On the importance of replicating research findings

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K.I.J Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op dinsdag, 10 oktober 2017, te 12:00 uur

door Wouter Eduard Boekel

geboren te Purmerend

Promotiecommissie:

Promotor:	Prof. Dr. B.U. Forstmann	Universiteit van Amsterdam
Copromotor:	Dr. M. Mittner	Universitetet i Tromsø
Overige leden:	Prof. Dr. H.L.J. van der Maas	Universiteit van Amsterdam
	Prof. Dr. H.M. Huizenga	Universiteit van Amsterdam
	Prof. Dr. K.R. Ridderinkhof	Universiteit van Amsterdam
	Prof. Dr. L.M. Bouter	Vrije Universiteit Amsterdam
	Dr. D. Matzke	Universiteit van Amsterdam
	Dr. L. van Maanen	Universiteit van Amsterdam
	Dr. M.K. van Vugt	Rijksuniversiteit Groningen

Faculteit der Maatschappij- en Gedragswetenschappen

Table of Contents

Introduction

Replication is one of the central aspects of the scientific process (Aarts, 2015). In science, we replicate research findings in order to test their reliability. Most simply, replications may answer the question: If we perform this experiment again, will we get the same result?

In cognitive neuroscience as well, replication is a powerful tool aimed at filtering out individual and other biases (e.g., bias might occur when researchers have conflicts of interests), through mutual agreement between results of similar experiments performed by independent scientists. In order to discover the generalized principles that govern the workings of the human nervous system, experiments must be repeated over and over again so that their outcome may be tested in different samples and under different conditions (Jasny et al., 2011). If empirical science does not replicate, erroneous findings (such as false positives; finding a statistically significant result when there is none) remain uncorrected. In the fields of experimental psychology and cognitive neuroscience, the suggestion has been made that replication attempts are scarce, and erroneous findings may indeed exist. Moreover, these erroneous findings would remain uncorrected as long as replication is absent (Ioannidis, 2005). Recent unsuccessful replication efforts (Galak et al., 2012; Boekel et al., 2015) have substantiated this idea, and have lead the associated fields to self-evaluate their scientific practices. This self-evaluation has occurred not only in terms of replication (Wagenmakers and Forstmann, 2014; Boekel et al., 2016), but also in terms of openness of research materials and methods and data-sharing (Wicherts 2006; 2011), statistical methodology (e.g., Cohen, 1994; Wagenmakers, 2007), and conventional but questionable research practices (John et al., 2012, Simmons et al., 2011). The general concerns arising from these investigations center around an unexpected low reliability of research findings, which have lead some researchers to suggest that the field is experiencing a ‘crisis of confidence’ (Pashler and Wagenmakers, 2012).

Some aspects of conventional scientific practice might have contributed to this crisis of confidence. Below I discuss three of these aspects in terms of their influence on reliability; Questionable research practices (QRP), the file drawer problem, and low

sample sizes. Because this thesis mainly centers around replication, I will also discuss each of these harmful aspects of conventional research in terms of their relation to replication, firstly in terms of how replication can be hampered by these aspects, and secondly in terms of how replication can aid in the cessation of these aspects.

Questionable research practices

Simmons et al., (2011) suggested that false positives might be more prevalent in the literature than the generally accepted 5% false positive rate. This false positive rate stems from the traditional frequentist alpha=0.05, denoting a 5% chance of finding an effect when the null-hypothesis is true (thus, one would expect only one in twenty findings/papers to constitute a false-positive). Simmons et al., (2011) argued that the large variety of choices and options available to the researcher - *researcher degrees of freedom* - combined with the general tendency of researchers to over-explore their data, followed by the application of certain conventional yet harmful practices (QRPs; John et al., 2012), could lead to an increase in false positives. QRPs are conventional research practices which (1) may increase the chance of finding a significant effect, or which (2) may lead researchers to unfairly report their findings, overstating their reliability. QRPs commonly obscure the truth about the acquisition and analysis of data, sometimes by omission, sometimes by misinformation, or even deceit. One example of this is failing to report all of a study's dependent measures. This is problematic for the false-positive rate because it leads to an overstating of the reliability of a research finding. Normally, when one single significant result is reported, the assumption is made that the false-positive rate of the associated statistical test is 5%. If multiple tests were run, the overall false positive rate is no longer 5% if each of these tests is taken as evidence for the researcher's hypothesis. In these instances researchers often correct their statistical result for multiple comparisons (Bennett et al, 2009). However, if instead the researchers ignore one non-significant test and include the other significant one without applying a correction for multiple comparisons, an unfair view of the results is presented. Another example of a QRP is "rounding off" a *p* value (e.g., reporting a *p* value of 0.054 as less than 0.05). This QRP inflates the 5% positive rate because it presents non-significant (i.e. $p > 0.5$) findings as significant (p

< 0.05). John et al. (2012) found that over half of all researchers reported using at least one QRP. Since QRPs lead to the misinformed idea that the presented findings are statistically significant at the $p<0.05$ level, it stands to reason that the actual false positive rate in the literature could be higher than previously expected. For a list of common QRPs, see John et al., (2012).

The reliability of replication efforts may be negatively affected by QRPs, depending on the researchers' lack of expertise or bias. If researchers who perform a replication are biased towards the confirmation of a previous finding, they are at risk of employing QRPs. Indeed, it has been shown that replications by independent researchers were less likely to be successful than replications performed by the researchers who discovered the original effect (Makel et al., 2012). While this certainly implies an effect of bias on replication succes, the direction of this bias is uncertain (i.e., both the original study and the replication might be biased, i.e., the independent researchers might be biased towards a non-existence of the effect). Because of this it is important to perform many subsequent replications while trying to avoid the use of QRPs. In addition, adversarial collaborations (Matzke et al., 2015) may help in this regard because opposing biases may be mutually regulated (e.g., a replication in which half of the data are supplied by the original authors, and half of the data are supplied by new authors who are skeptical about the original effect, may be more reliable because researchers collaborate in terms of analyses and their interpretations).
Replication may be a powerful tool with which to prevent QRPs, especially when combined with a pre-registration protocol (such as in Boekel et al., 2015; Chapter 2). The method of pre-registration prevents QRPs because it entails determining and publicly pre-registering methods and analyses plans of an experiment, prior to performing the experiment (Chambers, 2013). Any results from the pre-planned analyses are then taken as is and will not influence the acceptance or rejection of a resulting paper (thereby aiding in the cessation of the file-drawer problem). Instead, the study is judged based on its adherence to the pre-registration protocol and the sensible interpretation of the research findings. This method may be used to prevent QRPs in repli-

cations, and to strengthen the ability of replications to retroactively correct erroneous findings in the literature.

The file drawer problem

Rosenthal (1979) suggested that many non-significant findings might not be published, thereby metaphorically disappearing into the file drawer. This file drawer problem could range from near-absence to the extreme in which journals are filled with the 5% of false positive findings implied by the conventional significance cutoff, with the remaining 95% nonsignificant findings tucked away in file drawers. Under some conditions, the funnel plots of meta-analyses can be used to make inferences regarding the magnitude of the file drawer effect in a certain field. Funnel plots are scatter plots of effect size against sample size. In the absence of selective publication, one would expect stronger variability in the effect sizes for lower then for higher powered studies, resulting in the characteristic "funnel" shape of the plot. The file drawer problem can be detected in funnel plots by an asymmetry of the scatter plot arising out of the absence of high-powered low-effect size experiments (Sterne et al., 2011). However, the estimation of the true severity of the problem remains as of yet elusive.

Failed replications may in some situations be at risk of being placed into the file drawer: Bias in researchers might lead to the unfair disposal of failed replications into the file drawer. As before, the combination of replication with pre-registration can remedy this problem. In this format a pre-registration protocol is made publicly accessible, and reviewed prior to data acquisition (Chambers et al., 2013). Should no result ever be published (for instance due to technical errors), this increased transparency will at least facilitate the realization that results are missing.

Low sample sizes

Low sample sizes are problematic especially in neuroimaging (Button et al. 2013), as they decrease statistical power. With MRI scan costs up to $500 an hour (Poldrack and Gorgolewski, 2014), scanning any more than 20 participants for one hour each may not be an option for many scientists. Button et al. (2013) argue that researchers must

publish to succeed, and that in order to increase their chances of publishing many papers, they divide their research money over several underpowered experiments.

In terms of replication, an ambiguous result due to insufficient statistical power is one of the worst outcomes, because it precludes any kind of substantial conclusion either for the presence or the absence of the effect.. These situations may be avoided by performing a power analysis which determines the preferred number of participants before running the experiment (Mumford, 2012), and/or setting a Bayesian stopping rule (e.g., BF > 10) prior to the start of the experiment, which means that as data come in, the Bayesian evidence for a set of hypotheses is tracked until it reaches a pre-determined threshold, after which data acquisition is halted.
Replications such as the one described in chapter 2 (Boekel et al., 2015) reveal a general effect of attenuation of effect sizes relative to the original studies. This attenuation suggests that the original, usually underpowered studies might result in inflated effect sizes, because only the over-estimated effect sizes are high enough to cross the high statistical thresholds prescribed to low-power data. As such, replications may contribute to reliability in research findings, by increasing the sample sizes of future neuroimaging studies.

These three aspects of conventional research; QRPs, the file-drawer, and low sample sizes, may in combination damage the reliability of research findings. Journals could be filled with questionable research performed on insufficiently large sample sizes, whereas the file drawer contains the more methodologically sound, reliable null-findings. Of course, this is an extreme view, and empirical studies will have to investigate to what extent these reliability-decreasing forces permeate the scientific literature.
This thesis contains four chapters aimed at investigating issues of reliability in the cognitive neuroscience literature, specifically centered around the practice of replication.

In chapter 1 we present a replication of a novel and exciting finding that training action video games transfers to perceptual decision making, making it possible to in-

crease a person's performance on a random-dot motion task by training them on an action video game (Green and Bavelier, 2012). We could not find this effect in our replication data set and used Bayesian statistics to show that our data was more in favor of the absence of this transfer effect: Our participants showed no signs of transfer effects from action video game training to perceptual decision making performance (van Ravenzwaaij et al., 2014).

Having questioned the reliability of research findings, we continue in chapter 2 with a pre-registered confirmatory replication study of recent findings in cognitive neuroscience. Here we attempted to replicate a total of 17 effects from 5 studies which reported structural brain-behavior (SBB) correlations (Boekel et al., 2015). Our results were in favor of the null-hypothesis of the absence of an effect for 8 out of 17 effects. In total we found no reliable evidence for the presence of the hypothesized effects in any of our 17 tests. This study was the first of its kind in the sense that we replicated effects of multiple studies simultaneously, while preventing QRPs using a pre-registration protocol (Chambers, 2013).

Chapter 3 presents a rejoinder to two commentaries we received on the replication study presented in chapter 2. Several points of critique that were raised by skeptics are discussed in this chapter. We add some nuance to the interpretation of our replication findings, but still argue that the field would benefit greatly from more confirmatory replication attempts (Boekel et al., 2016).

In chapter 4 we provide an example of an analysis which can be done to identify sources of variance which might decrease reliability. We perform a test-retest reliability analysis of DWI measures in several tracts relevant for the field of cognitive control (Boekel et al., 2017).

The work presented in these four chapters represent some initial steps made to increase the reliability and replicability of research findings in the cognitive neurosciences. Some ideas and suggestions for change have sprung from this work, which will be discussed further in the discussion.

Chapter 1

Action Video Games Do Not Improve the Speed of Information Processing in Simple Perceptual Tasks

Authors

Don van Ravenzwaaij[1], Wouter Boekel[2], Birte U. Forstmann[2], Roger Ratcliff[3], and Eric-Jan Wagenmakers[2]

[1]University of Newcastle [2]University of Amsterdam [3]Ohio State University

Abstract

Previous research suggests that playing action video games improves performance on sensory, perceptual, and attentional tasks. For instance, Green, Pouget, and Bavelier (2010) used the diffusion model to decompose data from a motion detection task and estimate the contribution of several underlying psychological processes. Their analysis indicated that playing action video games leads to faster information processing, reduced response caution, and no difference in motor responding. Because perceptual learning is generally thought to be highly context-specific, this transfer from gaming is surprising and warrants corroborative evidence from a large–scale training study. We conducted two experiments in which participants practiced either an action video game or a cognitive game in five separate, supervised sessions. Prior to each session and following the last session, participants performed a perceptual discrimination task. In the second experiment we included a third condition in which no video games were played at all. Behavioral data and diffusion model parameters showed similar practice effects for the action gamers, the cognitive gamers, and the non–gamers and suggest that, in contrast to earlier reports, playing action video games does not improve the speed of information processing in simple perceptual tasks.

Keywords: Action Games, Diffusion Model, Probabilistic Inference, Moving Dots Task.

Video games are immensely popular: they are played in 67% of US households and keep the average gamer occupied for an estimated eight hours a week.[1] One of the most popular genres in video gaming is the so-called action video game. Typically, action video games revolve around violent battles in war-like situations. Although this type of video game has been developed for entertainment purposes only, a growing body of research suggests that playing these games improves performance on a wide range of perceptual and cognitive tasks (e.g., Green & Bavelier, 2012). This suggestion is surprising — the effects of perceptual learning tend to be highly context-dependent and therefore fail to generalize broadly (Fahle, 2005; but see Liu & Weinshall, 2000; Jeter, Dosher, Petrov, & Lu, 2009). The presence of transfer effects from action video game playing may have profound societal, financial, and logistic ramifications; for instance, action video game playing may moderate cognitive decline in the elderly, assist the recovery of stroke patients, or be part of special needs educational programs. These real-life ramifications mean it is incumbent on the field to critically assess the existing evidence for the benefit of action video game playing.

In a letter to Nature, Green and Bavelier (2003) discussed advantages of video game players (VGPs) over non-video game players (NVGPs) in a compatibility task, an enumeration task, a spatial attention task, and an attentional blink task. Furthermore, to rule out the possible confound of pre-existing differences between VGPs and NVGPs, Green and Bavelier (2003) conducted a training experiment in which NVGPs were required to play an action game, Medal of Honor, for one hour per day on ten consecutive days. As a control condition, another group of NVGPs played Tetris, a game requiring visuo-spatial skills instead. The authors concluded that compared to the control group, the NVGPs trained on the action game improved more on the enumeration task, the spatial attention task, and the attentional blink task.

A number of subsequent studies have examined the effect of training on action video games. For example, in a study by Li, Polat, Makous, and Bavelier (2009), playing action video games led to enhanced visual contrast sensitivity whereas playing a non-action control video game (henceforth referred to as cognitive game) did not. In addi-

[1]For these and other statistics on gaming see for instance http://www.esrb.org/about/video-game-industry-statistics.jsp.

tion, Green and Bavelier (2006) showed how training on action video games, compared to training on cognitive games, improves performance on multiple object tracking tasks.

A further study by Feng, Spence, and Pratt (2007) reported that pre–existing gender differences in spatial attention disappear after as little as ten hours of action video game training. Specifically, women benefitted more from training in action video games than did men. Participants who were trained on a cognitive game showed no such improvements.

Moreover, in a study by Schlickum, Hedman, Enochsson, Kjellin, and Fell¨ander-Tsai (2009), it was reported that training on action video games improved performance of medical students on a virtual reality surgery task. Note though that the authors also found some benefits for training on a cognitive game.

Finally, video games also appear to benefit the elderly. For instance, a study by Drew and Waters (1986) showed that playing arcade video games improved manual dexterity, eye–hand coordination, RTs, and other perceptual–motor skills among residents of an apartment house for elderly citizens. Clark, Lanphear, and Riddick (1987) found that playing video games improved performance of elderly adults in a two–choice stimulus–response compatibility paradigm.

In sum, video gaming seems to improve performance on a range of different tasks. In an attempt to pinpoint the locus of the improvement, Green et al. (2010) conducted a study in which they compared performance of VGPs and NVGPs with the help of two mathematical decision making models: the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) and a neural decision maker model (Beck et al., 2008). The authors found that VGPs outperformed NVGPs on a visual motion discrimination task and an auditory discrimination task. Furthermore, the authors concluded that the advantage of VGPs over NVGPs is caused by a higher rate of information processing, whereas VGPs had a lower response caution than NVGPs, and motor processing was unaffected.

At first glance, the evidence in favor of action gaming benefits seems compelling. However, Boot, Blakely, and Simons (2011) warn against confounds and pitfalls that most of the above studies fall prey to. Such pitfalls include overt recruiting (creating differing demand characteristics), unspecified recruiting methods, no tests of per-

ceived similarity between tasks and games, and possible differential placebo effects. All studies without a training regimen furthermore suffer from the confound of possible pre-existing differences between VGPs and NVGPs in aptitude on perceptual learning tasks. The studies that did use a training paradigm often feature only two measurement occasions (i.e., prior to training and after all training), a coarse design in which any impact of video game playing on performance cannot be traced over time as it develops. Finally, many training studies do not supervise game-play, making it near impossible to confirm the extent to which participants have fulfilled their training requirements.

Aside from these issues, a number of studies fail to find any reliable benefit of playing video games (Boot, Kramer, Simons, Fabiani, & Gratton, 2008; Irons, Remington, & McLean, 2011; Murphy & Spencer, 2009). Perhaps most importantly, a recently published meta-analysis by Powers, Brooks, Aldrich, Palladino, and Alfieri (in press) provides a summary of effect sizes by game type. They report Cohen's ds for the benefit of a range of performance measures for the following game types: action/violent = 0.22 [0.13 – 0.30], mimetic = 0.95 [0.66 – 1.23], non-action = 0.52 [0.31 – 0.73], and puzzle = 0.30 [0.16 – 0.45]. Thus, according to the meta-analysis, effect sizes for action video games are smaller than for any other type of video game. It appears, therefore, that the final verdict on the benefit of action video game playing is still pending.

If action video games are to be used as a training method to improve perceptual and cognitive abilities, we need to be certain that they result in a tangible benefit; after all, we do not want to force grandmothers, stroke patients, and children with autism to spend their time shooting up aliens for nothing. Thus, the purpose of this study is to investigate the two claims made by Green et al. (2010): Does action video game playing improve performance on perceptual tasks? And if so, does this benefit reside in a higher rate of information processing?

In two experiments, we addressed these fundamental issues by administering a training design to two or three groups of randomly assigned participants. Two of the groups played video games under supervision; one group trained on an action video game and one group trained on a cognitive game. In the second experiment, an additional third

group served as a control. During training, we repeatedly measured performance on a perceptual discrimination task. We analyze the behavioral data of this task but also examine the data through the lens of the diffusion model (Ratcliff, 1978). In the next section, we first introduce the diffusion model. Then, we will discuss Experiments 1 and 2 and conclude by discussing the general ramifications of our results.

The Diffusion Model

In the diffusion model for speeded two–choice tasks (Ratcliff, 1978; Wagenmakers, 2009; van Ravenzwaaij & Oberauer, 2009), stimulus processing is conceptualized as the accumulation of noisy information over time. A response is initiated when the accumulated evidence reaches a predefined threshold (Figure 1).

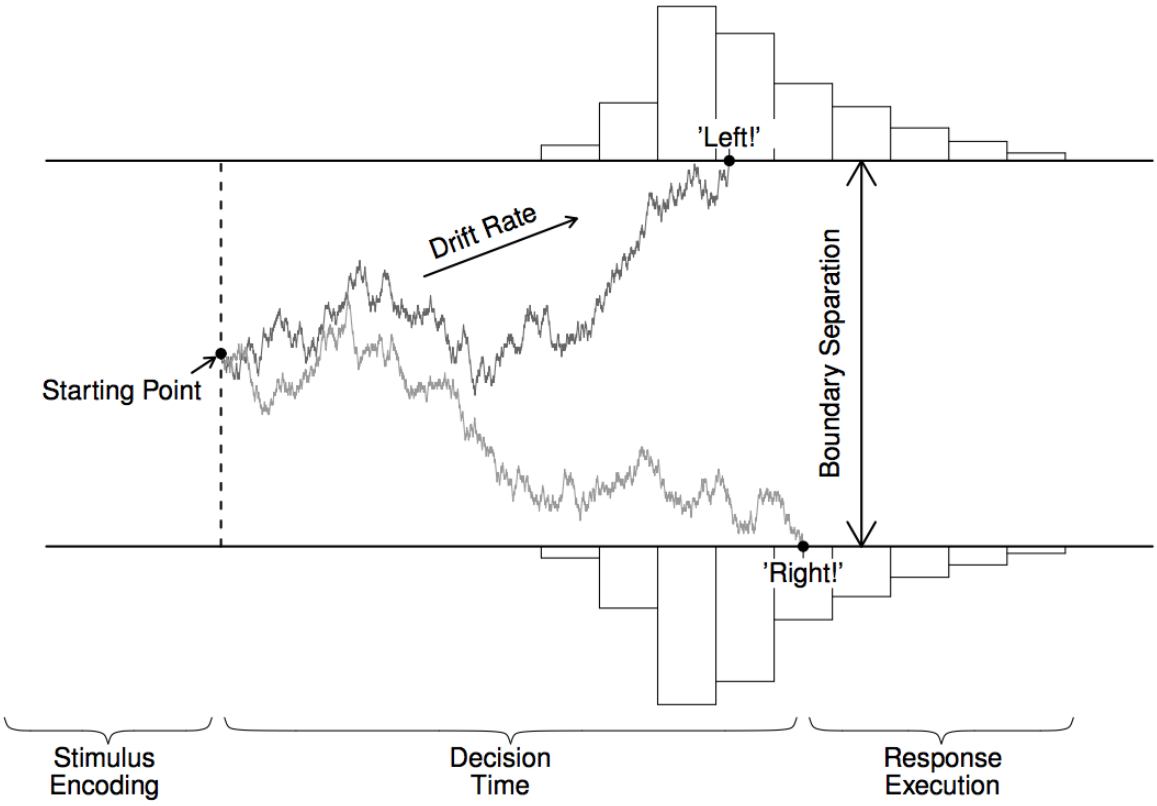


Figure 1. The diffusion model and its parameters as applied to the moving dots task. Evidence accumulation begins at z, proceeds over time guided by drift rate v, and halts whenever the upper or the lower boundary is reached. Boundary separation a quantifies response caution. Observed RT is an additive combination of the time during which evidence is accumulated and non–decision time T_{er}.

The diffusion model assumes that the decision process starts at z, after which information is accumulated with a signal–to–noise ratio that is governed by drift rate ξ, normally distributed over trials with mean v and standard deviation η.[2] Values of ξ near zero produce long RTs and high error rates. Boundary separation a determines the speed–accuracy tradeoff; lowering a leads to faster RTs at the cost of a higher error rate. Together, these parameters generate a distribution of decision times DT. The observed RT, how-

[2]Mathematically, the change in evidence X is described by a stochastic differential equation $dX(t) = \xi \cdot dt + s \cdot dW(t)$, where $s \cdot dW(t)$ represents the Wiener noise process with mean 0 and variance $s^2 \cdot dt$. Parameter s is a scaling parameter and is usually set to 0.1.

ever, also consists of stimulus–nonspecific components such as response preparation and motor execution, which together make up non–decision time T_{er}. The model assumes that T_{er} simply shifts the distribution of DT, such that RT = DT + T_{er} (Luce, 1986). The model specification is completed by including parameters that specify across–trial range in starting point, s_z, and non–decision time, s_t (Ratcliff & Tuerlinckx, 2002). Hence, the four key components of the diffusion model are (1) the speed of information processing, quantified by mean drift rate v; (2) response caution, quantified by boundary separation a; (3) a priori bias, quantified by starting point z; and (4) mean non–decision time, quantified by T_{er}.

