability in our tracts because the shrinking procedure should solve the problem of overlap between our tracts and nonwhite matter tissue. It seems that, overall, the ICCs are indeed higher in these more conservative versions of our tracts (all ICCs > 0.87; but most even > 0.95). More notably, the low ICC of 0.33 in the AD measure of the right STN-vmPFC tracts has disappeared; after applying the shrinkage operator, this ICC was brought to 0.94. Bayes factors were not noticeably affected.

In sum, our data revealed that, using an initial thresholding procedure of 10%, we find convincing test-retest reliability in most measures in most tracts. Several measures, mostly tract strength and volume, showed decreased reliability, although Bayesian statistical tests still largely support the notion of stability. This stability was further enhanced by our shrinkage operator, which aimed to remove nonwhite matter voxels from our tracts.

Table 8. *Reliability of DTI Measures in the STN-vmPFC Tract After the Shrinking Procedure*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	4.36	0.96	2.53	0.88	4.49	0.96
	R	2.90	0.98	3.31	0.92	7.68	0.97
MD	L	4.04	0.95	2.02	0.90	2.13	0.96
	R	5.41	0.96	2.27	0.96	6.53	0.97
AD	L	4.51	0.96	1.85	0.96	1.82	0.98
	R	5.13	0.94	1.49	0.91	3.57	0.95
RD	L	4.05	0.95	2.15	0.87	2.86	0.96
	R	5.03	0.97	3.51	0.95	9.26	0.97

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

Table 9. *Reliability of DTI Measures in the STR-Pre-SMA Tract After the Shrinking Procedure*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	3.56	0.97	2.81	0.97	8.57	0.98
	R	4.89	0.96	3.24	0.98	3.64	0.98
MD	L	3.39	0.98	2.53	0.98	6.75	0.99
	R	4.84	0.97	3.26	0.98	8.89	0.99
AD	L	1.82	0.97	3.27	0.98	7.81	0.99
	R	2.85	0.97	3.01	0.94	5.18	0.98
RD	L	5.15	0.99	2.51	0.98	7.07	0.99
	R	5.33	0.97	3.69	0.99	7.63	0.99

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; RD: radial = diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

Discussion

We set out to assess the test-retest reliability of DTI measures in four tracts that have recently been the subject of SBB investigations. We delineated these tracts using standard probabilistic tractography and subsequently tested FA, MD, AD, RD, tract strength, and tract volumes for stability using ICCs augmented by a Bayesian statistical framework.

Table 10. *Reliability of DTI Measures in the IFOF After the Shrinking Procedure*

		t test S1–S2		*t* test S1–S4		ANOVA	
Measure	Hemisphere	BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC
FA	L	5.29	0.94	2.88	0.95	8.52	0.98
	R	1.06	0.97	1.49	0.97	3.31	0.98
MD	L	3.96	0.98	3.41	0.93	6.80	0.98
	R	4.80	0.98	3.49	0.96	9.66	0.98
AD	L	2.64	0.98	3.81	0.94	4.41	0.98
	R	5.07	0.97	2.47	0.92	5.89	0.97
RD	L	5.0	0.98	3.09	0.93	7.49	0.98
	R	3.62	0.98	3.81	0.97	10.28	0.99

Note. FA = fractional anisotropy; MD = mean diffusivity; AD = axial diffusivity; RD = radial diffusivity; ICC = intraclass correlation coefficient; BF_{01} = Bayes factors representing relative evidence in favor of the null hypothesis.

Our analyses showed general stability in our initial tracts that were thresholded at 10% of robust range. Tract strength and tract volume seemed overall to be the least stable, although Bayes factors were still largely in favor of stability. The tracts in our initial analysis were still rather large and showed overlap with gray matter. In order to restrict our analyses to the main white matter tracts, we applied a shrinking operator. The pattern of results in our more conservative tracts improved in the sense that stability (at least in terms of ICC) was generally greater after applying the shrinkage operator. We are left with an overall reliable data set, suggesting that the cognitive control tracts we tested here are stable over time, in a small ($n = 15$) as well as larger ($n = 34$) sample size, and thus can readily be correlated to (stable) behavior measures.