The diffusion model has been applied to a wide range of experimental paradigms, including perceptual discrimination, letter identification, lexical decision, recognition memory, and signal detection (e.g., Ratcliff, 1978; Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2006; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; van Ravenzwaaij, van der Maas, & Wagenmakers, 2011; Ratcliff, Thapar, & McKoon, 2010). Recently, the diffusion model has also been applied in clinical settings featuring sleep deprivation (Ratcliff & van Dongen, 2009), anxiety (White, Ratcliff, Vasey, & McKoon, 2010), and hypoglycemia (Geddes et al., 2010). The model has also been extensively applied in the neurosciences (Ratcliff, Hasegawa, Hasegawa, Smith, & Segraves, 2007; Philiastides, Ratcliff, & Sajda, 2006; Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012). The advantages of a diffusion model analysis are twofold. First, the model takes into account entire RT distributions, both for correct and incorrect responses. This contrasts with a traditional analysis that considers only the mean RT for correct responses, and perhaps error rate, but ignores entirely the shape of the RT distributions and the speed of error responses. Second, the model allows researchers to decompose observed RTs and error rates into latent psychological processes such as processing speed and response caution. In the traditional analysis no attempt is made to explain the observed data by means of a psychologically plausible process model.

In their study on the effects of action video game playing on RT tasks, Green et al. (2010) report an increase in drift rate and a decrease in boundary separation for participants who practiced action games relative to participants who practiced cognitive

games. Below we report two experiments that feature supervised video game play, repeated measurements of perceptual task performance, and a diffusion model decomposition of the data.

Experiment 1

As a first test of the results of Green et al. (2010), we set out to compare NVGPs in an action game condition versus NVGPs in a cognitive game condition. Both groups were trained on video games for ten hours, divided equally over five separate sessions. Prior to every two hours of gaming and in a sixth and final behavioral session, participants performed a perceptual discrimination task. See Figure 2 for a graphical depiction of the events in all six sessions.

Participants

Twenty students from the University of Amsterdam (18 women, 2 men), aged 18 to 25 (mean = 20.6, SD = 2.4), participated on six separate days in exchange for course credit or a monetary reward of 112 euros. Participants were screened for gaming experience. In order to qualify for participation, students could not play video games for more than two hours per week on average at the present and for no more than five hours per week on average during any time in their life. Participants were randomly assigned to either the “action” or the “cognitive” condition, under the restriction that each condition contain ten participants.

Materials

Video Games. In the action condition, participants played *Unreal Tournament 2004*. This game is a so–called *first–person shooter*. The aim of the protagonist is to navigate a three–dimensional world using keyboard and mouse, shooting scores of enemies and avoiding to be shot himself. This game requires response speed, anticipation, and planning. In the cognitive condition, participants played *The Sims 2*. This game is a *strategy*

game. The aim of the protagonist is to live a virtual life and manage its basic "requirements" such as fulfilling social obligations, obtaining food, and keeping a job. This game does not rely on response speed, or at least not to the extent that *Unreal Tournament 2004* does.

Moving Dots Task. Each of the six sessions featured a moving dots task (Ball & Sekuler, 1982; Newsome & Par´e, 1988; Britten, Shadlen, Newsome, & Movshon, 1992) with two blocks of 200 trials each. On every trial, the stimulus consisted of 120 dots, 40 of which moved coherently and 80 of which moved randomly. After each 50–ms frame, the 40 coherently moving dots moved 1 pixel in the target direction. The other 80 dots were relocated randomly. On the subsequent frame, each dot might switch roles, with the constraint that there were always 40 dots moving coherently between a given set of frames. The moving dots stimulus gives the impression that the cloud of dots is systematically moving or turning in one direction, even though the cloud remains centered on the screen. Each dot consisted of 3 by 3 pixels, and the entire cloud of dots had a diameter of 250 pixels. Dots were randomly distributed over this pixel range. Participants indicated their response by pressing one of two buttons on an external device with their left or right index finger.

Immediately prior to each stimulus, a fixation cross was displayed for a random interval of 500, 800, 1000 or 1200 ms. Participants had 1500 ms to view the stimulus and give a response. The stimulus disappeared as soon as a response was made. If, for a given trial, the participant's response was slower than 1000 ms, participants saw the message "te langzaam" (too slow) at the end of the trial. In contrast to Experiment 2, coherence level was not calibrated on an individual basis.

Procedure

Upon entering, if it was the participant's first session, he or she received a general instruction about the procedure and signed an informed consent form. The participant

then completed the moving dots task and a lexical decision task.[3] The moving dots task and the lexical decision task each took approximately 20 minutes to complete. Then, in all but the sixth and final session, the participant played the video game (action or cognitive, depending on the condition) for an hour. An experimenter was in the room while the participants played the action video games, had the computer screen in sight at all times, and was available for assistance in the unlikely event a participant got stuck (this rarely happened). Following the first hour of gaming, the participant had a break of 10 to 15 minutes, after which he or she played the video game for a second hour, completing the session. The first, second, third, fourth, and fifth session all took approximately three hours each. For the sixth session, the participant first completed the moving dots task and the lexical decision task, and then filled out a payment form. The sixth session took approximately one hour. In total, all sessions took approximately 16 hours per participant, 10 hours of which were spent gaming.

[3] For consistency with Experiment 2, only the results for the moving dots task are reported. The results for the lexical decision task may be found in the online appendix, available at http://www.donvanravenzwaaij.com/Papers.html. Compared to the cognitive game, playing the action video game did not lead to improved performance on the lexical decision task.

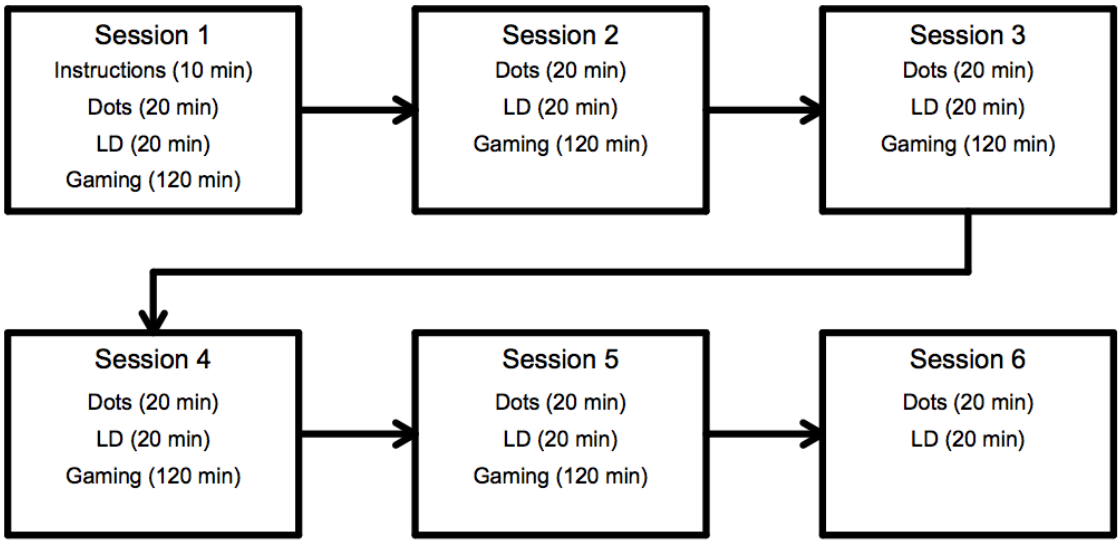


Figure 2. A flowchart of the design for Experiment 1. Dots = Moving Dots Task, LD = Lexical Decision Task. Supervised sessions took place on different days, spanning at most seven days.

Results

The next three subsections present the behavioral results (mean RT and accuracy), the diffusion modeling decomposition, and the diffusion model fit.[4] For the remainder of the statistical analyses, we report not only conventional *p*-values but also Bayes factors (e.g., Jeffreys, 1961; Kass & Raftery, 1995; Hoijtink, Klugkist, & Boelen, 2008). Bayes factors represent "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378); in contrast to *p*-values, Bayes factors allow researchers to quantify evidence in favor of the null hypothesis vis-a-vis the alternative hypothesis. For instance, when the Bayes factor BF_{01} = 10 the observed data are 10 times more likely to have occurred under H_0 than under H_1. When BF_{01} = 1/5 = 0.20 the observed data are 5 times more likely to have occurred under H_1 than under H_0. In the following, Bayes factors for analysis of variance are based on the BIC approximation (e.g., Wagenmakers, 2007; Masson, 2011), and Bayes factors for *t*-tests are based on the default Bayesian *t*-test proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009).

Behavioral Results

[4] Data from both experiments is available at http://www.donvanravenzwaaij.com/Papers.html.

One participant withdrew from the experiment after the first session and was replaced. For each participant, we excluded all RTs below 275 ms, as these were likely to be guesses. This led to the exclusion of 0.1% of all RTs and did not affect the results.

Figure 3 shows the within-subject effects for mean RT and accuracy. Across conditions, participants' mean RTs shortened in subsequent sessions, as confirmed by the presence of a negative linear trend over sessions on mean RT ($F(1, 96) = 53.0$, $p < .001$, $BF_{01} = 2.8 \cdot 10^{-9}$). Thus, practice on the moving dots task resulted in faster responding. Importantly, this session effect for mean RT did not interact with gaming condition ($F(1, 96) = 0.26$, $p > .05$, $BF_{01} = 8.75$). From the first to the last session, the overall speedup in mean RT was 51 milliseconds for the action condition and 59 milliseconds for the cognitive condition.

In addition to speeding up, participants also made more mistakes in subsequent sessions across conditions (linear trend: $F(1, 96) = 6.6$, $p < .05$, $BF_{01} = 0.37$). There was no evidence for an interaction between session and gaming condition for accuracy ($F(1, 96) = 1.8$, $p > .05$, $BF_{01} = 3.86$).

In sum, practice on the moving dots task decreased mean RT for both the action and the cognitive condition. Playing the action video game did not result in better performance compared to playing the cognitive video game. Response accuracy decreased slightly over sessions, hinting at the possibility that participants became less cautious as they improved with practice (see also Dutilh, Wagenmakers, Vandekerckhove, & Tuerlinckx, 2009). In order to quantify the psychological factors that drive the observed effects we now turn to a diffusion model decomposition.

Diffusion Model Decomposition

The diffusion model was fit to the data using the DMAT software package (Vandekerckhove & Tuerlinckx, 2007), which minimizes a negative multinomial log- likelihood function. Each participant was fit separately. We fixed starting point z to be half of boundary separation a, as there was no reason to expect a bias for either the left or right direction in the moving dots stimuli. We estimated a separate mean drift rate v, boundary separation a, and non-decision time T_{er} for each session. Furthermore, we constrained the standard deviation of drift rate η, range of starting point s_z, and range of non-decision time s_t to be equal across sessions.

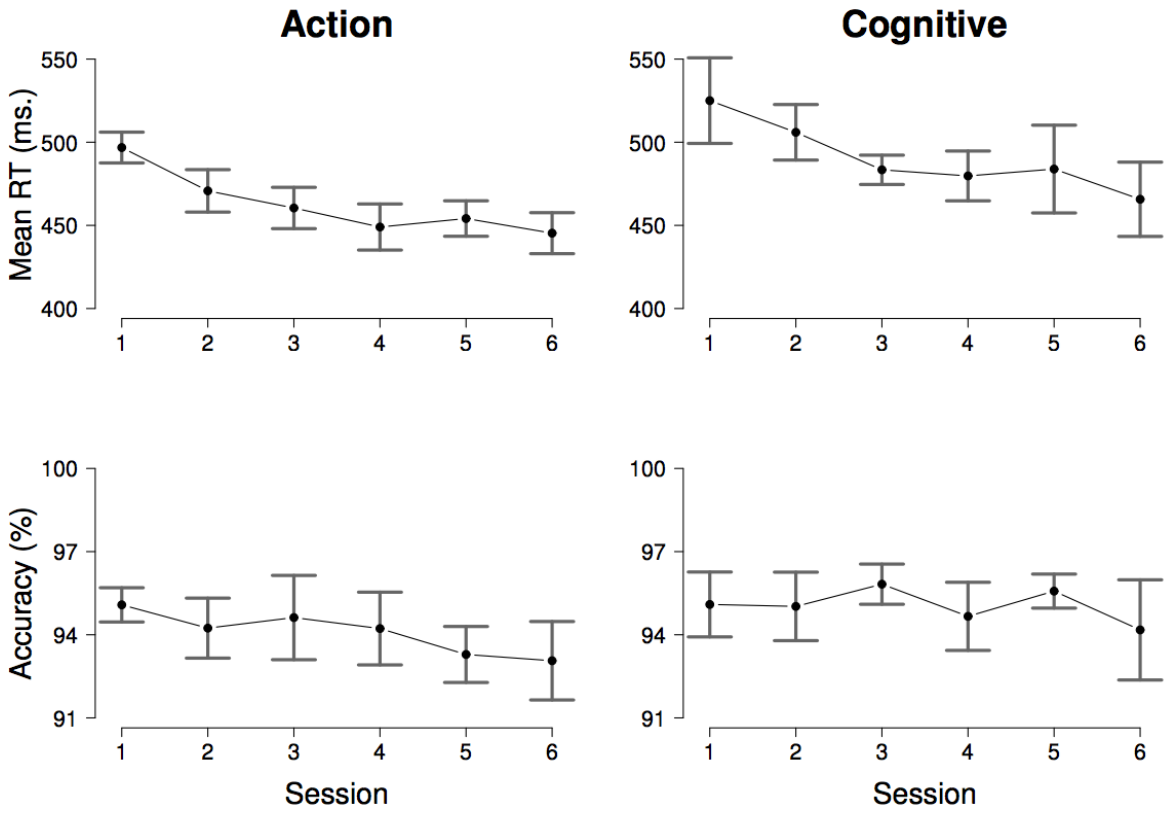


Figure 3. The within-subject effects of the action condition (top panels) and the cognitive condition (bottom panels) on mean RT (left panels) and response accuracy (right panels) for the moving dots task from Experiment 1. Error bars represent 95% confidence intervals.

The diffusion model captured the error rates okay for the cognitive condition and somewhat poorly for the action condition. The RTs were captured well on average.[5] Figure 4 shows the within-subject effects for drift rate v, boundary separation a, and non-decision time T_{er}. Across conditions, participants processed information faster in subsequent sessions, but the effect leveled off for later sessions; this visual impression is confirmed by the presence of a positive linear trend over sessions for drift rate v ($F(1, 96) = 24.7$, $p < .001$, $BF_{01} = 1.1 \cdot 10{-}4$). From the first to the last session, the overall training effect on drift rate v was 0.13 for the action condition and 0.23 for the cognitive condition. Importantly, there was no interaction between session and gaming condition for drift rate ($F(1, 96) = 0.77$, $p > .05$, $BF_{01} = 6.71$).

[5] For details, see Figure 3 of the online appendix

For boundary separation a, there was no evidence for a linear trend over sessions across conditions ($F(1, 96) = 3.5$, $p > .05$, $BF_{01} = 1.71$) and there was no evidence for an interaction between session and gaming condition ($F(1, 96) = 0.03$, $p > .05$, $BF_{01} = 9.82$). For non-decision time T_{er} there was also no evidence for the presence of a linear trend over sessions across conditions ($F(1, 96) = 0.64$, $p > .05$, $BF_{01} = 7.16$) and no evidence for an interaction between session and gaming condition ($F(1, 96) = 0.09$, $p > .05$, $BF_{01} = 9.59$).

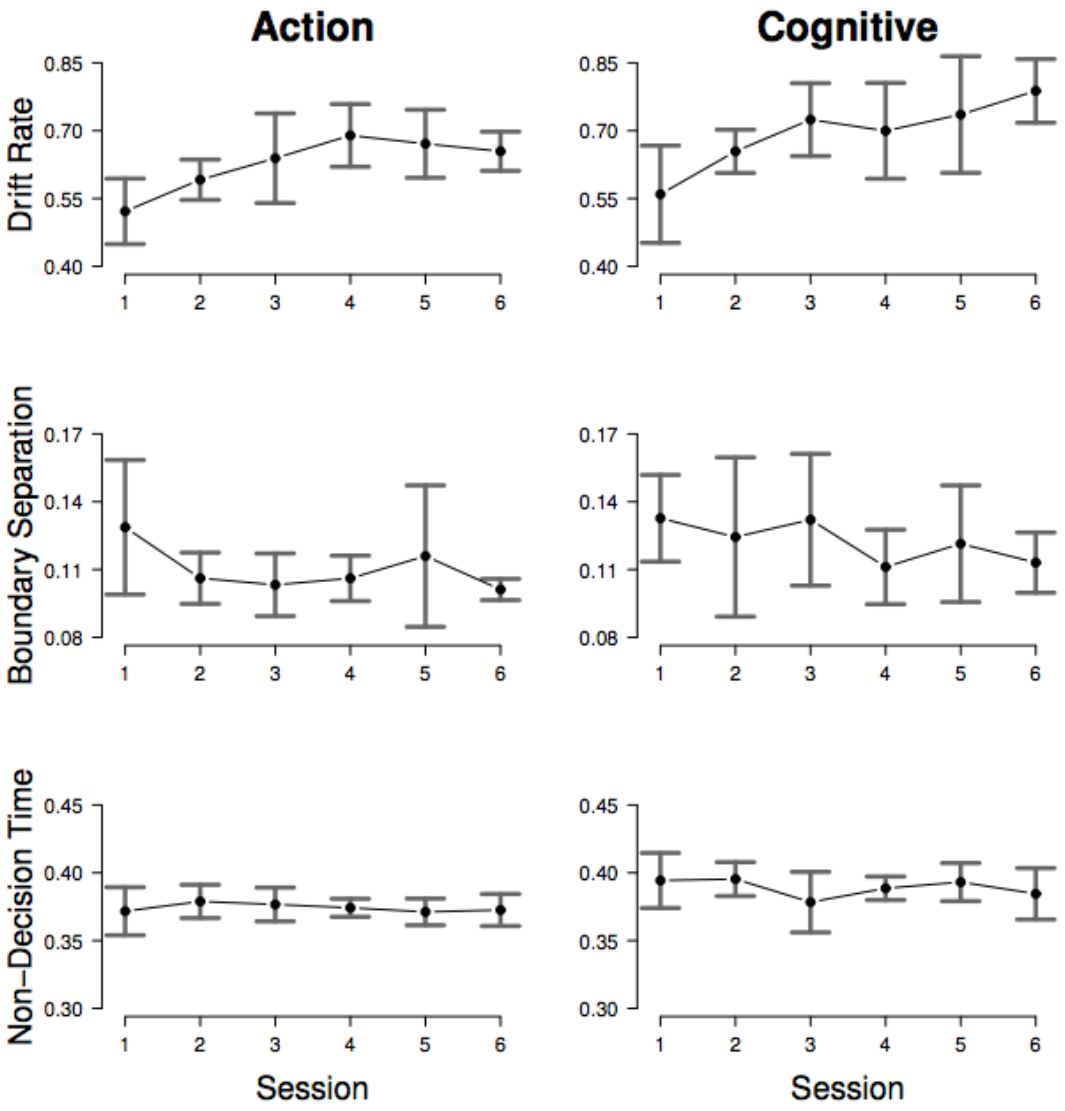


Figure 4. The within-subject effects of the action condition (top panels) and the cognitive condition (bottom panels) on drift rate *v* (left panels), boundary separation *a* (middle panels), and non- decision time T_{er} (right panels) for the moving dots task from Experiment 1. Error bars represent 95% confidence intervals.

In sum, practice on the moving dots task increased the rate of information processing. Response caution and non-decision time were unaffected. The practice-induced increase in the rate of information processing was unaffected by the type of game played. Hence, contrary to

the results by Green et al. (2010), playing action video games did not yield an increased benefit on information processing.[6]

Interim Conclusion

Experiment 1 showed no benefit of action video game playing, neither in the behavioral data nor in the diffusion model parameters. Hence, the results from Experiment 1 are at odds with the findings from Green et al. (2010). However, one may argue that we failed to find the effect because ten hours of game-play training are insufficient to elicit a reliable effect. Of course, earlier training studies also used ten hours of game-play (e.g., Green & Bavelier, 2003; Feng et al., 2007, see Powers et al., in press for a review). Moreover, Experiment 1 showed not even a hint of an effect, making its hypothetical appearance after additional training hours somewhat implausible. Nevertheless, we decided to conduct a second experiment in which we doubled the number of hours spent gaming, for both the action and the cognitive game condition. To verify that participants actually improved in the action video game, we monitored their skill level. As an additional safeguard, we also included a no-gaming condition which served as a baseline for both the action and the cognitive game condition. We also calibrated the moving dots task to each individual to produce response accuracies that were not at ceiling. In addition, we tested 45 participants instead of 20 and we increased the number of moving dots trials in each session from 400 in Experiment 1 to 1000 in Experiment 2. Finally, in an attempt to reduce the impact of potential confounds due to differential expectations (e.g., Boot et al., 2011), we told participants a believable cover story as to the goal of the experiment.

Experiment 2

As a second test of the results of Green et al. (2010), we set out to compare NVGPs in an action game condition to NVGPs in a cognitive game condition and to NVGPs in a no- gaming condition. The action and cognitive groups were trained on video games for twenty hours, divided equally over five separate sessions. Prior to every four hours of gaming and in a sixth

[6] The conclusions from the diffusion model parameters can only be relied upon, however, when the model provides a satisfactory fit to the data. In order to reassure the reader that the diffusion model gives a good description of the data, we present model predictives in the online appendix.

and final behavioral session, participants performed a perceptual discrimination task. In the first session, we calibrated the coherence level of the moving dots task to obtain comparable and off-ceiling response accuracies for each participant. On two separate days, one prior to the first session and one after the final session, participants underwent a diffusion tensor imaging scan. Both of these scans were compared to examine improvements with practice on the moving dots task. The results of these scans are unrelated to the video gaming and will be reported elsewhere. See Figure 5 for a graphical depiction of the events in all six sessions.

Participants

Forty–five students from the University of Amsterdam (19 women, 24 men), aged 17 to 24 (mean = 20, SD = 1.8), participated on six separate days in exchange for course credit and a monetary reward of 63 euros. Participants were screened for gaming experience. In order to qualify for participation, students could not play video games for more than one hour per week on average at the present and for no more than ten hours per week on average during any time in their life. Participants were randomly assigned to either the "action", the "cognitive", or the "control" condition, under the restriction that each condition contained fifteen participants. As pointed out by Boot et al. (2011), participants in the action condition may expect to improve on the experimental tasks, whereas participants in the cognitive condition may not. This confound has the potential to create spurious benefits from action video–gaming, and in order to attenuate its influence we (falsely) informed participants that they participated in an experiment that examined the effect of a perceptual task on their performance in video game playing.

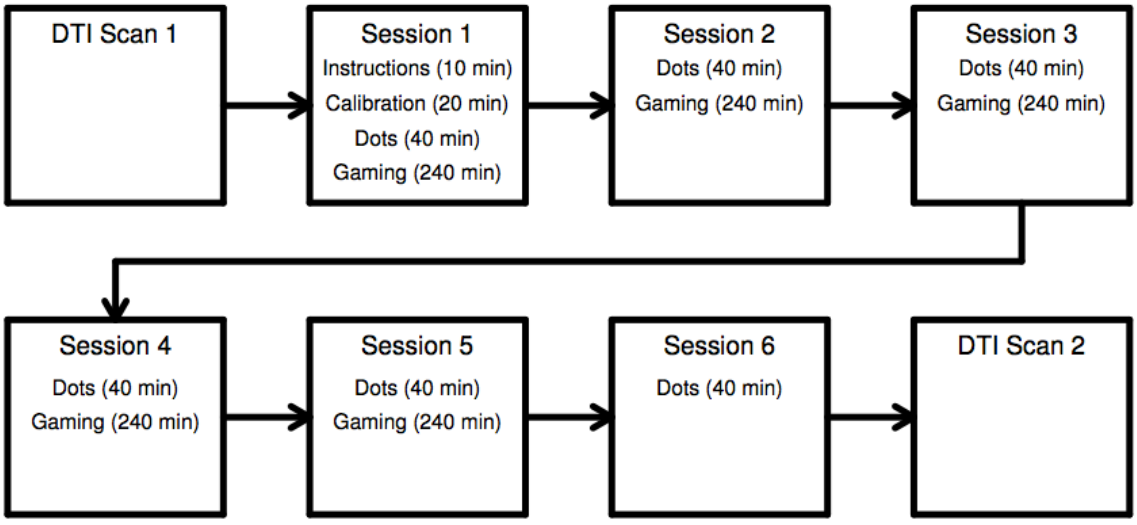


Figure 5. A flowchart of the design for Experiment 2. Calibration = Practice and calibration block for the moving dots task, Dots = Moving dots task. Supervised sessions took place on different days, spanning exactly seven days. In the control condition, participants terminated their session upon completion of the moving dots task.

Materials

Video Games. See the section "Video Games" in Experiment 1 for details.

Moving Dots Task. For Experiment 2 we modified the moving dots task used in Experiment 1. Specifically, we calibrated task difficulty (i.e., coherence level) and increased the number of trials. For calibration purposes, each participant performed a practice block of 400 trials with stimuli of varying difficulty (i.e., 0%, 10%, 20%, 40%, and 80% coherence, for 80 trials each in a randomly interleaved order). The mean RTs and accuracy data from this practice block were then fit with the Palmer diffusion model where drift rate is constrained to be proportional to the coherence level (Palmer, Huk, & Shadlen, 2005). The psychometric curve predicted by the Palmer diffusion model was then used to determine for each participant the coherence level that corresponds to 75% accuracy. This coherence level was then fixed throughout all of the experimental blocks.

Participants had 2000 ms to view the stimulus and give a response. The stimulus disappeared as soon as a response was made. If, for a given trial, the participant's response was slower

than 2000 ms, participants saw the message “No Response” at the end of the trial. If the participant’s response was faster than 200 ms, participants saw the message “Too Fast”. The fixation cross was present on screen at all times. The moving dots task took approximately 40 minutes to complete.

Procedure

The procedure in Experiment 2 differed from that of Experiment 1 in a number of ways. Firstly, the initial session commenced with a 400-trial moving dots practice block that was used for the individual calibration of task difficulty. After completing this practice block participants had a five minute break, during which the experimenter set up the main task. Secondly, the moving dots task used in sessions one through six consisted of 1000 trials during which participants had three self-paced breaks after 250, 500, and 750 trials. Thirdly, in sessions one through five, participants played the video game for four hours instead of two (except in the control condition in which each session ended as soon as the participant completed the moving dots task). Participants had three self-paced breaks after each hour of gaming. In the action condition, the difficulty level was adapted to the ability of the participant.[7] Finally, Experiment 2 did not feature a lexical decision task.

Results

There were no significant gender main effects or interactions with condition for mean RT or accuracy. In what follows, we have collapsed across gender in our presentation of the results. Figure 6 shows the improvement in difficulty level over consecutive games, averaged over participants. Note that levels range from 1, “Novice”, to 8, “Godlike”. The figure shows substantial training effects for the first 10 hours (up to approximately game number 50, depending on the individual), and a substantially reduced training effect on the action video game in the second 10 hours of gaming. This law of diminished returns is characteristic of virtually all human learning (e.g., Newell & Rosenbloom, 1981).

[7] Participants continuously played in “Deathmatch” mode. If the participant had more kills than any of the AI opponents after 10 minutes of play, the experimenter increased the difficulty level by one, and if the participant did not, the experimenter decreased the difficulty level by one.

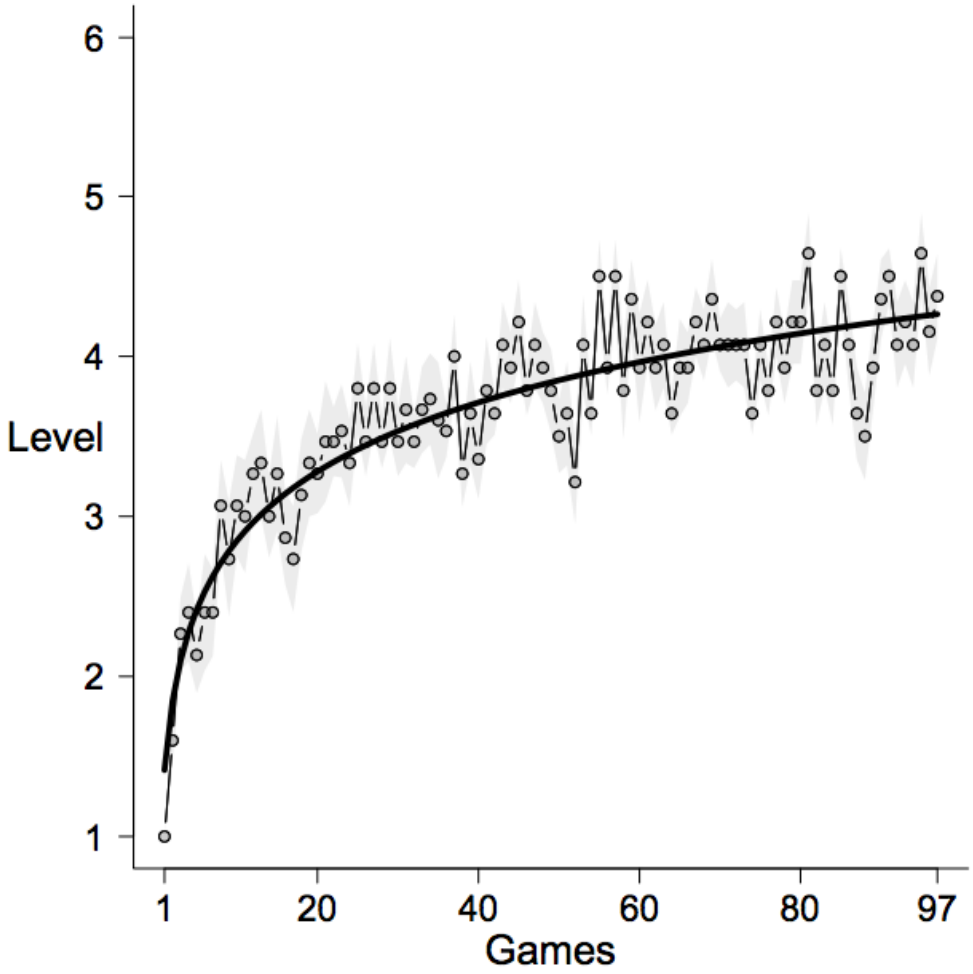


Figure 6. Average skill level of participants over consecutive games (10 minutes per game). The dark line represents the best fitting linear model through the log-transformed skill level data, with the grey area representing the standard error of the mean data points. Most participants were unable to play more than 97 games, given the time required for starting up new games, bathroom breaks, etc.

The next three subsections present the behavioral results (mean RT and accuracy), the diffusion modeling decomposition, and the diffusion model fit.

Behavioral Results

One participant from the action condition and one participant from the cognitive condition withdrew their participation. For each participant, we excluded all RTs below 275 ms, as

these were likely to be guesses. This led to the exclusion of 0.7% of all RTs and did not affect the results.

Figure 7 shows the within-subject effects for mean RT and accuracy. Across conditions, participants' mean RTs shortened in subsequent sessions, as confirmed by the presence of a negative linear trend over sessions on mean RT ($F(1, 209) = 60.5$, $p < .001$, $BF_{01} = 2.0 \cdot 10^{-11}$). Thus, practice on the moving dots task resulted in faster responding. Interestingly, this session effect for mean RT interacted with gaming condition ($F(2, 209) = 6.8$, $p < .01$, $BF_{01} = 0.24$), but not in the expected direction: as the number of sessions increases, participants speeded up both in the cognitive game condition ($t(13) = 5.0$, $p < .001$, 95% c.i. = [105.4 – 266.8], $BF_{01} = 0.01$) and in the no-game condition ($t(14)= 3.9$, $p < .01$, 95% c.i. = [52.1 – 176.5], $BF_{01} = 0.04$) whereas they did not speed up in the action game condition ($t(13) = 1.4$, $p > .05$, 95% c.i. = [-17.9 – 79.2], $BF_{01} = 2.11$). From the first to the last session, the overall session effect on mean RT was 31 milliseconds for the action condition, 186 milliseconds for the cognitive condition, and 114 milliseconds for the no-game condition.

In contrast to Experiment 1, participants in Experiment 2 became more accurate in subsequent sessions across conditions; there was a significant positive linear trend over sessions for accuracy ($F(1, 209) = 179.3$, $p < .001$, $BF_{01} = 1.8 \cdot 10-28$). This discrepancy is most likely due to the individual calibration of task difficulty. There was no evidence for an interaction between session and gaming condition for accuracy ($F(2, 209) = 0.97$, $p > .05$, $BF_{01} = 130.3$) – in fact, the Bayes factor suggests that there is strong evidence in favor of the absence of such an interaction.

In sum, practice on the moving dots task led to faster performance for the no-game condition and the cognitive game condition, but not for the action game condition. This result is at odds with that reported by Green et al. (2010). In order to quantify the psychological factors that drive the observed effects we now turn to a diffusion model decomposition.

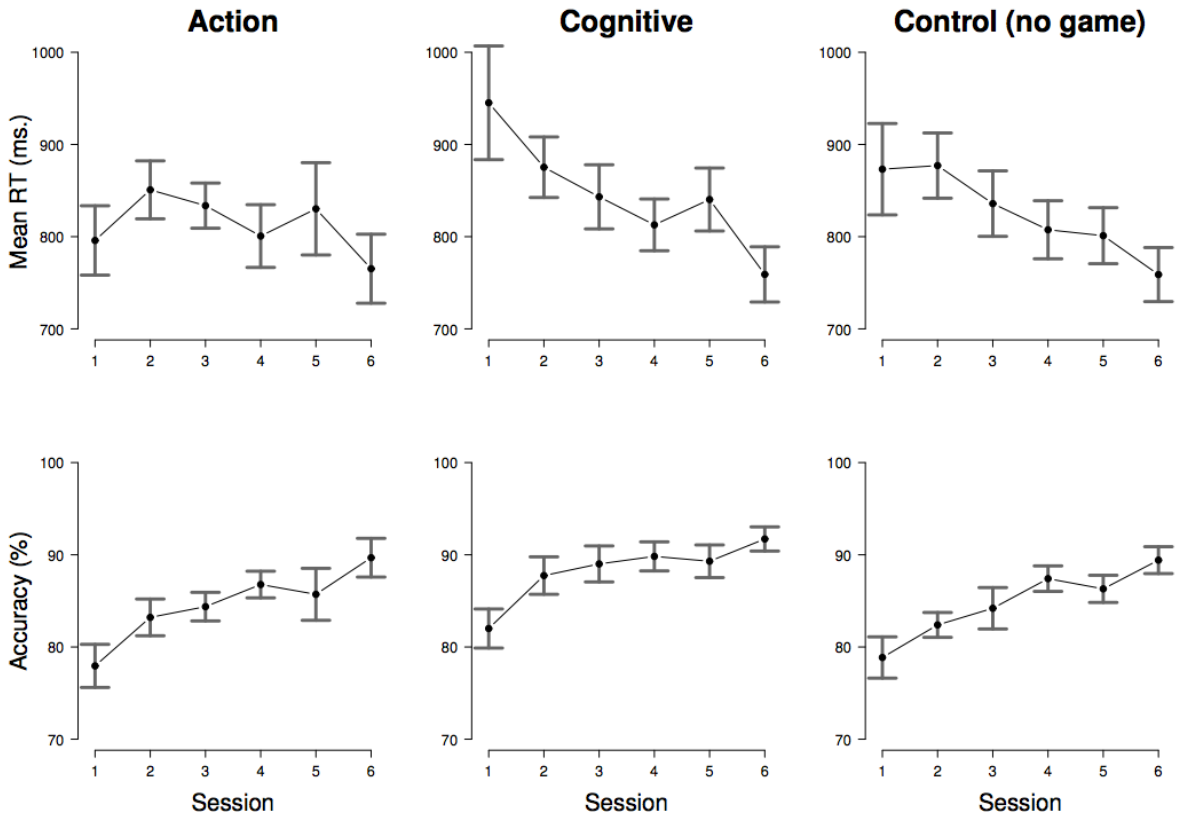


Figure 7. The within-subject effects of the action condition (left panels), the cognitive condition (middle panels), and the no-game conditions (right panels) on mean RT (top panels) and response accuracy (bottom panels) for the moving dots task from Experiment 2. Error bars represent 95% confidence intervals.

Diffusion Model Decomposition

The diffusion model analyses followed the outline provided under Experiment 1 above. The diffusion model captured the data well.[8] Figure 8 shows the within-subject effects for drift rate *v*, boundary separation *a*, and non-decision time *Ter*. Across conditions, participants processed information faster in subsequent sessions; this visual impression is confirmed by the presence of a positive linear trend over sessions for drift rate *v* ($F(1, 209) = 201.1$, $p < .001$,

[8] For the detailed model predictives, see Figure 4 of the online appendix.

BF_{01} = 5.0 · 10^{-31}).[9] For all conditions the increase in drift rate was about 0.1. Importantly, there was no interaction between session and gaming condition for drift rate (F(2, 209) = 2.2, p > .05, BF_{01} = 23.5). For boundary separation a, there was no evidence for a linear trend over sessions across conditions (F(1, 209) = 2.3, p > .05, BF_{01} = 4.50). There was a significant interaction between session and game condition (F(2, 209) = 9.3, p < .001, BF_{01} = 0.02); boundary separation a was increasing over sessions in the action condition (t(13) = 2.4, p < .05, 95% c.i. = [0.002 – 0.028], BF_{01} = 0.51), whereas boundary separation did not change in the cognitive condition (t(13) = 2.0, p > .05, 95% c.i. = [-0.001 – 0.030], BF_{01} = 0.95) nor in the no-game condition (t(14) = -1.8, p > .05, 95% c.i. = [-0.039 – 0.003], BF_{01} = 1.21). In other words, participants became more cautious in subsequent sessions for the action condition. Note that all Bayes factors for these specific comparisons suggest the evidence is ambiguous.

For non-decision time T_{er} there was no linear trend over sessions across conditions (F(1, 209) = 0.28, p > .05, BF_{01} = 12.76) and no evidence for an interaction between session and gaming condition (F(2, 209) = 1.8, p > .05, BF_{01} = 33.95).

In sum, practicing the moving dots task led to benefits in terms of mean RT for the no-game and cognitive game conditions, bot not for the action game conditions. In addition, practicing the moving dots task led to an increase in response accuracy for all three conditions. When viewed through the lens of the diffusion model, it became clear that these practice effects were caused by an increase in the speed of information processing.

The benefits of practice were no greater for participants playing action video games than for participants playing cognitive games or for participants who did not play video games at all, a statement that holds true both for the behavioral measures and for the diffusion model drift rate parameters.

[9] Note that the absolute size of the drift rates is lower than for Experiment 1, a phenomenon that reflects that the moving dots paradigm for Experiment 2 was more difficult in order to prevent ceiling effects.

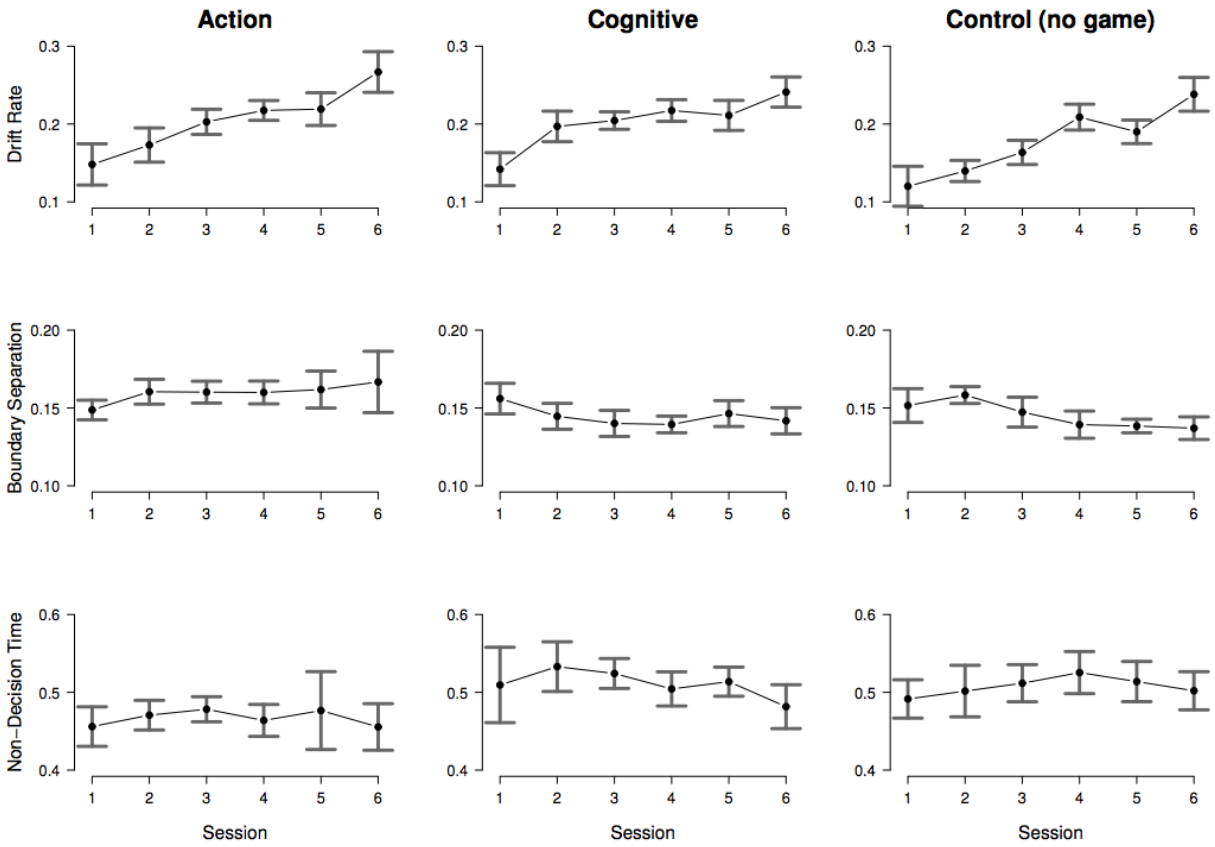


Figure 8. The within-subject effects of the action condition (left panels), the cognitive condition (middle panels), and the no-game conditions (right panels) on drift rate *v* (top panels), boundary separation *a* (middle panels), and non-decision time T_{er} (bottom panels) for the moving dots task from Experiment 2. Error bars represent 95% confidence intervals.

Conclusion

Two training experiments showed that performance on a perceptual discrimination task improves with practice. This is hardly surprising. More noteworthy is that we failed to find any performance benefit for participants who played an action video game compared to participants who played a cognitive game or no game at all. Neither the behavioral data nor a diffusion model analysis revealed even a trace of a performance increment due to playing action video games. The Bayesian analyses supported the null hypothesis of equal performance for all gaming conditions.

So why do we fail to find a gaming-specific benefit on the moving dots task? Was the amount of hours playing video games insufficient? Recall that our participants played for 10 hours in

Experiment 1 and for 20 hours in Experiment 2. The training study of Green et al. (2010), however, employed 50 hours of video game training; perhaps the difference is due to these last 30 hours?

There are multiple reasons why this alternative explanation is unlikely. First and foremost, a recent meta-analysis by Powers et al. (in press) concluded: "In true experiments, effect sizes were comparable across studies utilizing varying amounts of training (under 10 h vs. over 10 h), which suggests that learners quickly adapt cognitive processes to the design features of specific games, and may not need extensive practice to accrue training benefits." Second, as underscored by Powers et al. (in press), the number of training hours in our two experiments is average and high, respectively. Third, the critique is valid only when the effect is of a rather special nature, namely: (1) it does not manifest itself at all during the first 20 hours, and then develops quite strongly over the next 30 hours; and (2) it is somehow different from other effects of gaming, because, in general, "length of training (…) failed to moderate the effects" (Powers et al., in press).

It may be argued that the repeated testing on the moving dots task produced such a strong practice effect that it swamped any improvements due to the gaming itself. We believe this interpretation of our results to be implausible. There are no indications of a ceiling effect in the data from our second experiment. Testing six times instead of two greatly increases the power to detect a beneficial gaming effect, should it exist (see section "Power Analysis Experiment 2" of the online appendix for a demonstration). On top of that, if the power of our design had been too low, the Bayesian analyses would have indicated ambivalent evidence. However, we found clear evidence in favor of the null hypothesis.

Nevertheless, one could argue that the gaming transfer effects are so fragile and specific that they are only present when the target task has received little practice. For instance, assume that test–retest benefits are of two kinds, (a) improvement in visuo- perceptual processing (unfolding over a long time scale); and (b) improvement in pressing response buttons (unfolding over a very short time scale). It could possibly be the case that the transfer effect from gaming influences only the peripheral process (b), and that process (b) is at ceiling relatively quickly

Of course, this account is speculative, contradicts prior theorizing in which the benefit is thought to represent general facilitation of cognitive processing, and raises the question of the relevance of the transfer effect in practical application. So even if this alternative account is true, it would render the transfer effect of gaming rather uninteresting, as it may represent a temporary adjustment of a peripheral response process rather than a change in cognitive processing.

Our findings contradict the claim that action–video game playing selectively improves cognitive processing as argued by Green et al. (2010). It is possible that overt recruiting, unspecified recruiting methods, or the lack of supervised game–play has caused the results of Green et al. (2010) to be different from ours see e.g., Boot et al. (2011). However, our findings are consistent with the outcome of the recent meta–analysis by Powers et al. (in press), who concluded that "In true experiments, action/violent game training was no more effective than game training utilizing nonaction or puzzle games, but mimetic games showed large effects." Powers et al. (in press) did conclude that there is a reliable transfer effect of gaming in general, and this is inconsistent with our result from Experiment 2 – our control group improved just as much as the two groups of gamers.

In order to resolve the contradictory findings it may be essential to engage in adversarial collaborations or at least state the analysis plan in advance of data collection (Chambers, 2013; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Wolfe, 2013). Furthermore, we suggest that researchers who plan to study the effects of video game playing consider the following guidelines:

1. Despite the considerable effort required, conduct a training study on non–gamers rather than compare gamers to non–gamers. Training studies eliminate confounds due to pre-existing differences between gamers and non–gamers.
2. Design experiments with sufficient power. By testing many participants and including many trials for the psychological task at hand you increase the chances of confidently confirming or disconfirming the effects of interest.
3. Calibrate your psychological task on an individual basis to ensure error rates that are sizable and similar across participants. Training effects on response accuracy are easier to detect when accuracy is not near ceiling. A sufficient number of errors also benefits RT modeling with the diffusion model (e.g., compare the model predictives for Experiment 1 and Experiment 2).
4. Use Bayes factors to quantify evidence. Only by using Bayes factors can researchers quantify the evidence favoring the null hypothesis.
5. Use the diffusion model to decompose the behavioral effects in their underlying psychological processes such as speed of information processing, response caution, and non–decision time.

We find the claim enticing that people can boost their cognitive capacities by playing violent action video games. However, our result urge caution and suggest that before the video games find application, a series of purely confirmatory experiments is in order. We hope that our experiments will encourage such confirmatory work.

References

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science, 218*, 697–698.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J. D., et al. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60*, 1142–1152.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Boot, W. R., Blakely, D. P., & Simons, D. J. (2011). Do action video games improve perception and cognition? *Frontiers in Psychology, 2*, 1–6.

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica, 129*, 387–398.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *Journal of Neuroscience, 12*, 4745–4765.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*, 609–610.

Clark, J. E., Lanphear, A. K., & Riddick, C. C. (1987). The effects of videogame playing on the response selection processing of elderly adults. *Journal of Gerontology, 42*, 82–85.