With regard to our smaller subset of 15 subjects, we find a notable pattern of results when comparing the first and last sessions. While ICCs show general stability, Bayes factors associated with this comparison only provide anecdotal evidence in favor of stability. We believe this to be due to the smaller sample size. Contrary to this comparison between only the first and last sessions, a comparison taking into account all four sessions (also using $N = 15$) resulted in higher Bayes factors, probably because of the increased number of sessions (and therefore data) used per subject.

Continuing on sample sizes, previous studies have shown stability of DWI measures in as little as fewer than 10 participants (Fox et al., 2012; Heiervang et al., 2006; Vollmar et al., 2010). Previous studies have also shown this stability using, compared to the present analysis, overall fewer data per participant (Buchanan et al., 2014; Vollmar et al., 2010). With our comparatively larger (although still somewhat small; see Button et al., 2013) sample size, and our greater amount of data per participant through multiple repetitions of the same DWI sequence, we provide additional evidence for the stability of DTI measures.

We demonstrate this stability specifically in tracts of the cognitive control network. In light of the preexisting body of literature, we are tempted to also make the generalized claim that DWI measures, obtained using a standard acquisition and analyses, show general stability. Further investigations into test-retest reliability of DWI measures could systematically vary parameters in the acquisition and analysis stages in order to investigate the extent to which these parameters can influence DWI stability.

Some DWI reliability studies have already investigated the impact of specific acquisition parameters on test-retest reliability (Buchanan et al., 2014; Celik, 2016; Vollmar et al., 2010; Wang et al., 2012). Parameters such as the amount of volumes acquired per diffusion direction and the amount of diffusion directions have been shown to impact test-retest reliability. Different parameters such as b values and voxel resolution might also affect test-retest reliability. Comprehensive test-retest reliability studies that systematically vary these parameters may start to elucidate the conditions under which the most reliable DWI signal can be acquired and processed.

Such comprehensive studies can be found in the fMRI literature (Bennet & Miller, 2010; Laumann et al., 2015) and could serve as templates for future DWI reliability studies and meta-analyses. At this time, extensive investigations of this kind for DWI data seem to be absent. Recent efforts in promoting transparency and data sharing could also help to increase the availability of data and subsequently facilitate large-scale investigations into DWI reliability and its relationship to acquisition parameters (Poline et al., 2012). Some examples include openfMRI (https://openfmri.org/), Open Science Framework (https://osf.io/), and the human connectome project (http://www.humanconnectomeproject.org/data/). More efforts to increase public availability of data are sure to come, and will open the door to large-scale reliability analyses.

References

Aron, A. R., Behrens, T. E., Smith, S. Frank, M. J., & Poldrack, R. A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience*, *27*(14), 3743–3752. doi: 10.1523/jneurosci.0519-07.2007

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences*, *18*(4), 177–185. doi: 10.1016/j.tics.2013.12.003

Avants, B., Duda, J. T., Kim, J., Zhang, H., Pluta, J., Gee, J. C., & Whyte, J. (2008). Multivariate analysis of structural and diffusion imaging in traumatic brain injury. *Academic Radiology*, *15*(11), 1360–1375. doi: 10.1016/j.acra.2008.07.007

Behrens, T. E. J., Woolrich, M. W., Jenkinson, M., Johansen-Berg, H., Nunes, R. G., Clare, S., … Smith, S. M. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, *50*(5), 1077–1088. doi: 10.1002/mrm.10609

Bennet, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, *1191*, 133–155. doi: 10.1111/j.1749-6632.2010.05446.x

Boekel, W., Forstmann, B. U., & Wagenmakers, E.-J. (2016). Challenges in replication brain-behavior correlations: Rejoinder to Kanai (2015) and Muhlert and Ridgway (2015). *Cortex*, *74*, 348–352. doi: 10.1016/j.cortex.2015.06.018

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behaviour correlations. *Cortex*, *66*, 115–133. doi: 10.1016/j.cortex.2014.11.019