Drew, D., & Waters, J. (1986). Video games: Utilization of a novel strategy to improve perceptual motor skills and cognitive functioning in the non-institutionalized elderly. *Cognitive Rehabilitation, 4*, 26–31.

Dutilh, G., Wagenmakers, E.-J., Vandekerckhove, J., & Tuerlinckx, F. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review, 16*, 1026–1036.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology, 15*, 154–160.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*, 850–855.

Geddes, J., Ratcliff, R., Allerhand, M., Childers, R., Wright, R. J., Frier, B. M., et al. (2010). Modeling the effects of hypoglycemia on a two–choice task in adult humans. *Neuropsychology, 24*, 652–660.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*, 534–537.

Green, C. S., & Bavelier, D. (2006). Enumeration versus multiple object tracking: The case of action video game players. *Cognition, 101*, 217–245.

Green, C. S., & Bavelier, D. (2012). Learning, attentional control, and action video games. *Current Biology, 22*, R197–R206.

Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current Biology, 20*, 1573–1579.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). Bayesian evaluation of informative hypotheses that are of practical value for social scientists. New York: Springer. Irons, J. L., Remington, R. W., & McLean, J. P. (2011). Not so fast: Rethinking the effects of action video games on attentional capacity. *Australian Journal of Psychology, 63*, 224–231.

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

Jeter, P. E., Dosher, B. A., Petrov, A., & Lu, Z.-L. (2009). Task precision at transfer determines specificity of perceptual learning. *Journal of Vision, 9*, 1–13.

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

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the implicit association test: A diffusion–model analysis. *Journal of Personality and Social Psychology, 93*, 353–368.

Li, R., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience, 12*, 549–551.

Liu, Z., & Weinshall, D. (2000). Mechanisms of generalization in perceptual learning. *Vision Research, 40*, 97–109.

Luce, R. D. (1986). Response times. New York: Oxford University Press.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null–hypothesis significance testing. *Behavior Research Methods, 43*, 679–690.

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*, 2335–2343.

Murphy, K., & Spencer, A. (2009). Playing video games does not make for better visual attention skills. *Journal of Articles Supporting the Null Hypothesis, 6*, 1–20.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Newsome, W. T., & Par´e, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *The Journal of Neuroscience, 8*, 2201–2211.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision, 5*, 376–404.

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision–making during perceptual categorization: A timing diagram. *Journal of Neuroscience, 26*, 8965–8975.

Powers, K. L., Brooks, P. J., Aldrich, N. J., Palladino, M. A., & Alfieri, L. (in press). Effects of video-game play on information processing: A meta-analytic investigation. *Psychonomic Bulletin & Review*.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). Diffusion model account of lexical decision. *Psychological Review, 111*, 159-182.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, Y. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology, 97*, 1756-1774.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873-922.

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging, practice, and perceptual tasks: A diffusion model analysis. *Psychology and Aging, 21*, 353-371.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*, 127-157.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*, 438-481.

Ratcliff, R., & van Dongen, H. P. A. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin & Review, 16*, 742-751.

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*, 225-237.

Schlickum, M. K., Hedman, L., Enochsson, L., Kjellin, A., & Fell¨ander-Tsai, L. (2009). Systematic video game training in Surgical novices improves performance in virtual reality endoscopic surgical simulators: A prospective randomized study. *World Journal of Surgery, 33*, 2360-2367.

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53*, 463–473.

van Ravenzwaaij, D., van der Maas, H. L. J., & Wagenmakers, E.-J. (2011). Does the Name-Race Implicit Association Test Measure Racial Prejudice? *Experimental Psychology, 58*, 271–277.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14*, 1011–1026.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review, 14*, 779–804.

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology, 21*, 641– 671.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140–159.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 627–633.

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology, 54*, 39–52.

Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception, & Psychophysics. *Attention, Perception, & Psychophysics, 75*, 781–783.

Chapter 2[10]

A purely confirmatory replication study of structural brain-behavior correlations.

Authors

Wouter Boekel[1], Eric-Jan Wagenmakers[1], Luam Belay[1], Josine Verhagen[1], Scott Brown[2], Birte U. Forstmann[1]

[1]University of Amsterdam, Amsterdam, The Netherlands, [2]University of Newcastle, Australia

Abstract

A recent 'crisis of confidence' has emerged in the empirical sciences. Several studies have suggested that questionable research practices (QRPs) such as optional stopping and selective publication may be relatively widespread. These QRPs can result in a high proportion of false-positive findings, decreasing the reliability and replicability of research output. A potential solution is to register experiments prior to data acquisition and analysis. In this study we attempted to replicate studies that relate brain structure to behavior and cognition. These structural brain-behavior (SBB) correlations occasionally receive much attention in science and in the media. Given the impact of these studies, it is important to investigate their replicability. Here, we attempt to replicate five SBB correlation studies comprising a total of 17 effects. To prevent the impact of QRPs we employed a preregistered, purely confirmatory replication approach. For all but one of the 17 findings under scrutiny, confirmatory Bayesian hypothesis tests indicated evidence in favor of the null hypothesis ranging from anecdotal (Bayes factor < 3) to strong (Bayes factor > 10). In several studies, effect size estimates were substantially

[10] This chapter includes the paper published in Cortex 2015 (Boekel, W, Wagenmakers, EJ, Belay, L, Verhagen, J, Brown, S, Forstmann, BU, A purely confirmatory replication study of structural brain-behavior correlations, Cortex, 2015, 66, 115-133). Unfortunately, an error in part of the preprocessing for these data was found later, and a corrigendum is in press and will appear in Cortex (Keuken MC, Boekel WE, Wagenmakers E-J, Belay L, Verhagen J, Brown S, BU Forstmann BU, Corrigendum for: A purely confirmatory replication study of structural brain-behavior correlations). Specifically, the DWI preprocessing of replication 1: Forstmann et al. (2010), and replication 3: Xu et al. (2012) were performed incorrectly. Re-analyses showed minor adjustments of values but no changes in terms of interpretation of results.

lower than in the original studies. To our knowledge, this is the first multi-study confirmatory replication of SBB correlations. With this study, we hope to encourage other researchers to undertake similar replication attempts.

Keywords: preregistration; confirmatory; replication; brain-behavior correlations

1. Introduction

In the last few years, the need for confirmatory replication studies has become increasingly evident. Recent studies have suggested that the empirical sciences are bedeviled by the use of questionable research practices (QRPs; Simmons, Nelson, & Simonsohn, 2011; John, Loewenstein, & Prelec, 2012). These practices include, for instance, optional stopping (i.e., continuing data collection until $p<.05$) and cherry-picking (e.g., reporting only those variables, conditions, or analyses that yield the desired result). In combination with the ubiquitous file drawer problem (Rosenthal, 1979), the use of these QPRs results in a high false-positive rate, such that many significant findings may in fact be false (Ioannidis, 2005). This realization has brought about a crisis of confidence in the replicability and reliability of published research findings (Ioannidis, 2012; MacArthur, 2012; Pashler & Wagenmakers, 2012). A recent study by Button, Ioannidis, Mokrysz, Nosek, Flint et al. (2013) showed that this crisis of confidence extends to the neurosciences. The crisis of confidence can be reduced in several ways. One powerful remedy is to eliminate QRPs by preregistering experiments prior to data acquisition and analysis, resembling the standard operating procedure mandated in the case of clinical trials (De Groot, 1969; Goldacre, 2009; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Chambers, 2013; Wolfe, 2013). In this article we apply study preregistration to assess the replicability of a series of findings in cognitive neuroscience.

Research in cognitive neuroscience aims to investigate the link between brain and behavior. Recently, researchers have exploited significant advances in anatomical magnetic resonance imaging (MRI) to detect subtle differences in brain structure

associated with differences in behavioral measures (Kanai and Rees, 2011). For example, in a study that received much attention in science and the media, Kanai, Bahrami, Roylance, and Rees (2012) found that individuals with a relatively large grey matter volume in specific brain regions have more Facebook friends . Other studies have reported structural brain-behavior (SBB) correlations between properties of grey and/or white matter and behavioral measures such as choice reaction time (Tuch, Salat, Wisco, Zaleta, Hevelone et al., 2005), control over speed and accuracy in decision making (Forstmann, Anwander, Schäfer, Neumann, Brown et al., 2010), percept duration in perceptual rivalry (Kanai, Bahrami, & Rees, 2010; Kanai, Carmel, Bahrami, & Rees 2011a), components of attention (i.e., executive control and alerting; Westlye, Grydeland, Walhovd, & Fjell, 2011), response inhibition (King, Linke, Gass, Hennerici, Tost et al., 2012), metacognitive ability (i.e., the ability to evaluate one's perceptual decisions; Fleming, Weil, Nagy, Dolan, & Rees, 2010), aspects of social cognition (i.e., social network size; Bickart, Wright, Dautoff, Dickerson, & Barrett, 2011, and social influence; Campbell-Meiklejohn, Kanai, Bahrami, Bach, Dolan et al., 2012), distractibility (Kanai, Dong, Bahrami, & Rees, 2011b), political orientation (Kanai, Feilden, Firth, & Rees, 2011c), sensitivity to reward and approach motivation (Xu, Kober, Carroll, Rounsaville, Pearlson et al., 2012), moral values (Lewis, Kanai, Bates, & Rees, 2012), and empathy (Banissy, Kanai, Walsh, & Rees, 2012).

Motivated by the increase in number and prominence of SBB correlations, as well as the general uncertainty regarding the reliability of non-preregistered research findings, we attempted to replicate a subset of the above-mentioned studies in a purely confirmatory fashion. It should be noted that conceptual replications, wherein a hypothesis from the original study is tested in a different experimental paradigm, do not provide reliable evidence for or against the robustness of the respective finding. Instead, only direct replications, wherein all relevant aspects of the original study are repeated can support or oppose the original finding (Pashler and Harris, 2012).

Here, we report a preregistered, purely confirmatory replication of a subset of five SBB correlation studies selected from recent literature based on the brevity of their

behavioral data acquisition. The transparency conveyed by a confirmatory design helps to avoid common pitfalls in neuroscience (and other sciences) such as the use of nonindependent analysis (Vul, Harris, Winkielman, & Pashler, 2009), double dipping (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009), obscure data collection and analysis techniques which increase false-positive rates (Simmons et al., 2011), confirmation and hindsight bias on the part of the researcher (i.e., the tendency to confirm instead of disconfirm one's beliefs and the tendency to judge events more predictable after they have occurred, respectively; Wagenmakers et al., 2012). A strictly confirmatory framework was ensured by publishing a 'Methods and Analyses document' (M&A; http://confrepneurosci.blogspot.nl/2012/06/advanced-methods-and-analyses_26.html) online before any data were inspected or analyzed (as recommended by several researchers, e.g., De Groot, 1969; Goldacre, 2009; Wagenmakers et al., 2012; Chambers, 2013; Wolfe, 2013). This M&A document was sent to the corresponding authors of the original studies. All authors agreed to the replication attempt and the processing pipeline as outlined in the M&A document. Any analysis not outlined in the M&A document will be labeled 'exploratory' (as recommended by Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). We confined our hypotheses to the direction and location of the SBB correlations reported in the original articles. For instance, Kanai et al. 2012 reported a positive SBB correlation between grey matter density in left amygdala and the number of friends on Facebook; consequently the to-be-replicated hypothesis postulates a positive SBB correlation between the same variables in our sample. This order-restriction of the hypotheses has two benefits. First, it allowed us to use one-sided as opposed to two-sided hypothesis tests, which are more specific and statistically more powerful. Second, it allowed us to focus our analyses on specific regions in the brain, i.e., regions of interest (ROI), instead of searching the whole brain for SBB correlations. This way we circumvent the need for multiple comparisons corrections that are required in whole-brain analyses.

In order to quantify the evidence that the data provide for and against the null-hypothesis, we opted for a Bayesian hypothesis test for correlations and computed Bayes factors (BF; Jeffreys, 1961) instead of p-values (for a discussion of problems with

p-values, see Edwards, Lindman, & Savage, 1963; Wagenmakers, 2007). Note that in contrast to Bayes factors, p-values are unable to quantify support in favor of the null hypothesis; a non-significant p-value indicates no more than a “failure to reject the null hypothesis”. The replication attempts will be considered successful if the corresponding Bayes factor supports the hypothesized relationship. Accordingly, a Bayes factor that supports the null hypothesis suggests a failed replication. In addition to this preregistered analysis, exploratory analyses examine estimates of effect size. It is possible that the Bayes factor supports the null hypothesis, but the estimated effect size is nevertheless close to the original effect size. To address this concern, an additional exploratory Bayes factor analysis compares the null hypothesis to an alternative hypothesis that incorporates the knowledge obtained from the original study (cf. Verhagen and Wagenmakers, 2014). These exploratory analyses occasionally provide a more nuanced perspective on the extent to which SBB correlations can be replicated.

2. Materials and Methods

2.1. General methods

Prior to inspection of the data, a preregistration protocol was published online (http://confrepneurosci.blogspot.nl/2012/06/advanced-methods-and-analyses_26.html). This ‘Methods and Analyses’ (M&A) document described all data acquisition and analysis steps. Below we summarize the subparts of this M&A document which are applicable to the results described in this article.

2.1.1. Participants

36 undergraduate psychology students (mean age = 20.12, SD = 1.73; 18 females) with normal or corrected-to-normal vision were recruited from the participant pool of a previous 43-participant MRI study. The MRI study was recently conducted by Forstmann and Wagenmakers’ research group at the University of Amsterdam and featured extensive Diffusion Weighted Imaging (DWI) and T1-weighted imaging. Hence, the additional effort involved in replicating the five studies consisted primarily in having

participants complete a battery of behavioral tests. The experiments were approved by the local ethics committee of the University of Amsterdam. Participants received a monetary compensation for their time and effort.

2.1.2. Study selection

We aimed to perform replications of a series of recent studies reporting correlations between brain structure and behavior. A review by Kanai and Rees (2011) provided us with many topical SBB correlation findings. In addition, several other studies were selected from previous literature. Brevity of behavioral data acquisition was the main selection criterion, to ensure that we would be able to replicate many SBB correlations while minimizing total acquisition time.

2.1.3. Study exclusion

Several studies, although selected and described in the M&A document, were omitted from the final analyses based on several reasons: Kanai et al. (2011c) found an SBB correlation between political orientation and brain structure in young adults, using a simple 5-point self-report measure ranging from very liberal to very conservative. The data that we acquired to replicate this contained insufficient variability in this self-report measure, and thus we excluded this study (mean: 2.26, SD: 0.57, range: 1-3; Supplementary figure S1 shows scatterplots of these data). The other three studies (Kanai et al., 2010; Kanai et al., 2011a; Bickart et al., 2011) were excluded from final replication based on problems with the ROI masks sent by the authors of the original papers (e.g., missing masks, or masks which did not match coordinates reported in the original papers). Five studies remained for the final replication attempt.

2.1.4. General procedure

The time between MRI-scanning and behavioral testing ranged from 25 to 50 days. Prior to the behavioral test session, participants received an information brochure and signed an informed consent form. Participants were tested in individual computer booths. All instructions were shown on the computer screen or printed on top of the questionnaires. Participants began by filling out the following questionnaires: BIS/BAS

(Carver and White, 1994), social network index (Cohen, 1997) social network size questionnaire (Stileman & Bates, 2007), cognitive failures questionnaire (Broadbent, Cooper, Fitzgerald, & Parkes, 1982), political orientation questionnaire (Kanai et al., 2011c), moral foundations questionnaire (Graham, Haidt, & Nosek, 2009), and the interpersonal reactivity index (Davis, 1980). After completing the questionnaires, participants continued with the computerized tasks: Bistable SFM task (Wallach & O'Connel, 1953), random dot motion task (Britten, Shadlen, Newsome, & Movshon, 1992; Gold and Shadlen, 2007), and the attention network test (Fan, McCandliss, Sommer, Raz, & Posner, 2002). The order of both questionnaires and computer tasks was randomized across participants. The total duration of the test session was 1 hour and 30 minutes. A subset of these tasks and questionnaires (i.e., the ones connected to the five studies that were included in the final replication attempt) were analyzed.

2.1.5. MRI data acquisition

DWI and T1-weighted images were collected on a 3T Philips scanner using a 32-channel head coil. For each participant, four repetitions of a multi-slice spin echo (MS-SE), single shot DWI scan were obtained using the following parameters: TR = 7545 ms, TE = 86 ms, 60 transverse slices, 2 mm slice thickness, FOV: 224 x 224 mm2, voxel size 2 mm isotropic resolution. For each slice, 32 diffusion-weighted images (b = 1000 s/mm2) along 32 directions were acquired, along with one image without diffusion weighting (b0 image, where b = 0). In addition, a T1-weighted anatomical scan was acquired (T1 turbo field echo, 220 transverse slices of 1 mm, with a resolution of 1 mm^3, TR = 8.2 ms, TE = 3.7 ms).

2.1.6. ROI-based analysis

Our purely confirmatory approach allowed us to circumvent the multiple comparison problems present in whole-brain analyses. We extracted measures of brain structure from ROIs provided to us by the authors of the original papers. These measures were then correlated to the respective behavioral measure. This approach would not have been possible if the authors of the original authors had not provided us with the ROI

masks of their findings. We would like to thank these authors for their cooperation and openness.

2.1.7. Diffusion weighted imaging analyses

All DWI data (pre-)processing and analyses were carried out using FMRIB's Software Library (FSL, version 4.0; www.fmrib.ox.ac.uk/fsl). Per participant, all four runs of DWI were concatenated and corrected for eddy currents. Affine registration was used to register each volume to a reference volume (Jenkinson and Smith, 2001). A single image without diffusion weighting (b0; b-value = 0 s/mm2) was extracted from the concatenated data and non-brain tissue was removed using FMRIB's Brain Extraction Tool (BET; Smith, 2002) to create a brain-mask which was used in subsequent analyses.

DTIFIT (Behrens, Johansen-Berg, Woolrich, Smith, wheeler-Kingshott et al., 2003) was applied to fit a tensor model at each voxel of the data (Smith, Jenkinson, Woolrich, & Beckmann, 2004). Tract-Based Spatial Statistics (TBSS) were performed using FSL's default TBSS pipeline (Smith, Jenkinson, Johansen-Berg, Reuckert, Nichols et al., 2006; http://www.fmrib.ox.ac.uk/fsl/tbss/index.html). First, fractional anisotropy (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 non-linear registration to the FMRIB58_FA standard-space image. Affine registration was then used to align images into 1 x 1 x 1 mm MNI152 space, and a skeletonization procedure was subsequently applied to a mean FA image resulting from averaging all individual MNI-aligned images. Subsequently, the mean skeletonized FA image was thresholded at FA > 0.2 in order to accurately represent white-matter tracts. Participants FA data were then projected onto the mean skeletonized FA image and concatenated. In addition to using FA images, we repeated this processing pipeline for mean diffusivity (MD) and parallel eigenvalue (λ1) images using the tbss_non_FA function in order to generate skeletonized MD and λ1 files.

As opposed to using voxel-wise permutation tests for significance, our purely confirmatory approach allowed us to extract and average FA/MD/λ1 from ROIs based on spatial maps provided by the original authors. For the TBSS procedure, the spatial maps provided by the original authors were registered to the mean FA template generated by our TBSS procedure. This was done to maximize the overlay between the spatial maps and our study-specific skeletonized FA template. In order to exclude the possibility that this registration step might impact the final hypothesis test, additional exploratory analyses were performed without registering the spatial maps to our FA template. These analyses are not reported here, as their results did not differ from our main analyses in terms of interpretation (i.e., Bayes factors were comparable).

After extracting FA/MD/λ1 signal from the ROIs, we then used one-sided Bayesian correlation tests (described below) to quantify evidence in favor of either the null hypothesis (H_0) or the alternative hypothesis (H_1). In our analyses, H_1 represents the presence of either a positive or a negative correlation (depending on the predicted direction of the correlation), and the H_0 represents the absence of the predicted correlation.

2.1.8. Probabilistic tractography

Bayesian estimation of diffusion parameters obtained using sampling techniques (BedpostX) was applied to the pre-processed DWI data. BedpostX uses a dual fibre model which can account for crossing fibres. Estimation of tract strengths (for the replication attempt of Forstmann et al., 2010) was conducted using probabilistic tractography (Behrens et al., 2003). Five thousand tracts were sampled from each voxel in the seed mask (right pre-supplemetary motor area; Pre-SMA) at a curvature threshold of 0.2. Next, the number of samples that reach the classification target mask (e.g., right striatum) was measured. In addition, contralateral exclusion masks were used to discard pathways crossing over to the contralateral hemisphere before traveling to the classification target mask. The number of voxels for which a minimum of 10 samples reached the classification mask 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.

2.1.9. Voxel-Based Morphometry

Voxel-Based Morphometry (VBM) was performed using FSL's default VBM pipeline (Douaud, Smith, Jenkinson, Behrens, Johansen-Berg et al., 2007; http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLVBM). First, non-brain tissue was removed from T1 images using BET. Second, brain-extracted images were segmented into grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). GM images were non-linearly registered to GM ICBM-152, and averaged to create a study-specific template at 2 mm resolution in standard space. All GM images were then non-linearly registered to the study-specific template. During this stage, each voxel of each registered grey matter image is divided by the Jacobian of the warp field (Good, Johnsrude, Ashburner, Henson, & Friston et al., 2001). Images were smoothed using a Gaussian kernel with a sigma of 3 mm.

As opposed to using voxel-wise permutation tests for significance, our purely confirmatory approach allowed us to extract and average GM volume from ROIs based on spatial maps provided by the original authors. We then used one-sided Bayesian correlation tests (described below) to quantify evidence in favor of either H_0 or H_1.

2.1.10. Cortical thickness analysis

Cortical reconstruction and volumetric segmentation was performed with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/). The technical details of these procedures are described elsewhere (Dale and Sereno, 1993; Dale, Fischl, & Sereno, 1999; Fischl and Dale, 2000; Fischl, Lio, & Dale, 2001; Fischl, Salat, Busa, Albert, Dieterich et al., 2002; Fischl, Salat, van der Kouwe, Makris, Ségonne, Dale, Busa, Glessner, Salat et al., 2004a; Fischl, Sereno, & Dale, 1999a; Fischl, Sereno, Tootell, & Dale, 1999b; Fischl, van der Kouwe, Destrieux, Halgren, Ségonne et al., 2004b; Han,

Jovicich, Salat, van der Kouwe, Quinn et al., 2006; Jovicich, Czanner, Greve, Haley, van der Kouwe et al., 2006; Reuter, Rosas, & Fischl, 2010; Reuter, Schmansky, Rosas, & Fischl, 2012; Ségonne, Dale, Busa, Glessner, Salat et al., 2004). FreeSurfer pre-processing included motion correction (Reuter et al., 2010) of volumetric T1-weighted images, removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Ségonne et al., 2004), automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, and ventricles; Fischl et al., 2002; 2004a) intensity normalization (Sled, Zijdenbos, & Evans, 1998), tessellation of the gray/ white matter boundary, automated topology correction (Fischl et al., 2001; Ségonne, Pacheco, & Fischl, 2007), and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class (Dale and Sereno, 1993; Dale et al., 1999; Fischl and Dale, 2000). Reconstruction of the GM/WM boundary and pail surface was manually checked for inaccuracies. Subsequently, ROI-labels were mapped onto individual brains and average cortical thickness (Fischl and Dale, 2000) was extracted per ROI, per participant.

2.1.11. General outlier rejection criterion

In the M&A document that we published online prior to inspection of the data, we specified a general outlier rejection criterion. Any deviation of more than 2.5 standard deviations (SDs) from the respective mean results in an exclusion of the participant from the replication in which it is classified as an outlier (as such, a participant can still be included in a different replication, for which he or she was not classified as an outlier).

2.1.12. Confirmatory Bayesian hypothesis test for correlations

Our main analysis goal was to grade the decisiveness of the evidence that the data provide for and against the presence of a correlation between the structural brain measures and the behavioral measures. This goal can be achieved by computing Bayes factors (Dienes, 2008; Jeffreys, 1961; Kass & Raftery, 1995; Lee &

Wagenmakers, 2013; Rouder et al. 2009; Rouder et al., 2012). The Bayes factor compares the adequacy of two models; in our case, the first model is the null hypothesis H_0 that postulates the absence of a correlation between the structural brain measures and the behavioral measures. The second model is the alternative hypothesis H_1 that postulates the presence of a positive (or negative) correlation between the two measures.

The Bayes factor quantifies the odds that the observed data occurred under H_0 versus H_1. For example, a Bayes factor equal to 5.2 indicates that the observed data are 5.2 times as likely to occur under H_0 than under H_1. In this way the Bayes factor provides a continuous measure of evidential support, and its interpretation does not require recourse to actions, decisions, or criteria of acceptance.

To compute the Bayes factor for the Pearson correlation coefficient, we need to specify both H_0 and H_1. Jeffreys (1961) proposed a default test by assigning uninformative priors to the nuisance parameters (i.e., parameters common to H_0 and H_1) and a uniform prior distribution from -1 to 1 to the correlation coefficient ρ that is unique for H_1 (Jeffreys, 1961, p. 291). Consequently, under Jeffreys' alternative hypothesis H_1, each value of the correlation coefficient ρ is a priori equally likely.

Inspired by Jeffreys' test we grade the decisiveness of the evidence by computing BF_{10}, that is, the probability of the observed data under H_1 versus H_0:

$$BF_{10} = \int_0^1 \frac{(1-\rho^2)^{\frac{1}{2}(n-1)}}{(1-\rho r)^{n-\frac{3}{2}}} d\rho \qquad (1)$$

The number of data pairs is denoted by *n,* and *r* is the sample Pearson correlation coefficient. As indicated by the range of integration in Equation 1, we have adjusted Jeffreys test such that the alternative hypothesis is one-sided. The one-sided nature of this test is appropriate, since we intend to replicate SBB correlations, thereby committing to specific directions (as reported in the original studies).

In Equation 1, the integration is from 0 to 1 implying a test for a positive correlation. In case of a test for a negative correlation we simply multiply one of the observed variables with -1. An R function to compute the BF in the above-mentioned way is freely available at http://www.josineverhagen.com/?page_id=76

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 evidence categorization for the BF_{01}, edited by and taken from Wetzels and Wagenmakers (2012; Table 1, p. 1060). In short, a BF_{01} greater than 1 indicates that the data are more likely to occur under H_0 than under H_1. Equivalently, a BF_{01} lower than 1 indicates that the data are more likely to occur under H_1 than under H_0. The evidence categories apply to the BF_{10} (= 1/ BF_{01}; reciprocal of the BF_{01}) in a reversed manner; e.g., a BF_{10} with a value between 10 and 30 provides strong evidence for H_1 and a BF_{10} with a value between 1/10 and 1/30 provides strong evidence for H_0. Thus, when we analyze data and find that, for instance, BF_{01} = 6.5, this means that the data are 6.5 times more likely to have occurred under H_0 than under H_1; similarly, BF_{01} = 0.2 means that the data are 5 times more likely to have occurred under H_1 than under H_0. The labels shown in Table 1 are useful because they facilitate scientific communication; nevertheless, the labels should not be over-interpreted. Many researchers may find the meaning of BF_{01} = 6.5 clear without the help of the labels from Table 1.

Table 1. Categories for the BF_{01}

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

Bayes factor BF_{01}			Interpretation
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

2.1.13. Posterior probability distributions

The posterior distribution is formed by combining the information or beliefs about the correlation available prior to the experiment (as expressed in the prior distribution), with the correlation observed in the data.

In a situation where nothing is known about the correlation prior to the experiment, an uninformative uniform prior distribution can be used, in which every correlation between -1 and 1 has equal probability (Figure 1 black line). In this situation, once a correlation has been observed, the posterior distribution will have a higher probability around the observed correlation and less probability at values further away (Figure 1 red line).The posterior distribution represents the knowledge we have about the correlation of interest after observing the data.

When we want to update this knowledge with a new experiment, the posterior from the previous experiment can be taken as the prior for the next experiment. This indicates that the correlation in the new study is expected to be similar to the correlation in the previous study, as the prior gives more probability to values closer to the previously observed correlation. When this informative prior distribution is updated by the correlation observed in a new experiment, the final posterior distribution will be identical to the posterior distribution had all data been analyzed together from the start (Figure 1 blue line).

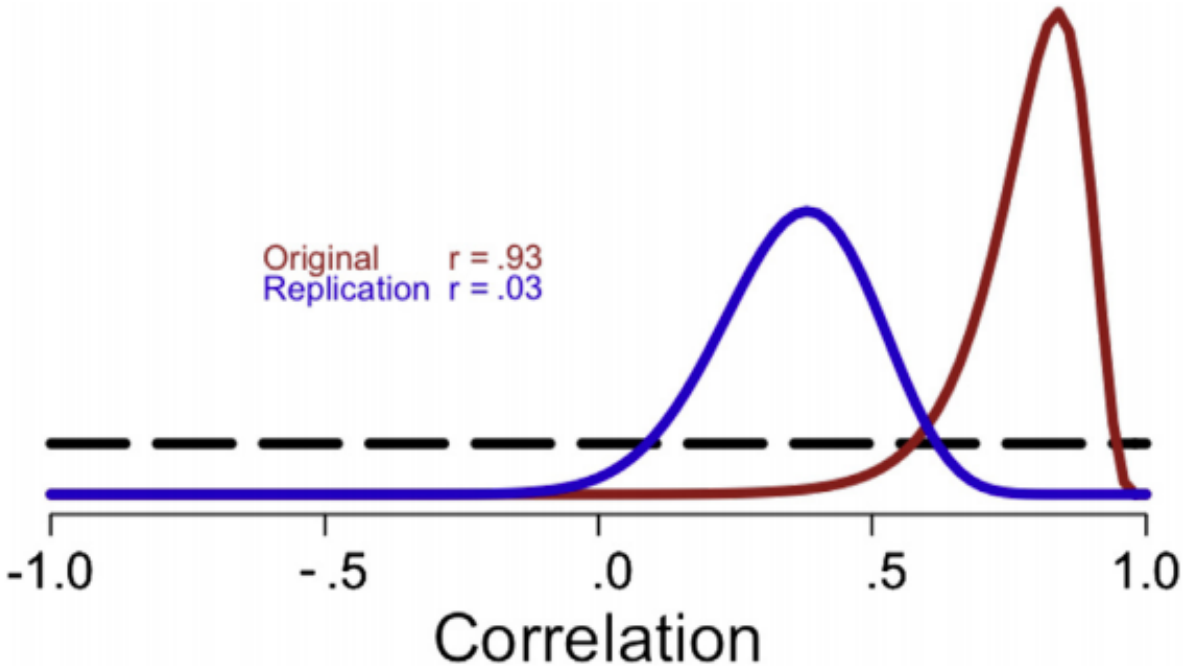


Fig 1. Posteriors plot.

Example of a posterior plot, showing uniform prior distribution (black line), the posterior after the original effect (red line), and the posterior after the replication effect (blue line), using the posterior as a prior distribution.

We will use the posterior distribution of the previous study in a different way, for model comparison. In this case, the posterior distribution from the original study is used to represent the hypothesis that the observed correlation is similar to the previous correlation.

2.1.14. Additional exploratory analyses

In addition to the Bayesian test described above, we computed an additional Bayesian test in which H_1 is specified not only to the direction of the effect found in the original study, but also to its effect size (Verhagen and Wagenmakers, 2014). In this way, this test answers the question 'Is the effect from the replication attempt comparable to what was found before, or is it absent?', whereas the original Bayesian test answers the question 'Is the effect present or absent in the data from the replication attempt?'. We

label this additional analysis exploratory as it was not described and published in the M&A document prior to inspection of the data.

The replication Bayes factor compares evidence in favor of the null hypothesis of no effect, H_0*: $\rho = 0$*, with the evidence in favor of the alternative hypothesis that the effect is equal to the effect found in the original study, H_r*: $\rho \sim$ posterior distribution from original study*. The resulting Bayes factor is similar to the Bayes factor in equation (1), with the only difference that the replication Bayes factor is obtained by integrating over the posterior distribution from the first study instead of a uniform distribution. A more detailed description of the replication Bayes factor can be found in Appendix A. R code to perform this analysis can be found in this link http://www.josineverhagen.com/?page_id=76

In addition to the Bayes factor tests, an intuitive assessment of the extent to which our results replicate the original studies can also be obtained by comparing the posterior distributions for the correlation coefficients in the original and replication studies. We facilitate such a comparison by plotting, for each of the five replication attempts, both the entire posterior distribution and a summary in terms of 95% credible intervals.

Finally, for frequentist readers we provide p-values. Once again, these are labeled as exploratory given that we did not preregister the use of frequentist statistics in our M&A document.

2.2. Study-specific methods

Below we describe study-specific methods for the five experiments included in the final replication attempt remaining after study exclusion. For each experiment we describe the stimuli and procedure, behavioral analyses, structural brain analyses, and statistical tests based on hypotheses generated by the original papers.

2.2.1. Replication 1: Forstmann et al. (2010)

2.2.1.1. Random dot motion task and procedure

We used the same random dot motion (RDM) task (Gold and Shadlen, 2007) as Forstmann and colleagues (2010). The task contained 360 trials in total, with 180 speed and 180 accuracy trials. The RDM cloud consisted of 60 coherently moving white dots and 60 randomly moving white dots, presented against a black background (see http://wouterboekel.com/CONFREP/dots_loop.gif). A single dot consisted of 3 pixels and the entire cloud spanned 250 pixels. At the start of each trial, either a speed cue or an accuracy cue was presented for 1000 ms. The speed cue instructed participants to respond as quickly as possible. The accuracy cue instructed participants to respond as accurate as possible. The cue was followed by a fixation cross presented at the center of the screen for 500 ms. Subsequently, the RDM stimulus was presented for 1500 ms or until a response was made. Responses outside of this time window were ignored. Participants responded on a keyboard by pressing 'a' with their left index finger when they perceived a leftward motion and 'l' with their right index finger when they perceived a rightward motion. Immediately after the response, participants received a feedback message for 400 ms. On speed trials, the feedback read either 'te traag' or 'op tijd' (i.e., Dutch for 'too slow' and 'in time'). On accuracy trials, the feedback read either 'fout' or 'goed' (i.e., Dutch for 'incorrect' and 'correct'). 45-second breaks were inserted after 120 and after 240 trials. The entire task lasted for approximately 20 minutes.

2.2.1.2. LBA model

The linear ballistic accumulator (LBA; Brown and Heathcote, 2008) model decomposes the response time and accuracy measures into latent psychological processes. It assumes that when given a choice between two alternatives, evidence accumulates from a start point (*A*), at a certain speed (drift rate *v*), for both alternatives separately. When one of these accumulators reaches its response threshold (*b*), a decision is made in favor of the associated alternative. Response time is determined by the time taken to reach the threshold, plus an offset time for stimulus encoding and motor processes (non-decision time *t0*) (Figure 2).

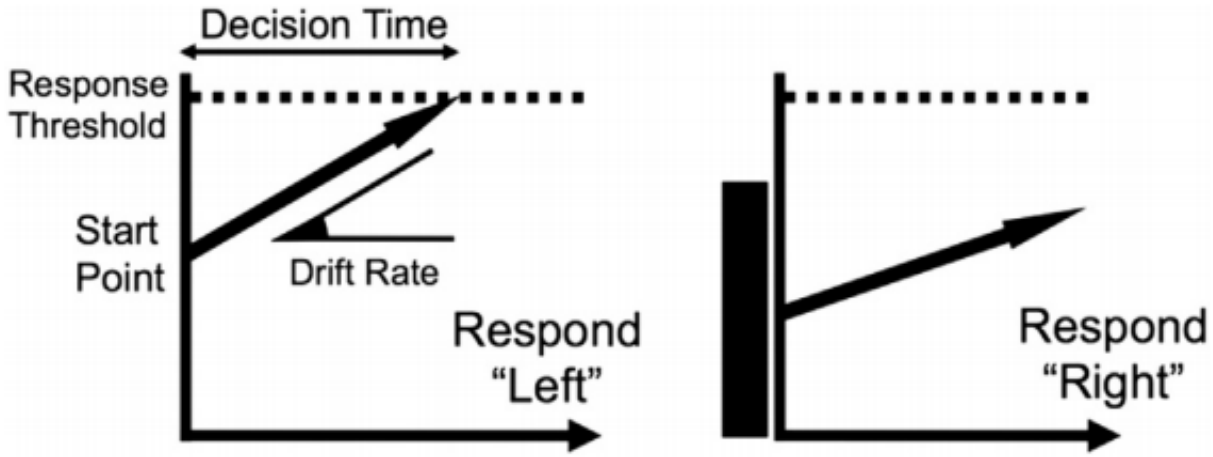


Fig 2. Schematic representation of the LBA model used in the replication of Forstmann et al (2010).

In the LBA model, the decision to respond either left or right is modeled as a race between 2 accumulators. Activation in each accumulator begins at a random point between zero and start point *A* and increases with time. The rate of increase is random from trial to trial, but is (on average) faster for the accumulator whose associated response matches the stimulus. A response is given by whichever accumulator first reaches the threshold *b*, and the predicted response time depends on the time taken to reach that threshold.

The element of central interest here is response caution, which can be quantified via the threshold height in the LBA. We applied the same parameter constraints as Forstmann and colleagues (2010). In this design only one parameter—response threshold *b*—is free to vary with the speed vs. accuracy cue, while all other parameters (width of start point distribution *A*, drift rate *v*, variability of the drift rate *s*, and nondecision time *t0*) are fixed. Response caution is measured by subtracting start point *A* from response threshold *b*.

2.2.1.3. Behavioral data analysis

The behavioral measure of interest is the LBA flexibility parameter, assessing efficacy of changing response caution. It is assumed that “changes in response caution originate from adjustments of response thresholds (Forstmann et al., 2010; page 1516)”. Therefore, LBA flexibility was computed as the difference between the LBA caution estimates for the accuracy and the speed conditions. We fit the LBA model to each participants accuracy and reaction time (RT) distributions on speed and accuracy trials separately. The only parameter allowed to vary was the response threshold *b*. The resulting individual LBA flexibility estimates were imported into R software (R Foundation for Statistical Computing, http://www.R-project.org) for the Bayesian correlation test.

2.2.1.4. Probabilistic tractography

We limited our tractography to delineate tracts that the authors found to correlate significantly with LBA flexibility. Hence, probabilistic tractography was performed only between right pre-SMA and right striatum. Here we used the same MNI-space masks for right pre-SMA and right striatum as were used in Forstmann et al. (2010). We performed the probabilistic tractography in accordance with the protocol stated in the general methods section (see above). Resulting tract strength values were corrected for age and gender using partial correlations, and were subsequently imported into R software for the Bayesian correlation test. Specifically, we tested for a positive correlation between right pre-SMA–right striatum tract strength and LBA flexibility.

2.2.2. Replication 2: Kanai et al. (2012)

2.2.2.1. Social Network Size Questionnaire and procedure

Participants completed a Dutch version of the Social Network Size questionnaire (Stileman & Bates, 2007). This questionnaire consists of 9 items. One of its items is: “How many friends do you have on ‘Facebook’?”. We asked participants to make a note of the number of friends they have on ‘Facebook’ or an alternative comparable social network site such as ‘myspace’ or the Dutch ‘Hyves’ and bring it to the test session. The administration time is approximately 10 minutes.

2.2.2.2. Behavioral data analysis

The behavioral measures of interest are online social network size (i.e., FBN) and real-world social network size. As was done in Kanai et al. (2012), answers to the 9 subquestions contained in this questionnaire were square-root transformed to correct for skewness. We computed the FBN as the square root of participants answer to the question: “How many friends do you have on ‘Facebook’?”. A normalized real-world social network size score (SNS) was computed per participant by averaging the z-scores for the questionnaire items 1, 2, 4, 5, 6, 8, and 9 after skewness correction. For each participant an online social network size score and a real-world social network size score was imported into R software for the Bayesian correlation test.

2.2.2.3. ROI generation

Kanai et al. (2012) reported significant positive correlations between online social network size and GM volume within left middle temporal gyrus (MTG), right superior temporal sulcus (STS), right entorhinal cortex (EC), and bilateral amygdala. In addition, real-world social network size was positively correlated with GM volume only within right amygdala. We defined all these regions as our ROIs. Dr. Kanai kindly provided us with the spatial maps of these regions.

2.2.2.4. Correlational analysis

For every participant, we extracted GM volume values from all voxels contained in the ROIs and averaged them. These GM volume measures were then corrected for age, gender and total grey matter volume. The corrected mean GM volume measures were imported into R software for the Bayesian correlation test. Specifically, we tested for positive correlations between FBN and mean GM volume within left MTG, right STS, right EC, and bilateral amygdala. Furthermore, we tested for a positive correlation between SNS and mean GM volume within right amygdala.

2.2.3. Replication 3: Xu et al. (2012)

2.2.3.1. BIS/BAS questionnaire and procedure

Participants completed a Dutch version of the Behavioral Inhibition System/Behavioral Activation System scale (BIS/BAS; Carver et al., 1994). The BIS/BAS is a 20-item questionnaire. Our interest was focused on the BAS scale, which comprises 13 items (BAS-Total) and has three sub-scales: Drive (BAS-Drive), Fun-Seeking (BAS-Fun), and Reward-Responsiveness (BAS-Reward).

2.2.3.2. Behavioral analysis

The behavioral measures of interest were BAS-Total scores and BAS-Fun scores. BAS-Total scores assess the sensitivity to signals of reward and non-punishment. BAS-Fun scores assess the tendency to seek out new potentially rewarding experiences. For each participant these scores were imported into R software for the Bayesian correlation test.

2.2.3.3. ROI generation

Xu et al. (2012) reported significant positive correlations between the BAS-Total scores and λ1 within left corona radiata (CR) and left superior longitudinal fasciculus (SLF). Furthermore, they reported positive correlations between the BAS-Fun scores and λ1 as well as FA within left CR and left SLF. The authors also reported significant positive correlations between the BAS-Fun scores and MD within left inferior longitudinal fasciculus (ILF) and left inferior fronto-occipital fasciculus (IFOF). We defined all these WM tracts as our ROIs. Dr. Xu kindly provided us with the spatial maps of these areas.

2.2.3.4. Correlational analysis

For every participant, we extracted FA, MD, and λ1 values from all voxels contained in the respective ROIs and averaged them. These values were then corrected for age and gender using partial correlations. Unlike Xu et al. (2012), we did not need to correct for differences in education because our participants were all first-year Psychology students. The corrected mean WM tract measures per ROI were imported into R software for the Bayesian correlation test. Specifically, we tested for positive correlations between BAS-Total scores and mean λ1 within left CR and left SLF. Furthermore, we

tested for positive correlations between BAS-Fun scores and mean λ1 as well as mean FA within left CR and left SLF. Finally, we tested for positive correlations between BAS-Fun scores and mean MD within left ILF and left IFOF.

2.2.4. Replication 4: Kanai et al. (2011)

2.2.4.1. Cognitive Failures Questionnaire and procedure

Participants completed a Dutch version of the Cognitive Failures Questionnaire (CFQ, Broadbent et al., 1982).

2.2.4.2. Behavioral data analysis

The behavioral measure of interest is distractibility as assessed by the CFQ. As in Kanai et al. (2011), we quantified distractibility by computing the standard loadings derived from a previous factor analysis (Wallace, Kass, & Stanny, 2002). Specifically, we used the following 9 items: 1, 2, 3, 4, 15, 19, 21, 22, and 25. Scores on these items were imported into R software for the Bayesian correlation test.

2.2.4.3. ROI generation

Kanai et al. (2011b) reported a significant positive correlation between CFQ scores and GM volume within left superior parietal lobe (SPL). Furthermore, the authors reported a negative correlation between CFQ scores and GM volume within left middle prefrontal cortex (mPFC). We defined these regions as our ROIs. Dr. Kanai kindly provided us with the spatial maps of these regions.

2.2.4.4. Correlational analysis

For every participant, we extracted GM volume values from all voxels contained in the respective ROIs and averaged them. These GM volume values were then corrected for age, gender and total grey matter volume using partial correlations. The corrected mean GM volume values were imported into R software for the Bayesian correlation test. Specifically, we tested for a positive correlation between CFQ scores and mean GM

volumes within left SPL, and for a negative correlation between these measures within left mPFC.

2.2.5. Replication 5: Westlye et al. (2011)

2.2.5.1. Attention Network Test

We used the same Attention Network Test as Westlye and colleagues (2011; downloaded from Dr. Jin Fan's website www.sacklerinstitute.org/users/jin.fan). The task included 2 runs of 96 trials and 20 practice trials. Each trial began with the presentation of a fixation cross in the centre of the screen for variable durations (400, 800, 1200, or 1600 ms). Subsequently, one of three cues was presented for 100 ms: (1) no cue, (2) centre cue (*, replacing fixation cross), or (3) spatial cue (*, above or below fixation cross). This was followed by the presentation of the target for a maximum duration of 1700 ms, or until a response was made. The target was an arrow in the centre of a row of 5 arrows, presented either below or above the fixation cross. The flanking arrows consisted of either (1) two congruent arrows (pointing in the same direction as the target), (2) two incongruent arrows (pointing in the opposite direction of the target), or (3) two lines on each side of the target (neutral). Participants were instructed to report the direction (left or right) of the target arrow by pressing the spatially compatible key ('left mouse button' or 'right mouse button') with their left or right thumb. The entire experiment took approximately 15 minutes.

2.2.5.2. Behavioral data analysis

The behavioral measures of interest are executive control and alerting network scores, assessing the executive control and the alerting components of attention, respectively. We applied the same processing steps as described by Westlye et al. (2011) prior to computing these scores: *"To remove outliers, all RTs > 1500 ms and < 200 ms were removed (...). Next, since error responses are assumed to originate from a different RT distribution than correct responses, we only analyzed correct responses. Also, because responses following erroneous responses typically are slower than responses following correct responses (posterror slowing), we also removed responses following erroneous*

responses. Since RTs are not normally distributed, we used median RT per condition as raw scores for each subject. (...). (page 348).” However, we did not adjust the component scores with the baseline RT in order to control for an effect of age on RT, because our participants form a homogenous age group (Psychology freshmen).

Based on median RT, the executive control and alerting scores will be computed as follows:

Executive control = $[RT_{incongruent} - RT_{congruent}] / RT_{congruent}$
Alerting = $[RT_{no\ cue} – RT_{center\ cue}] / RT_{center\ cue}$

For each participant, the resulting scores were imported into R software for the Bayesian correlation test.

2.2.5.3. ROI generation

For their subsample of young participants, Westlye et al. (2011) reported significant negative correlations between executive control scores and CT within left caudal anterior cingulate cortex (ACC), left superior temporal gyrus (STG), and right middle temporal gyrus (MTG). The alerting scores showed a significant negative correlation with CT within left superior parietal lobe (SPL). We defined all these regions as our ROIs. Dr. Westlye kindly provided us with the FreeSurfer labels of these areas.

2.2.5.4. Correlational analysis

For every participant, we extracted CT values from all voxels contained in the ROIs and averaged them. These CT measures were then corrected for age and gender using partial correlations. The corrected mean CT measures were imported into R software for the Bayesian correlation test. Specifically, we tested for negative correlations between executive control scores and mean CT within left caudal ACC, left STG and right MTG. Furthermore, we also tested for a negative correlation between alerting scores and mean CT within left SPL.