Buchanan, C. R., Pernet, C. R., Gorgolewski, K. J., Storkey, A. J., & Bastin, M. E. (2014), Test-retest reliability of structural brain networks from diffusion MRI. *NeuroImage*, *86*(1), 231–243. doi: 10.1016/j.neuroimage.2013.09.054

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Review Neuroscience*, *14*, 365–376. doi: 10.1038/nrn3475

Celik, A. (2016). Effect of imaging parameters on the accuracy of apparent diffusion coefficient and optimization strategies. *Diagnostic and Interventional Radiology*, *22*(1), 101. doi: 10.5152/dir.2015.14440

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi: 10.1037/1040-3590.6.4.284

Coxon, J. P., van Impe, A., Wenderoth, N., & Swimmen, S. P. (2012). Aging and inhibitory control of action: Cortico-subthalamic connection strength predicts stopping performance. *Journal of Neuroscience*, *32*(24), 8401–8412. doi: 10.1523/jneurosci.6360-11.2012

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York, NY: Palgrave MacMillan.

Drayer, B., Burger, P., Darwin, R., Riederer, S., Herfkens, R., & Johnson, G. A. (1986). MRI of brain iron. *American Journal of Roentgenology*, *147*, 103–110. doi: 10.2214/ajr.147.1.103

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., … Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, *107*(36), 15916?15920. doi: 10.1073/pnas.1004932107

Forstmann, B. U., Jahfari, S., Scholte, H. S., Wolfensteller, U. van den Wildenberg, W. P. M., & Ridderinkhof, K. R. (2008). Function and structure of the right inferior frontal cortex predict individual differences in response inhibition: A model-based approach. *Journal of Neuroscience*, *28*(39), 9790–9796. doi: 10.1523/jneurosci.1465-08.2008

Forstmann, B. U., Keuken, M. C., Jahfari, S., Bazin, P.-L., Neumann, J., Schaefer, A., … Turner, R. (2012). Cortico-subthalamic white matter tract strength predicts interindividual efficacy in stopping a motor response. *NeuroImage*, *60*(1), 370–375. doi: 10.1016/j.neuroimage.2011.12.044

Fox, R. J., Sakaie, K., Lee, J.-C., Debbis, J. P., Lio, Y., Arnold, D. L., … Fisher, E. (2012). A validation study of multicenter diffusion weighted imaging: Reliability of fractional anisotropy and diffusivity values. *American Journal of Neuroradiology*, *33*, 695–700. doi: 10.3174/ajnr.A2844

Gamer, M., Fellows, J., Lemon, I. & Singh, P. (2012). *Package "irr." Various coefficients of interrater reliability and agreement* (version 0.84) [computer software package].

Heiervang, E., Behrens, T. E. J., Mackay, C. E., Robson, M. D., & Johansen-Berg, H. (2006). Between session reproducibility and between participant variability of diffusion MR and tractography measures. *NeuroImage*, *33*, 867–877. doi: 10.1016/j.neuroimage.2006.07.037

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

Jansen, J. F., Kooi, M. E., Kessels, A. G., Nicolay, K., & Backers, W. H. (2007). Reproducibility of quantitative cerebral T2 relaxometry, diffusion tensory imaging, and 1H magnetic resonance spectroscopy at 3.0 Tesla. *Investigative Radiology*, *42*(6), 327–337.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. doi: 10.1016/S1361-8415(01)00036-6

Johansen-Berg, H., Behrens, T., Robson, M. D., Drobnjak, I., Rushworth, M., Brady, J. M., … Matthews, P. M. (2004). Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(36), 13335–13340. doi: 10.1073/pnas.0403743101

Jovicich, J., Marizzoni, M., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., … The PharmaCog Consortium. (2014). Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensory imaging of healthy elderly participants. *NeuroImage*, *101*, 390–403. doi: 10.1016/j.neuroimage.2014.06.075

Kanai, R. (2015). Open questions in conducting confirmatory replication studies: Commentary on "A purely confirmatory replication study of structural brain-behaviour correlations" by Boekel et al., 2015. *Cortex*, *74*, 343–347. doi: 10.1016/j.cortex.2015.02.020

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*, 231–242. doi: 10.1038/nrn3000

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Keuken, M. C., Bazin, P.-L., Crown, L., Hootsmans, J., Laufer, A., Müller-Axt, C., … Forstmann, B. U. (2014). Quantifying inter-individual anatomical variability in the subcortex using 7T structural MRI. *NeuroImage*, *94*(1), 40–46. doi: 10.1016/j.neuroimage.2014.03.032

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., … Petersen, S. E. (2015). Functional system and areal organization of a highly sampled individual human brain. *Neuron*, *86*, 657–670. doi: 10.1016/j.neuron.2015.06.037

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge, UK: Cambridge University Press.