3. Results

Below we describe study-specific results for the five experiments included in the final replication attempt remaining after study exclusion, comprising a total of 17 predicted effects. For each study, we briefly re-iterate the original findings, followed by our predictions based on these findings. We describe potential outlier exclusion and list the Bayes factors in favor of the null hypothesis (BF_{01}). Furthermore, we describe the outcome of the additional exploratory Bayes factor analysis that uses an informative prior distribution (cf. Verhagen and Wagenmakers, 2014).

3.1. Replication 1: Forstmann et al. (2010)

Forstmann et al. (2010) reported that individual differences in tract strength from right pre-SMA to right striatum predict individual differences in control over speed and accuracy in a perceptual decision making paradigm. The original authors replicated their effect in an independent data set. In line with the original authors' theorizing and results, we hypothesized the presence of a positive correlation between pre-SMA-striatum tract strength and LBA flexibility.

Three participants did not complete the behavioral task and were thus excluded from further analysis. Tract strengths of 2 out of 33 participants deviated more than 2.5 SDs from the group mean, and were thus excluded from this replication attempt. After outlier rejection, tract strength data ranged from 0.682 to 0.914, with a mean of 0.819 and a standard deviation of 0.061. LBA flexibility ranged from 0.020 to 1.554, with a mean of 0.578 and a standard deviation of 0.410. A one-sided Bayesian hypothesis test for positive correlations was performed on these data. Its result is shown in Table 2 and Figure 3. The Bayes factor shows that there is moderate support for the null hypothesis of no correlation. In order to provide a complete report of the SBB correlation found here in comparison with the original finding, Figure S2 shows posterior probability plots of this effect.

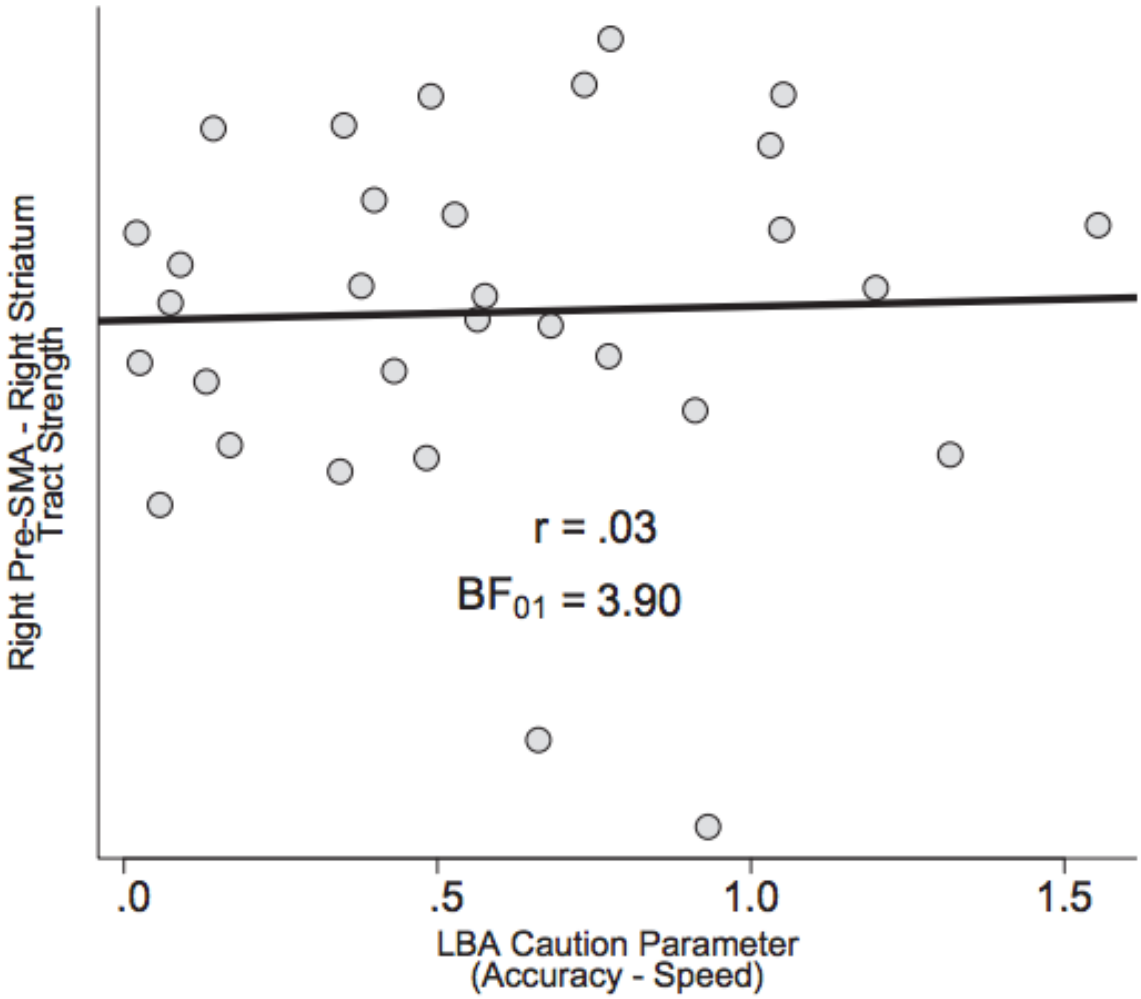


Fig 3. Scatterplot of replication 1: Forstmann et al. (2010).

The relationship between LBA caution parameter (quantified by taking the difference in response caution between the accuracy and speed condition) and tract strength between right Pre-SMA and right Striatum, quantified by probabilistic tractography.

Table 2. Results of the one-sided Bayesian hypothesis test for a positive correlation

Data pair					**Confirmatory**		**Exploratory**		
ROI	n_{ori} $_{g}$	n_{re} $_{p}$	r_{orig}	r_{rep}	BF_{01}	Evidence cat.	BF_{0r}	Evidence cat.	*p*-value
Tract strength and LBA flexibility									

Data pair					Confirmatory		Exploratory		
ROI	n_{orig}	n_{rep}	r_{orig}	r_{rep}	BF_{01}	Evidence cat.	BF_{0r}	Evidence cat.	*p*-value
Pre-SMA to striatum	9	31	0.93	0.03	3.90	Moderate (H_0)	180.20	Extreme (H_0)	0.431

The additional exploratory Bayes factor analysis with informative priors (Verhagen and Wagenmakers, 2014) shows that the data are extremely likely to have occurred under the null hypothesis compared to the proponent's hypothesis. Figure S2 (bottom) shows posteriors for this exploratory Bayes factor analysis. The p-value indicated a failed replication.

3.2. Replication 2: Kanai et al. (2012)

Kanai et al. (2012) showed that individual differences in the number of Facebook friends (FBN) and real-world social network size (SNS) are positively correlated with GM volume in several brain areas. The original authors replicated their effects in an independent data set. In line with the original authors' theorizing and results, we hypothesized positive correlations between FBN and GM volume in left MTG, right STS, right EC, and bilateral amygdala. In addition, we hypothesized a positive correlation between SNS and GM volume in right amygdala.

One participant did not complete the FBN and two participants did not complete the SNS questionnaire, and were thus excluded from further analysis. One participant was excluded in 4 out of 6 Bayesian correlations, due to a GM volume measure deviating more than 2.5 SDs from the group mean. After outlier rejection, the following summary statistics describe our data: FBN: range: 10.0499 – 24.7386, mean: 17.096, sd: 3.788. SNS: range: -1.05 – -0.44, mean: -0.650, sd: 0.153. GM in left MTG: range: 0.411 – 0.562, mean: 0.476, sd: 0.035. GM in right STS: range: 0.336 – 0.595, mean: 0.484, sd:

0.062. GM in right EC: range: 0.521 – 0.785, mean: 0.628, sd: 0.063. GM in left Amygdala: range: 0.636 – 0.770, mean: 0.707, sd: 0.033. GM in right Amygdala: range 0.603 – 0.772, mean: 0.670, sd: 0.035. One-sided Bayesian hypothesis tests for positive correlations were performed on these data. Results are shown in Table 3 and Figure 4. In 5 out of 6 cases we find support for the null hypothesis. The Bayes factors show that there is moderate support for the null hypothesis in 3 out of 6 effects (i.e., no correlations between FBN and GM volume in right EC, and bilateral amygdala). Our data are ambiguous with regard to the correlations between FBN and GM volume in left MTG and right STS. In order to provide a complete report of the SBB correlations found here in comparison with the original findings, Figures S3-8 show posterior probability plots of these effects.

Table 3. Results of the one-sided Bayesian hypothesis tests for positive correlations

Data pair					**Confirmatory**		**Exploratory**		
ROI	n_{orig}	n_{rep}	r_{orig}	r_{rep}	BF_{01}	Evidence cat.	BF_{0r}	Evidence cat.	*p*-value
FBN and GM volume									
left MTG	125	34	0.35	0.18	1.73	Anecdotal (H_0)	1.06	Anecdotal (H_0)	0.158
right STS	125	35	0.35	0.11	2.66	Anecdotal (H_0)	2.06	Anecdotal (H_0)	0.261
right EC	125	35	0.35	0.06	3.51	Moderate (H_0)	3.32	Moderate (H_0)	0.360
left amygdala	125	34	0.30	-0.14	7.76	Moderate (H_0)	9.56	Moderate (H_0)	0.779
right amygdala	125	34	0.32	0.02	4.35	Moderate (H_0)	3.88	Moderate (H_0)	0.462
SNS and GM volume									
right amygdala	65	33	0.26	0.30	0.57	Anecdotal (H_1)	0.27	Moderate (H_r)	0.041

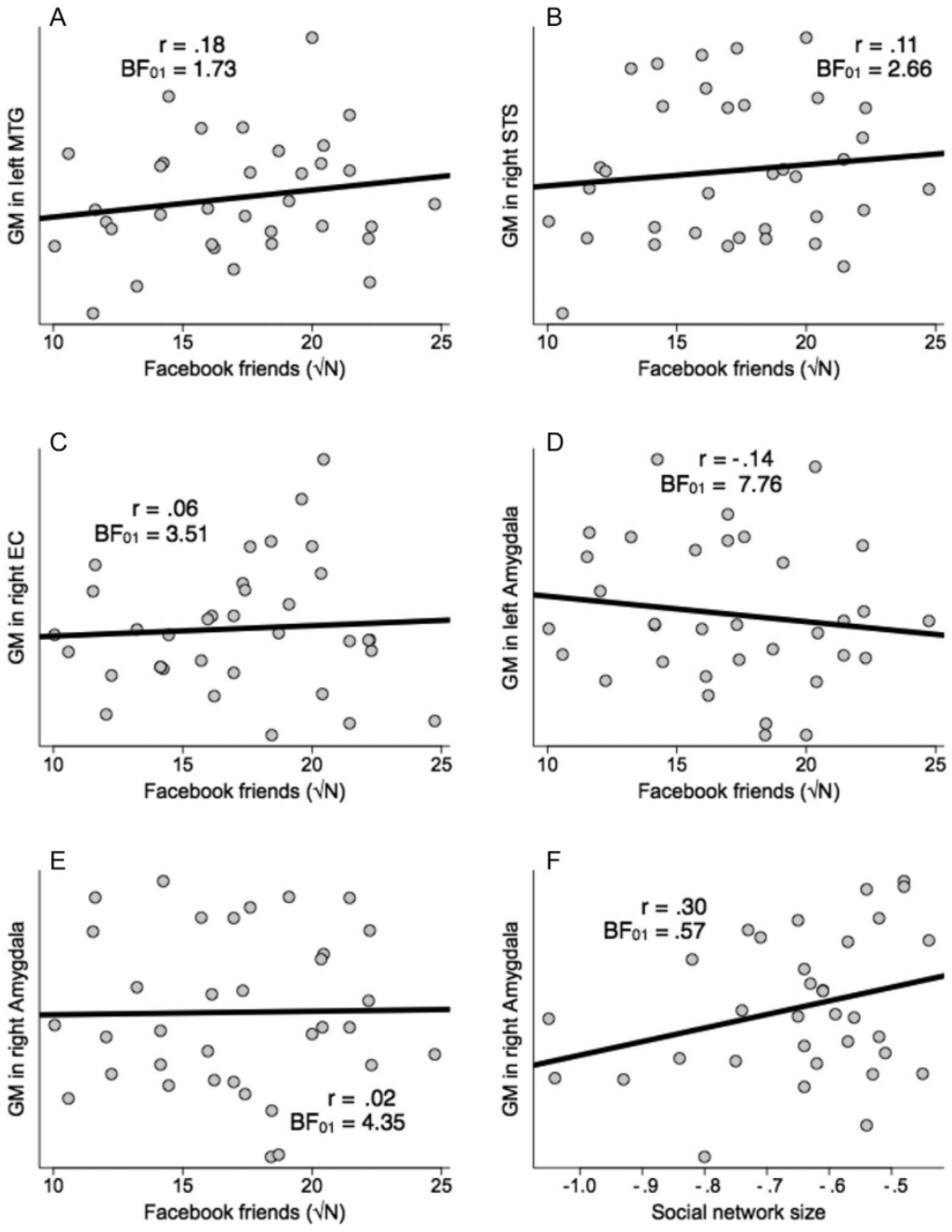


Fig 4. Scatterplots of replication 2: Kanai et al. (2012).

(A-E) The relationship between the number of Facebook friends and GM in (A) left MTG, (B) right STS, (C) right EC, (D) left amygdala, (E) right amygdala. (F) the relationship between real world social network size and GM in the right amygdala.

The additional exploratory Bayes factor analyses with informative priors (Verhagen and Wagenmakers, 2014) show that for two effects there is anecdotal evidence in favor of the null hypothesis compared to the proponent's hypothesis. For three effects there is moderate evidence in favor of H_0, and for one effect there is moderate evidence in favor of H_r, compared to H_0. Figures S3-8 (bottom) show posteriors for these exploratory Bayes factor analyses. P-values indicate failed replications for 5 out of 6 effects. For the correlation between SNS and GM volume in right amygdala, the p-value indicates a successful replication.

3.3. Replication 3: Xu et al. (2012)

Xu et al. (2012) reported that individual differences in diffusion measures of several white matter pathways are positively correlated with individual differences in the tendency to seek out new potentially rewarding experiences (i.e., BAS-Fun) and the sensitivity to signals of reward and non-punishment (BAS-Total). In line with the original authors' theorizing and results, we hypothesized a positive correlation between the BAS-Total scores and λ1 within left CR and left SLF, a positive correlation between BAS-Fun and FA in left CR and SLF, a positive correlation between BAS-FUN and λ1 in left CR and SLF, and a positive correlation between BAS-Fun and MD in left ILF and IFOF.

One participant was excluded from λ1 analyses due to white matter structural measures deviating more than 2.5 SDs from the group mean. After outlier rejection, the following summary statistics describe our data: BAS-Total: range: 14 – 31, mean: 22.833, sd: 3.783. BAS-FUN: range: 5 – 12, mean: 7.667, sd: 1.821. FA in left CR and SLF: range: 0.649 – 0.810, mean: 0.736, sd: 0.039. λ1 in left CR and SLF: range: 7.4E4 – 9.2E4, mean: 8.2E4, sd: 3.7E5. MD in left SLF and IFOF: range: 3.9E4 – 4.7E4, mean: 4.3E4, sd: 1.8E5. One-sided Bayesian hypothesis tests for positive correlations were performed on these data. Results are shown in Table 4 and Figure 5. In all cases we find support for the null hypothesis. The Bayes factors show that there is moderate or strong support for the null hypothesis in 3 out of 4 tests (i.e., no correlation between BAS-Total and λ1 in left CR and SLF, no correlation between BAS-FUN and FA in left

CR and SLF, and no correlation between Bas-FUN and λ1 in left CR and SLF). Our data are ambiguous with regard to the correlation between bas-FUN and MD in left ILF and IFOF. In order to provide a complete report of the SBB correlations found here in comparison with the original findings, Figures S9-12 show posterior probability plots of these effects.

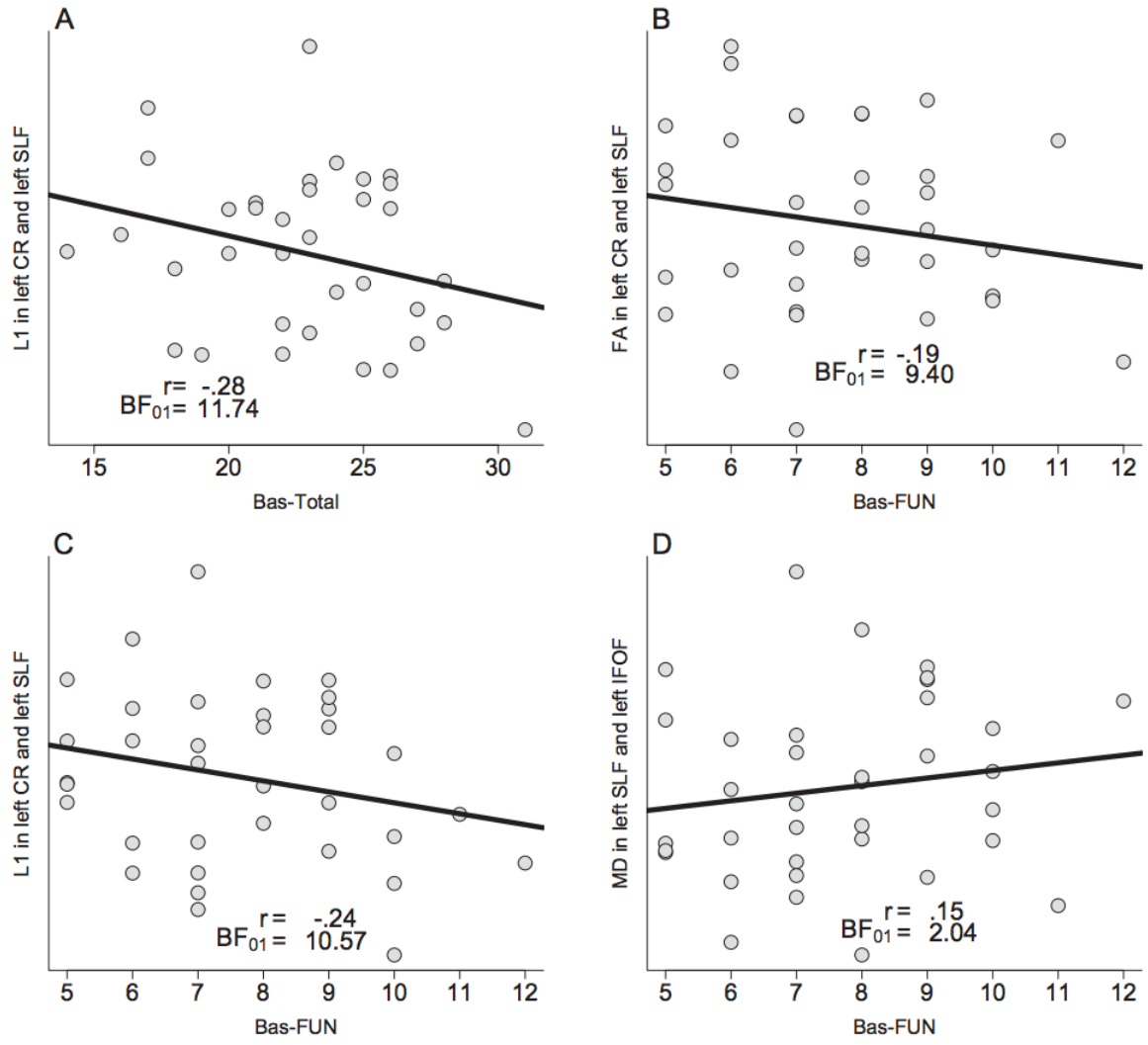


Fig 5. Scatterplots of replication 3: Xu et al. (2012).

(A) The relationship between Bas-total and λ1 in left CR and left SLF. (B-D) The relationship between Bas-FUN and (B) FA in left CR and left SLF, (C) λ1 in left CR and left SLF, and (D) MD in left SLF and left IFOF.

Table 4. Results of the one-sided Bayesian hypothesis tests for positive correlations

Data pair					Confirmatory		Exploratory		
ROI	n_{orig}	n_{rep}	r_{orig}	r_{rep}	BF_{01}	Evidence cat.	BF_{0r}	Evidence cat.	*p*-value
BAS-Total and λ1									
Left CR and SLF	51	35	0.51	-0.28	11.74	Strong (H_0)	249.41	Extreme (H_0)	0.948
BAS-FUN and FA									
Left CR and SLF	51	36	0.52	-0.19	9.40	Moderate (H_0)	170.51	Extreme (H_0)	0.861
BAS-FUN and λ1									
Left CR and SLF	51	35	0.58	-0.24	10.57	Strong (H_0)	848.06	Extreme (H_0)	0.915
BAS-FUN and MD									
Left SLF and IFOF	51	36	0.51	0.15	2.04	Anecdotal (H_0)	4.13	Moderate (H_0)	0.187

The additional exploratory Bayes factor analyses with informative priors (Verhagen and Wagenmakers, 2014) show that for three effects there is extreme evidence in favor of the null hypothesis compared to the proponent's hypothesis, and for one effect there is moderate evidence in favor of H_0. Figures S9-12 (bottom) show posteriors for these exploratory Bayes factor analyses. All p-values indicate failed replications.

3.4. Replication 4: Kanai et al. (2011)

Kanai et al. (2011) reported that individual differences in the degree of distractibility (CFQ) are correlated with GM volume in several brain areas. In line with the original authors' theorizing and results, we hypothesized a positive correlation between CFQ

scores and GM volume in left SPL, and a negative correlation between CFQ and GM volumes in left mPFC.

The following summary statistics describe our data: CFQ: range: 5 – 29, mean: 16.472, sd: 5.443. GM in left SPL: range: 0.378 – 0.812, mean: 0.545, sd: 0.113. GM in left mPFC: range: 0.342 – 0.693, mean: 0.499, sd: 0.101. Results of the one-sided Bayesian hypothesis tests for correlations are shown in Table 5 and Figure 6. In both cases we find anecdotal support ("not worth more than a bare mention", Jeffreys, 1961, Appendix B) for the null hypothesis. In order to provide a complete report of the SBB correlations found here in comparison with the original findings, Figures S13-14 show posterior probability plots of these effects.

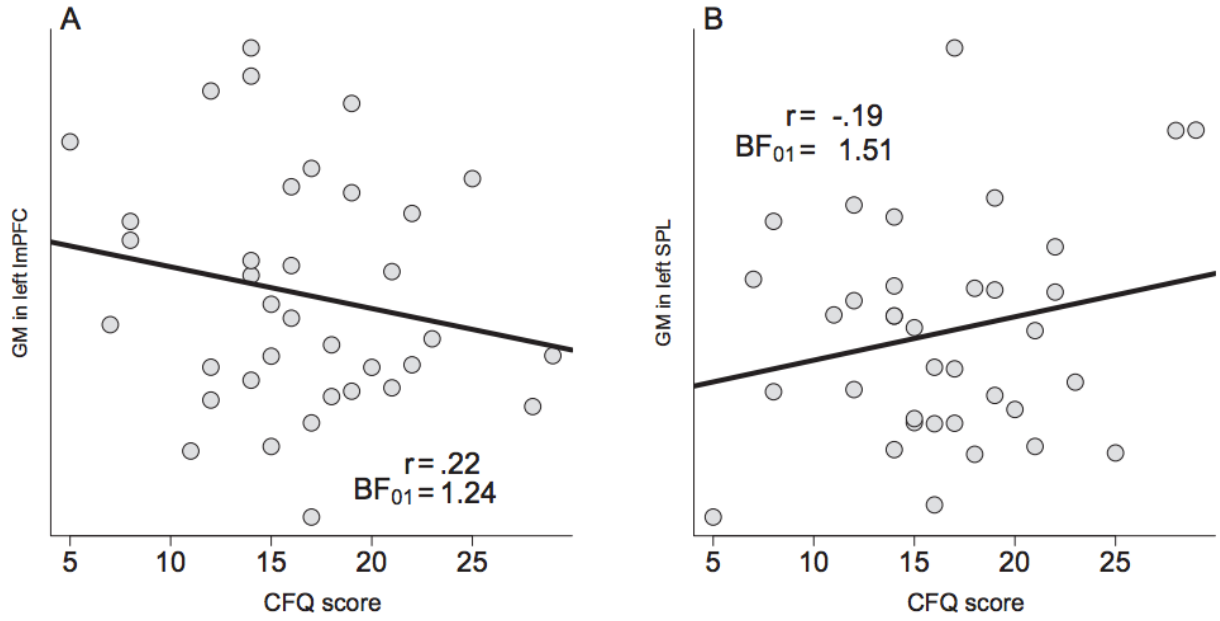


Fig 6. Scatterplots of replication 4: Kanai et al. (2011).
The relationship between CFQ score and GM in (A) left lmPFC, and (B) left SPL.