Madhyastha, T., Mérillat, S., Hirsiger, S., Bezzola, L., Liem, F., Grabowski, T., & Jäncke, L. (2014). Longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging. *Human Brain Mapping*, *35*(9):4544–4555. doi: 10.1002/hbm.22493

Matzke, D., Love, J., Wiecki, T. V., Brown, S. D., Logan, G. D., & Wagenmakers, E.-J. (2013). Release the BEESTS: Bayesian estimation of ex-Gaussian STop-signal reaction time distributions. *Frontiers in Psychology*, *4*. Retrieved from http://doi.org/10.3389/fpsyg.2013.00918

Morey, R. D., Rouder, J. N. & Jamil, T. (2015) Computation of Bayes factors for common designs [Computer software]. Retrieved from http://bayesfactorpcl.r-forge.r-project.org/

Muhlert, N., & Ridgway, G. R. (2016). Failed replications, contributing factors and careful interpretations: Commentary on "A purely confirmatory replication study of structural brain-behaviour correlations" by Boekel et al., 2015. *Cortex*, *74*, 338. doi:10.1016/j.cortex.2015.02.019.

Mulder, M. J., Boekel W., Ratcliff, R., & Forstmann, B. U. (2014). Cortico-subthalamic connection predicts individual differences in value-driven choice bias. *Brain, Structure and Function*, *219*, 1239–1249. doi: 10.1007/s00429-013-0561-3

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343. doi: 10.1523/jneurosci.4156-11.2012

Owen, J. P., Ziv, E., Bukshpun, P., Pojman, N., Wakahiro, M., Berman, J. I., … Mukherjee, P. (2013). Test-retest reliability of computational network measurements derived from the structural connectome of the human brain. *Brain Connectivity*, *3*(2), 160–176. doi: 10.1089/brain.2012.0121

Pfefferbaum, A., Adalsteinsson, E., & Sullivan, E. V. (2003). Replicability of diffusion tensory imaging measurements of fractional anisotropy and trace in brain. *Journal of Magnetic Resonance Imaging. 18*(4), 427–433. doi: 10.1002/jmri.10377

Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., … Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, *6*(9). doi: 10.3389/fninf.2012.00009

Rae, C. L., Hughes, L. E., Anderson, M. C., & Rowe, J. B. (2015). The prefrontal cortex achieves inhibitory control by facilitating subcortical motor pathway connectivity. *Journal of Neuroscience*, *35*(2), 786–794.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. doi: 10.1002/hbm.10062

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, K. E., … Behrens, T. E. (2006). Tract-based spatial statistics:voxelwise analysis of multi-participant diffusion data. *NeuroImage*, *31*(4), 1487–1505. doi: 10.1016/j.neuroimage.2006.02.02

Smith, S. M., Jenkinson, M., Woolrich, M. W., & Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H. … Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219. doi: 10.1016/j.neuroimage.2004.07.051

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., … Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. *NeuroImage*, *51*(4), 1384–1394. doi: 10.1016/j.neuroimage.2010.03.046

Wang, J. Y., Abdi, H., Bakhadirov, K., Diaz-Arrastia, R., & Devous, M. D. (2012). A comprehensive reliability assessment of quantitative diffusion tensory tractography. *NeuroImage*, *60*(2), 1127–1138. doi: 10.1016/j.neuroimage.2011.12.062

(Received November 29, 2015; Accepted August 10, 2016)

Supporting Information

Additional supporting information may be found in the online version of this article:

Table S1: STN-IFC

Table S2: STN-vmPFC

Table S3: STR-pre-SMA

Table S4: IFOF