Table 5. Results of the one-sided Bayesian hypothesis tests for positive correlations. In line with the prediction of a negative correlation, the test was flipped in sign for the correlation between CFQ and GM in left mPFC.

Data pair					**Confirmatory**		**Exploratory**		
ROI	n_{orig}	n_{rep}	r_{orig}	r_{rep}	BF_{01}	Evidence cat.	BF_{0r}	Evidence cat.	*p*-value
CFQ and GM volume									
Left SPL	144	36	0.38	0.22	1.24	Anecdotal (H_0)	0.73	Anecdotal (H_r)	0.102
Left mPFC	144	36	-0.28	-0.19	1.51	Anecdotal (H_0)	0.67	Anecdotal (H_r)	0.129

The additional exploratory Bayes factor analyses with informative priors (Verhagen and Wagenmakers, 2014) show that for both effects there is anecdotal evidence in favor of the proponent's hypothesis compared to the null hypothesis. Figures S13-14 (bottom) show posteriors for these exploratory Bayes factor analyses. All p-values indicate failed replications.

3.5. Replication 5: Westlye et al., 2011

Westlye et al. (2011) reported that individual differences in aspects of attention (executive control and alerting) are correlated with cortical thickness in several brain areas. In line with the original authors' theorizing and results, we hypothesized negative correlations between executive control scores and CT in left caudal anterior cingulate cortex, left superior temporal gyrus, and right middle temporal gyrus. In addition, we

hypothesized a negative correlation between alerting scores and CT in left superior parietal lobe.

One participant was excluded due to cortical thickness measures deviating more than 2.5 SDs from the group mean. After outlier rejection, the following summary statistics describe our data: Alerting: range: -0.068 – 0.157, mean: 0.064, sd: 0.050. Executive control: range: 0.057 – 0.402, mean: -0.229, sd: 0.082. CT in left caudal ACC: range: 2.464 – 2.979, mean: 2.671, sd: 0.121. CT in left STG: range: 2.692 – 3.075, mean: 2.901, sd: 0.083. CT in right MTG: range: 2.361 – 2.570, mean: 2.478, sd: 0.050. CT in left SPL: range: 2.116 – 2.610, mean: 2.360, sd: 0.103. One-sided Bayesian hypothesis tests for negative correlations were performed on these data. Results are shown in Table 6 and Figure 7. In all cases we find support for the null hypothesis. The Bayes factors show that there is moderate support for the null hypothesis in one out of four tests (i.e., no correlation between alerting scores and CT in left SPL). Our data are ambiguous with regard to the correlations between executive control scores and CT in left caudal ACC, left STG, and right MTG. In order to provide a complete report of the SBB correlations found here in comparison with the original findings, Figures S15-18 show posterior probability plots of these effects.

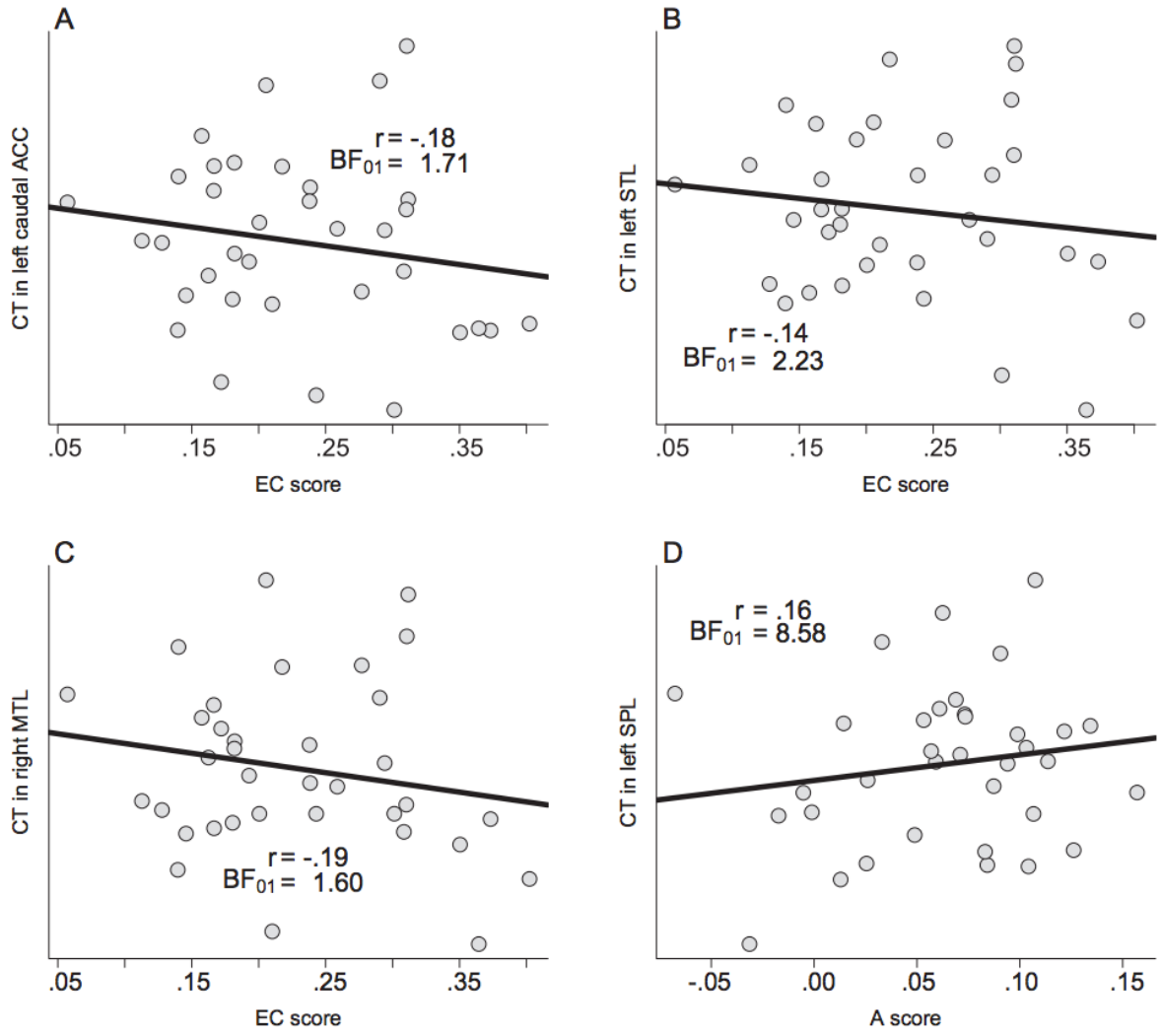


Fig 7. Scatterplots of replication 5: Westlye et al. (2011).

(A-C) The relationship between EC scores and CT in (A) left caudal ACC, (B) left STL, and (C) right MTL. (D) The relationship between A scores and CT in left SPL.

Table 6. Results of the one-sided Bayesian hypothesis tests for positive correlations. In line with the prediction of negative correlations, the tests were flipped in sign.

Data pair					**Confirmatory**		**Exploratory**		
ROI	n_{orig}	n_{rep}	r_{orig}	r_{rep}	BF_{01}	Evidence cat.	BF_{0r}	Evidence cat.	*p*-value
Executive control and CT									
left caudal ACC	132	35	-0.21	-0.18	1.71	Anecdotal (H_0)	0.67	Anecdotal (H_r)	0.153
left STG	132	35	-0.15	-0.14	2.23	Anecdotal (H_0)	0.81	Anecdotal (H_r)	0.211
right MTG	132	35	-0.13	-0.19	1.60	Anecdotal (H_0)	0.65	Anecdotal (H_r)	0.141
Alerting and CT									
left SPL	132	35	-0.26	0.16	8.58	Moderate (H_0)	7.70	Moderate (H_0)	0.824

The additional exploratory Bayes factor analyses with informative priors (Verhagen and Wagenmakers, 2014) show that for 2 effects there is anecdotal evidence in favor of the null hypothesis compared to the proponent's hypothesis. For 1 effect there is strong evidence in favor of H_0, and for 1 effect there is extreme evidence in favor of H_0 compared to H_r. Figures S15-18 (bottom) show posteriors for these exploratory Bayes factor analyses. All p-values indicate failed replications.

3.6 Summary of results

Our results show an attenuation in effect size for almost all effects. To illustrate this overall attenuation, Figure 8 shows the posterior probability distributions for all effects under scrutiny. Effect sizes seem to attenuate towards zero, or sometimes even shift to an opposite direction. However, for one effect from Kanai et al. (2012), the effect size is similar to the effect size found in the original study. For this effect our exploratory

analyses indicate successful replications. In addition, three effects from the Westlye et al. (2011) study also show similar effect sizes to the ones found in the original investigation. For these effects, the addition of data could narrow the posterior probability distributions, potentially resulting in a successful replication.

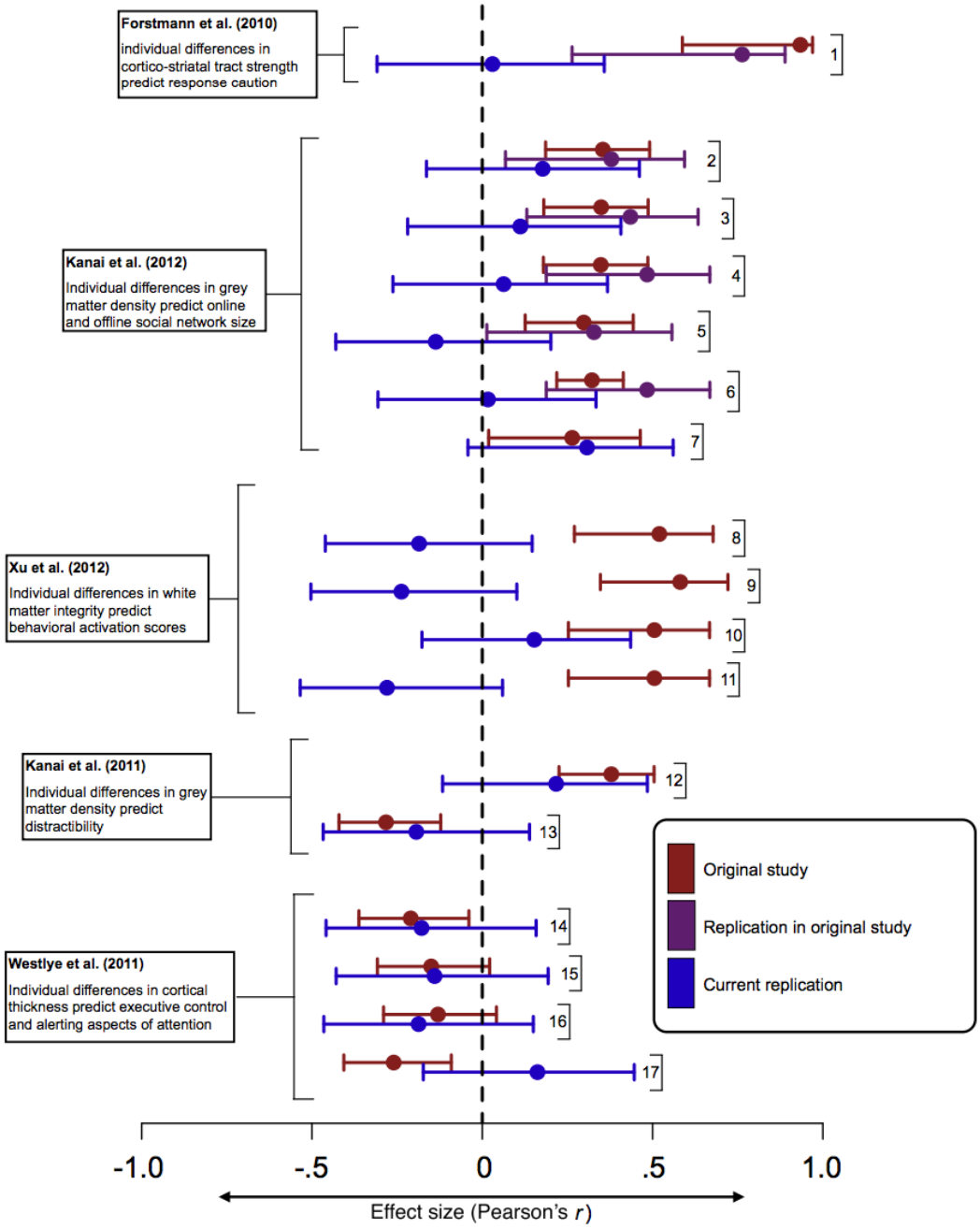


Fig 8. Summary image of our replication results.

95% confidence intervals of posterior probability distributions are shown for the original studies (red), replications within original studies (purple), and the current independent replication attempt (blue). individual effects: (1): LBA flexibility correlated to tract strength between pre-supplementary motor area and striatum. (2-6): FBN correlated to grey matter volume in (2) left middle temporal gyrus, (3) right superior temporal sulcus, (4) right entorhinal cortex, (5) left amygdala, and (6) right amygdala. (7) SNS correlated to grey matter volume in right amygdala. (8) BAS-total correlated to λ1 in left CR and SLF. (9) BAS-FUN correlated to FA in left CR and SLF. (10) BAS-FUN correlated to λ1 in left CR and SLF. (11) BAS-FUN correlated to MD in left SLF and IFOF. (12-13) CFQ correlated to grey matter volume in (12) left superior parietal lobe and (13) left middle prefrontal cortex. (14-16) Executive control correlated to cortical thickness in (14) left caudal anterior cingulate cortex, (15) left superior temporal gyrus, and (16) right middle temporal gyrus. (17) Alerting correlated to cortical thickness in left superior parietal lobe.

4. Discussion

In this study we set out to replicate five experiments showing structural brain-behavior correlations. We adopted a preregistered, purely confirmatory approach so as to avoid common pitfalls in neuroscience such as the use of nonindependent analysis (Vul et al., 2009), double dipping (Kriegeskorte et al., 2009), obscure data collection and analysis which increase false-positive rates (Simmons et al., 2011), and confirmation and hindsight bias on the part of the researcher (Wagenmakers et al., 2012). The five studies we attempted to replicate contained a total of 17 SBB correlations. The results from our confirmatory analyses show that we were unable to successfully replicate any of these 17 correlations. For all but one of the 17 findings under scrutiny, Bayesian hypothesis tests indicated evidence in favor of the null hypothesis. The extent of this support ranged from anecdotal (Bayes factor < 3) to strong (Bayes factor > 10).

Our additional exploratory analyses consisted of computing p-values, and a Bayes factor using an alternative method recently developed by Verhagen and Wagenmakers (2014). This method employs a more specific alternative hypothesis (termed the

proponent's hypothesis), which predicts that the effect size is similar to the effect size of the original finding, rather than just predicting the direction of the effect. This analysis generally provided similar or greater support for the null hypothesis. In addition, 16 out of 17 p-values were higher than threshold (0.05), indicating unsuccessful replications. For one effect in the Kanai et al. (2012), the p-value indicated a successful replication.

In the current replication attempt we aimed to replicate the original experiments as closely as possible. In order to adhere to this plan we adopted a strictly confirmatory framework by publishing a 'Methods and Analysis document' online before any data were inspected or analyzed. This M&A document described all acquisition and analysis plans. After data analysis was complete it became clear that for some analyses, better alternative methods are available. However, the current replication attempt was strictly confirmatory, and thus we choose to (1) not perform these alternative analysis methods, and (2) make the data publicly available[11], so that other researchers might perform these alternative analysis methods instead. It should be noted, however, that these alternative analysis methods can no longer be presented as strictly confirmatory.

Despite our best efforts to replicate the original experiments as closely as possible, this was partly not feasible and partly not desired. Thus, there are a number of deviations from the original study protocols. In the following section, deviations will be discussed with respect to the possibility that they contributed to spurious non-replication (i.e., a failure to detect a true correlation) of the investigated SBB correlations.

1. The sample characteristics of the present replication differed from the sample characteristics in the original studies (e.g., in terms of mean age). This might have led to systematic differences in the behavioral measures. We addressed this issue by correcting our data for age and gender, as was done in most original studies included in our replication attempt. Differences in sample characteristics might still have non-linear effects on our measures, or aging might have differing effects on different brain

[11] *The data set can be freely downloaded from the NITRC Neuroimaging data repository: https://www.nitrc.org/projects/confrep2014/*

regions. Future replication studies could take into account the characteristics of the sample used in the original study, and attempt to match participants in the replication sample to participants in the original sample more closely.

Despite the relevance of this concern, note that in cognitive neuroscience, one often makes claims with regard to a population of humans (i.e., generalizing towards an 'average person'). If the reported effects are indeed non-specific to the sample and its characteristics, there is no reason to assume a priori that a sample with different characteristics impairs our ability to detect the effect. For this reason we chose to acquire data from the current sample, and hypothesize effects as they were described in the original studies. In order to address the concern that (non-linear) effects of differences in sample characteristics might still impair our ability to find these effects, additional research is needed to investigate the specific sample characteristics for which these effects are present.

Similarly, our data differ from the data in the original studies, for instance in terms of the spread of some of the behavioral measures. However, these differences should have little impact on the correlational analyses, since these are not based on the values of the two measures of interest, but on their linear dependence. Only one behavioral measure (i.e., scores on the political orientation questionnaire) did not show enough variance in order to perform a replication attempt.

With respect to sample size, it should be noted that while our sample size was lower than most original studies, our results showed that in our data set, 8 out of a total of 17 hypothesized effects were contradicted with moderate or strong levels of evidence. Thus, even though larger samples are always better than smaller samples from a pre-experimental perspective, our Bayesian post-experimental perspective shows that even

with 36 participants it is possible to obtain informative results[12]. Nevertheless, we encourage additional replication attempts of SBB correlations using larger sample sizes in order to further decrease uncertainty about the replicability of these effects.

2. The MRI data used in the present replication were acquired using a different scanner and with slightly different scanning parameter settings than the MRI data of the original studies. However, recent multi-site reliability studies have shown that these differences have only little impact in both VBM/CT (Jovicich, Marizzoni, Sala-Llonch, Bosch, Bartrés-Faz et al., 2013; Schnack, van Haren, Brouwer, van Baal, Picchioni et al., 2010) and DTI analyses (Fox, Sakaie, Lee, Debbins, Liu et al., 2012).

3. In our TBSS analysis pipeline, another addition to the original protocols is the registration of the ROI spatial maps to our mean FA skeleton. We used spatial maps that were provided by the original authors, and comprised those voxels that correlated with the behavioral measure in the original study. As opposed to using comparably large atlas-based ROI, this approach minimizes the probability that the contribution of a small subset of voxels that correlate with the behavioral measure is canceled out due to averaging across all voxels within the atlas-based ROI. However, in order to be able to use the spatial maps from the original studies we had to register them into the skeleton space common to all participants in our sample. Following the principle of parsimony, we used affine-only (linear) registration with 12 degrees-of-freedom (DoF), which does not guarantee perfect alignment of even the major tracts (Smith et al., 2006). Residual misalignments would be reduced with the use of nonlinear registration. However, such high-DoF alternatives might warp the images so much that the overall structure is not preserved (Smith et al., 2006). It should be noted that, due to the residual misalignments from the linear registration, only a subset of the voxels contained in the registered spatial maps was used in the

---

[12] *In general, it is possible for low-power experiments to yield diagnostic results, and for high-power experiments to yield non-diagnostic results. By conditioning on the observed data, Bayes factors quantify the evidential impact of the information at hand, ignoring hypothetical outcomes that did not occur (Wagenmakers, Verhagen, Ly, Bakker, Lee et al., in press; http://ejwagenmakers.com/inpress/APowerFallacy.pdf).*

correlational analysis. Only voxels, overlapping with the mean FA skeleton were considered. The reduction in the size of ROI would be a concern if we had performed voxelwise statistics (Smith et al., 2006). However, since we aggregated only one value per ROI, it is unlikely that the smaller ROIs have led to spurious non-replication.

On a more general note, software packages may differ slightly in the statistical methods that they employ. These differences can have a relevant impact on the results (e.g., Gronenschild, Habets, Jacobs, Mengelers, Rozendaal et al., 2012; Rajagopalan, Yue, & Pioro, 2014). Our data are publicly available, so that other researchers can carry out additional analyses to probe the robustness of our results. However, such analyses can only be partly confirmatory. Here we restrict ourselves to reporting pre-registered, purely confirmatory analyses performed in FSL (Douaud et al., 2007).

4. The use of Bayesian hypothesis tests for correlations instead of the common null hypothesis significance tests was motivated by two compelling advantages. First, unlike p values the Bayes factor can quantify evidence in favor of the null hypothesis. Second, unlike p-values, Bayes factors do not have the tendency to over-estimate the evidence against the null hypothesis (Edwards, et al., 1963; Sellke, Bayarri, & Berger, 2001; Wetzels, Matzke, Lee, Rouder, Iverson et al., 2011). Note, however, that we have included p-values as exploratory tests. Another deviation concerning the correlational analyses is that we used one-sided instead of two-sided tests, incorporating our prior expectations about the direction of the SBB correlations based on the findings of the original studies. However, this approach provides more compelling evidence (e.g., Hoijtink, Klugkist, & Boelen, 2008; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) and should facilitate replication of true SBB correlations and not contribute to spurious non-replication.

5. While our ROI approach is specific with regard to the location at which we predict the SBB correlation, it does not take anatomical variability between data sets into account. In addition, we extracted the mean signal from the ROIs instead of performing voxel-wise correlations within the ROIs. This process, in combination with

anatomical variability between data sets, introduces noise into the structural measures, potentially concealing the SBB correlation. Future replication work might employ different approaches, which take into account potential anatomical variability, while still making clear predictions with regard to spatial locations of SBB correlations. Note that this point emphasizes the importance of replications within the current field of work. Given that there is random variation in the location of the effect as well as the size of the effect, replication studies are necessary in order to identify the precise location of the effect in addition to the precise effect size.

From the above discussion, one might be tempted to conclude that most of the SBB correlations tested here simply may not exist. However, as previously mentioned, a single replication cannot be conclusive in terms of confirmation or refutation of a finding. We acknowledge the recent replication efforts within the social sciences in general and psychology and neuroscience in particular; an excellent example is the Reproducibility Project of the Open Science Framework (http://opensciceframework.org/) and the first Registered Replication Report (Alogna et al., 2014). Still, to our knowledge, the present replication is the first independent attempt to replicate SBB correlations, despite the considerable number of publications on the matter. We believe that in order to establish correlations between behavior and structural properties of the brain more firmly, it is desirable for the field to replicate SBB correlations, preferably using preregistration protocols and Bayesian inference methods.

Acknowledgements

We would like to acknowledge the authors of the articles we attempted to replicate for sharing their spatial maps, without which we would not have been able to conduct this research.

Literature

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A., et al. (2014). Registered Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9(5), 556*-578.

Banissy, M. J., Kanai, R., Walsh, V., & Rees, G. (2012). Inter-individual differences in empathy are reflected in human brain structure. *NeuroImage*, *62*(3), 2034–2039.

Behrens, T. E. J., Johansen-Berg, H., Woolrich, M. W., Smith, S. M., Wheeler-Kingshott, C. A. M., Boulby, P. A., et al. (2003). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience*, *6*(7), 750–757.

Bickart, K. C., Wright, C. I., Dautoff, R. J., Dickerson, B. C., & Barrett, L. F. (2011). Amygdala volume and social network size in humans. *Nature Neuroscience*, *14*(2), 163–164.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, *12*(12), 4745–4765.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (2011). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, *21*(1), 1–16.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Campbell-Meiklejohn, D. K., Kanai, R., Bahrami, B., Bach, D. R., Dolan, R. J., Roepstorff, A., & Frith, C. D. (2012). Structure of orbitofrontal cortex predicts social influence. *Current Biology : CB*, *22*(4), R123–4.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, 67(2), 319-333.

Cohen, S. (1997). Social Ties and Susceptibility to the Common Cold. *JM&A: the Journal of the American Medical Association*, *277*(24), 1940.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610.

Dale, A. M., & Sereno, M. I. (1993). Improved Localization of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *Journal of Cognitive Neuroscience*, *5*(2), 162–176.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology, 10, 85-103.

Dienes, Z. (2008). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. New York: Palgrave MacMillan.

Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., et al. (2007). Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain : a Journal of Neurology*, *130*(Pt 9), 2375–2386

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193-242.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347.

Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11050–11055.

Fischl, B., Liu, A., & Dale, A. M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Transactions on Medical Imaging*, *20*(1), 70–80.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Fischl, B., Salat, D. H., van der Kouwe, A. J. W., Makris, N., Ségonne, F., Quinn, B. T., & Dale, A. M. (2004a). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23 Suppl 1*, S69–84.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999a). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*(2), 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*(4), 272–284.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., et al. (2004b). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, *14*(1), 11–22.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998), 1541-1543.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., et al. (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.

Fox, R. J., Sakaie, K., Lee, J. C., Debbins, J. P., Liu, Y., Arnold, D. L., et al. (2012). A Validation Study of Multicenter Diffusion Tensor Imaging: Reliability of Fractional Anisotropy and Diffusivity Values. *American Journal of Neuroradiology*, *33*(4), 695–700.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annu Rev Neurosci*, 30, 535-574.

Goldacre, B. (2009). Bad science. London: Fourth Estate.

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J., & Frackowiak, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, *14*(1 Pt 1), 21–36.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046.

Gronenschild, E., H., Habets, P., Jacobs, H., I., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012). The effectof FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One, 7(6).

Groot, A. D. (1969). Methodology: Foundations of inference and research in the behavioral sciences. The Hague: Mouton.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., et al. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). An Introduction to Bayesian Evaluation of Informative Hypotheses (pp. 1–3). New York, NY: Springer New York.

Ioannidis J. P. A. (2005). Why Most Published Research Findings Are False. PloS Med, 2(8).

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, *7*(6), 645–654.

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.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., et al. (2006). Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, *30*(2), 436–443.

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., et al. (2013). Brain morphometry reproducibility in multi-center 3T MRI studies: A comparison of cross-sectional and longitudinal segmentations. *NeuroImage*, *83*, 472–484.

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behavior and cognition. *Nature Reviews Neuroscience*, *12*(4), 231–242.

Kanai, R., Bahrami, B., & Rees, G. (2010). Human Parietal Cortex Structure Predicts Individual Differences in Perceptual Rivalry. *Current Biology*, *20*(18), 1626–1630.

Kanai, R., Bahrami, B., Roylance, R., & Rees, G. (2012). Online social network size is reflected in human brain structure. *Proceedings. Biological Sciences / the Royal Society*, *279*(1732), 1327–1334.

Kanai, R., Carmel, D., Bahrami, B., & Rees, G. (2011a). Structural and functional fractionation of right superior parietal cortex in bistable perception. *Current Biology*, 21(3), 106-107.

Kanai, R., Dong, M. Y., Bahrami, B., & Rees, G. (2011b). Distractibility in daily life is reflected in the structure and function of human parietal cortex. *Journal of Neuroscience*, *31*(18), 6620–6626.

Kanai, R., Feilden, T., Firth, C., & Rees, G. (2011c). Political orientations are correlated with brain structure in young adults. *Current Biology : CB*, *21*(8), 677–680.

Kass, R., E., & Raftery, A., E., (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795.

King, A. V., Linke, J., Gass, A., Hennerici, M. G., Tost, H., Poupon, C., & Wessa, M. (2012). Microstructure of a three-way anatomical network predicts individual differences in response inhibition: A tractography study. *NeuroImage*, *59*(2), 1949–1959.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540.

Lee, M., D., & Wagenmakers, E., J. (2013). Bayesian Modeling for Cognitive Science: A Practical Course. Cambridge: Cambridge University Press.

Lewis, G. J., Kanai, R., Bates, T. C., & Rees, G. (2012). Moral values are associated with individual differences in regional brain volume. *Journal of Cognitive Neuroscience*, *24*(8), 1657–1663.

MacArthur, D. (2012). Methods: Face up to false positives. *Nature*, *487*(7408), 427–428.

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Rajagopalan, V., Yue, G., H., & Pioro, E., P. (2014). Do preprocessing algorithms and statistical models influence voxel-based morphometry (VBM) results in amyotrophic lateral sclerosis patients? A systematic comparison of popular VBM analytical methods. *Journal of Magnetic Resonance Imaging*, 40(3), 662-667.

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *NeuroImage*, 53(3), 1181-1196.

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4), 1402-1418.

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.

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.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Schnack, H. G., van Haren, N. E. M., Brouwer, R. M., van Baal, G. C. M., Picchioni, M., Weisbrod, M., et al. (2010). Mapping reliability in multicenter MRI: Voxel-based morphometry and cortical thickness. *Human Brain Mapping*, *31*(12), 1967–1982.

Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, *22*(3), 1060–1075.

Ségonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Transactions on Medical Imaging*, *26*(4), 518–529.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366.

Sled, J. G., Zijdenbos, A. P., & Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, *17*(1), 87–97.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage*, *31*(4), 1487–1505.

Smith, S. M., Jenkinson, M., Woolrich, M. W., & Beckmann, C. F. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*. 23 Suppl 1, S208-219.

Stileman, E., & Bates, T. (2007) Construction of the Social Network Score (SNS) Questionnaire for undergraduate students, and an examination of the pre-

requisites for large social networks in humans. *Unpublished undergraduate thesis*. See http://hdl.handle.net/1842/2553.

Tuch, D. S., Salat, D. H., Wisco, J. J., Zaleta, A. K., Hevelone, N. D., & Rosas, H. D. (2005). Choice reaction time performance correlates with diffusion anisotropy in white matter pathways supporting visuospatial attention. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(34), 12212–12217.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143*, 1457-1475.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 274–290.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779-804.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (in press). A power fallacy. *Behavior Research Methods.*

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Wallace, J. C., Kass, S. J., & Stanny, C. J. (2002). The Cognitive Failures Questionnaire Revisited: Dimensions and Correlates. *The Journal of General Psychology*, *129*(3), 238–256.

Wallach, H., & O'Connel, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology*, *45*(4), 205–217.

Westlye, L. T., Grydeland, H., Walhovd, K. B., & Fjell, A. M. (2011). Associations between regional cortical thickness and attentional networks as measured by the attention network test. *Cerebral Cortex*, *21*(2), 345–356.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, *6*(3), 291–298.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064.

Wolfe, J. M. (2013). Registered Reports and Replications in Attention, Perception, & Psychophysics. *Attention, Perception, & Psychophysics*, *75*(5), 781–783.

Xu, J., Kober, H., Carroll, K. M., Rounsaville, B. J., Pearlson, G. D., & Potenza, M. N. (2012). White matter integrity and behavioral activation in healthy subjects. *Human Brain Mapping*, *33*(4), 994–1002.

Chapter 3

Challenges in replicating brain-behavior correlations: Rejoinder to Kanai (2015) and Muhlert & Ridgway (2015).

Authors

W. Boekel[1], B.U. Forstmann[1], E.-J. Wagenmakers[1]

[1]University of Amsterdam, the Netherlands

1. Introduction

In a recent study we attempted to replicate five studies that had previously reported a combined total of 17 significant structural brain behavior (SBB) correlations (Boekel et al., 2015). We preregistered our analysis plan and used confirmatory Bayesian hypothesis tests to quantify the evidence that our data provided for the presence or absence of the SBB correlations. For about half of the 17 SBB correlations that we set out to replicate the data suggested at least moderate evidence for their absence, and for 16 out of the 17 correlations the data produced no evidence for their presence. Subsequent exploratory analyses using Bayesian parameter estimation and a Bayesian replication test sketched a more nuanced perspective of the replication results. Nevertheless, our overall results suggest that confirmatory replication studies in the cognitive neurosciences deserve a more prominent role.

Our confirmatory replication has attracted two commentaries from researchers who are skeptical about our results. In “Failed replications, contributing factors and careful interpretations”, Muhlert and Ridgway (2015) critique our replication attempt for having low sample size and incomplete correction for nuisance variables. In addition, they note that there are differences in the VBM processing pipelines used by the original authors and those used in the replication attempt, and suggest that these differences may have contributed towards the discrepant results.

In “Open questions in conducting confirmatory replication studies”, Kanai (2015) also critiques our replication approach and points out that our confirmatory ROI approach may underestimate the SBB correlations. In addition, Kanai feels that the process of refereeing a preregistered study demands clearer guidelines.

We wish to thank the discussants for their interesting suggestions and constructive comments. At the moment little guidance exists with respect to the design and interpretation of purely confirmatory replication studies in the cognitive neurosciences, and we hope this discussion can help stimulate the development of common goals and guidelines.

Below, we discuss the key concerns raised in the commentaries. We also suggest ways in which future replication studies can take into account the issues raised by these commentaries.

2.1. Concern 1: Low sample size

Both Muhlert and Ridgway (2015), and Kanai (2015) point out that the sample size of our replication attempt was lower than the sample sizes of the original findings for 16 out of 17 effects. For 9 out of these 17 effects, our data remained evidentially ambiguous (i.e., $1/3 < BF_{01} < 3$). If we had gathered more data, these replication attempts might have provided us with more evidence, in favor of either hypotheses. Larger sample sizes generally provide a more accurate account of the size and location of the effect that is investigated.

We acknowledged the sample size issue explicitly in our original article: “With respect to sample size, it should be noted that while our sample size was lower than most original studies, our results showed that in our data set, 8 out of a total of 17 hypothesized effects were contradicted with moderate or strong levels of evidence. Thus, even though larger samples are always better than smaller samples from a pre-experimental perspective, our Bayesian post-experimental perspective shows that even with 36

participants it is possible to obtain informative results. Nevertheless, we encourage additional replication attempts of SBB correlations using larger sample sizes in order to further decrease uncertainty about the replicability of these effects." (p. 130)

Wagenmakers, Verhagen, Ly, Bakker, Lee et al. (in press) provide concrete illustrations of situations in which low-powered experiments yield diagnostic results, and situations in which high-powered experiments yield nondiagnostic results. In addition, Wagenmakers, Verhagen, and Ly (in press) show that some real, high-powered data sets can produce evidence that is only anecdotal. The Bayesian bottom line is that a power analysis is useful for planning a study, as it takes into account all possible outcomes that can be expected for an intended sample size. However, once a specific data set is observed the power analysis is inferentially irrelevant, as all that counts is the evidence for the data that have been observed.

In our data, for 8 out of 17 effects under scrutiny, our confirmatory Bayesian test yielded non-anecdotal evidence in favor of the null-hypothesis, despite our relatively modest sample size. Thus, after the data have been observed, all that matters is the evidence. Low samples sizes mean that one can expect the evidence to be inconclusive, but that need not always be the case, and our data demonstrate that something can be learned even when sample size is low.

To provide a different perspective on what our data reveal despite the relatively low sample size, Figure 1 plots effect sizes of the original studies against those of our replication attempt. The blue line represents effect size equality. In general the effects cluster in the area to the right of the line, representing an attenuation of effect sizes in the replication studies. Thus, our results suggest that overall, the effect sizes from our studies are lower than those reported in the original studies.

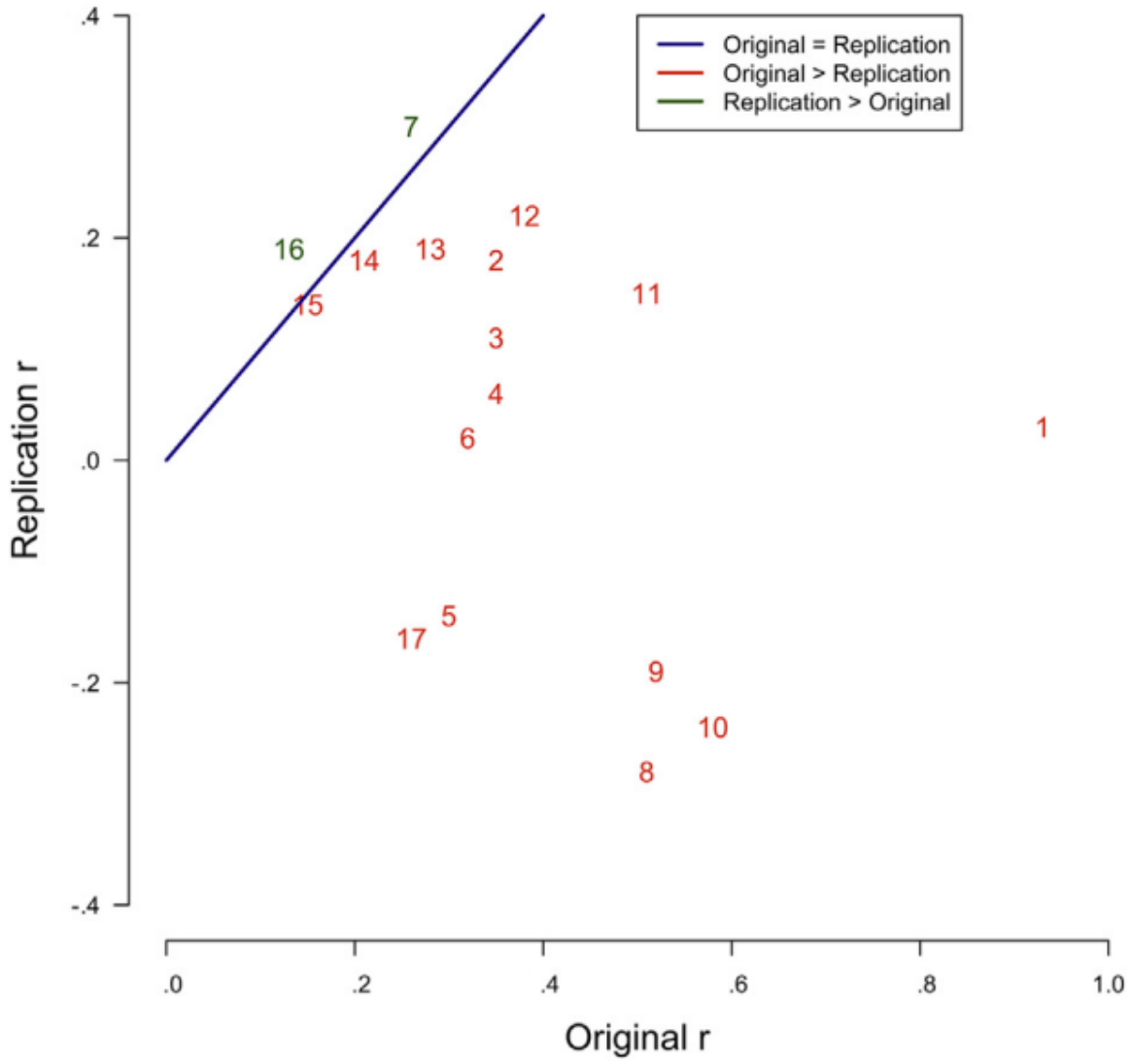


Fig. 1 - Original and replication effect sizes plotted against each other, in order to show the attenuation or amplification of effect sizes. Effects 13 through 17 were flipped in sign for illustration purposes. An effect plotted on the blue line indicates no change in effect size from the original finding to the replication attempt. Effects in green plotted to the left of the blue line indicate amplified effect sizes from the original finding to the replication attempt, whereas effects in red plotted to the right of the blue line indicate attenuation of effect sizes

Muhlert and Ridgway (2015) are concerned that readers might interpret the term “failed replications” to mean that there is compelling evidence for the absence of these effects. When we used the term, we meant to convey the fact that there was no evidence in favor of the presence of the SBB correlations. We agree that the term may be easily misunderstood, and that the ultimate assessment of a replication attempt requires a combination of testing and estimation, coupled with good judgment. The absence of evidence is not the same as evidence of absence, and one of the main advantages of a Bayesian analysis is that it can discriminate between the two possibilities. In our data, for some SBB correlations we find evidence of absence, and for others we find that the evidence is absent. We hope and expect that readers will turn to the concrete results of the Bayes factors and credible intervals to form their own opinion about the extent to which our results constitute a failure to replicate the original findings.

In order to provide a more nuanced perspective on the results from our replication attempts, we reported exploratory replication Bayes factor analyses and plotted posterior distributions for the correlations under scrutiny. These exploratory results can be used to identify potential candidates for further investigation. For example, the exploratory replication Bayes factor analysis of the correlation between social network size (SNS) and grey matter volume (GM) shows moderate evidence in favor of an effect similar to the one found in the original study (Table 1; effect #7). This correlation could be further examined in a new data set, using the combined data of previous findings and replications as a prior distribution for an informed Bayesian hypothesis test. In a similar way, the correlation between executive control and cortical thickness in right MTG (Table 1; effect #16) was larger in our replication sample (-0.19) than in the original sample (-0.13). Despite this, our confirmatory Bayes factor analysis suggests that the data are ambiguous (BF_{01}=1.60) for this SBB correlation. The reason for this is the large difference in sample size between the original sample (n=132) and the replication (n=35) sample. While the point estimate of the correlation might be larger in our replication sample, the posterior probability distribution is wider given the lower sample size, and has more surface area over r = zero (see Figure 8 and Figure S17 in Boekel et al., 2015). The replication Bayes factor for this effect tilts more toward the

alternative hypothesis (BF_{01}=0.65), providing another example for which the addition of data could result in a successful replication. It should be noted that while our exploratory Bayes factor analyses add a layer of nuance to our replication attempt, the general results are similar to our confirmatory analyses. For 9 out of 17 effects, the replication Bayes factor provides non-anecdotal evidence in favor of the null-hypothesis (Table 1).

Table 1. Summary of the results from 17 replication attempts from Boekel et al. (2015). The data show an overall attenuation of effect size. Both confirmatory (BF_{01}) and exploratory (BF_{0r}) Bayes factors suggest non-anecdotal evidence in favor of the null hypothesis for about half of the replication attempts.

Effect #	r_{orig}	r_{rep}	BF_{01}	BF_{0r}
1	0.93	0.03	3.90	180.20
2	0.35	0.18	1.73	1.06
3	0.35	0.11	2.66	2.06
4	0.35	0.06	3.51	3.32
5	0.30	-0.14	7.76	9.56
6	0.32	0.02	4.35	3.88
7	0.26	0.30	0.57	0.27
8	0.51	-0.28	11.74	249.41
9	0.52	-0.19	9.40	170.51
10	0.58	-0.24	10.57	848.06
11	0.51	0.15	2.04	4.13
12	0.38	0.22	1.24	0.73
13	-0.28	-0.19	1.51	0.67
14	-0.21	-0.18	1.71	0.67
15	-0.15	-0.14	2.23	0.81
16	-0.13	-0.19	1.60	0.65
17	-0.26	0.16	8.58	7.70

2.2. Concern 2: Attenuation of effect sizes due to the exploratory nature of discovery

As summarized by Figure 1, our replication attempts yielded a general attenuation of effect sizes. There are several possible explanations for this attenuation. Muhlert and Ridgway point out that given the exploratory nature of many of the original studies, the attenuation of effect sizes in a replication attempt is not surprising. Specifically, the methods used for detecting an effect in an exploratory study may result in an overestimation of the true effect size (Kriegeskorte et al., 2010). This is especially likely in experiments where the sample size is small, such that the effect sizes need to be relatively large in order to pass the classical .05 level of significance. Consequently these effect sizes will likely reduce with subsequent replication attempts. This attenuation, although often a disappointing feature of a confirmatory replication attempt, is a necessary step in identifying the true size of an effect.

2.3. Concern 3: Correction for nuisance variables

Muhlert and Ridgway (2015) suggest that our incomplete correction for nuisance variables is another potential contributor to the attenuation of effect size. Specifically, we corrected the structural brain measures for nuisance variables such as age and sex, but we did not do this for the behavioral measures. This means that if our behavioral measures are correlated with the nuisance variables, our effect sizes may be underestimated. In order to investigate this possibility, we re-computed our results, this time also correcting behavioral data for nuisance variables. Table 2 shows the results. As Muhlert and Ridgway suggested, there seems to be a small overall increase in effect sizes. However, as indicated by the Bayes factors, none of our interpretations were altered in any meaningful way.

Table 2. Partial vs. complete correction for nuisance variables. There is a small general increase in effect sizes after complete correction for nuisance variables. However, Bayes factors were not affected in any meaningful way.

Effect #	Pearson's *r*		BF_{01}	
	Partial nuisance correction	Complete nuisance correction	Partial nuisance correction	Complete nuisance correction
1	.0324	.0679	3.8962	3.2978
2	.1774	.1833	1.7300	1.6571
3	.1118	.1153	2.6645	2.6093
4	.0627	.0646	3.5142	3.4787
5	-.1365	-.1401	7.7605	7.8485
6	.0167	.0171	4.3511	4.3424
7	.3076	.3209	0.5659	0.4905
8	-.2798	-.3030	11.7388	12.3844
9	-.1855	-.1927	9.3958	9.5917
10	-.2374	-.2472	10.5733	10.8397
11	.1528	-.1588	2.0415	1.9576
12	.2169	.2212	1.2437	1.1977
13	-.1937	-.1976	1.5055	1.4591
14	-.1782	-.1783	1.7094	1.7076
15	-.1400	-.1401	2.2306	2.2290
16	-.1869	-.1870	1.6015	1.5997
17	.1621	.2017	8.5804	9.6141

2.4. Concern 4: VBM processing pipeline differences

Muhlert and Ridgway (2015) provide an interesting example of the differences in VBM signal intensity between different preprocessing pipelines, using our replication data (for a recent investigation of pipeline differences in SBB research, see Martinez et al., 2015). Their Figure 2 shows a collection of disconnected regions containing high between-method correlations ($r > 0.8$). This figure unfortunately does not show the entire distribution of correlations across the brain. However, it is safe to say that there are indeed differences between analysis methods, which may have impaired our ability to detect a true effect. It should be noted that the same holds true for the original findings: The difference in analysis methods can potentially also result in an overestimation of an effect size. Because of this, replications are an essential tool to investigate the reliability of empirical findings. Furthermore, by specifying pipelines before data are collected, preregistration reduces analytical freedom, and consequently the rate of false-positives.

2.5. Concern 5: ROI approach potentially leads to underestimation of effect sizes

Figure 1 of Kanai (2015) illustrates the concern that our confirmatory ROI approach might have resulted in an underestimation of effect size. For the convenience of the reader, with permission we have inserted Kanai's figure here (Figure 2). The mechanism underlying this underestimation is the spatial uncertainty that is introduced in the discovery of the effect (Figure 2A). When we define our confirmatory ROIs based on the original findings, there is a chance that due to the spatial uncertainty in the discovery, some voxels of the ROI are in fact not part of the true effect location. Because we extracted the average signal from the entire ROI, this could result in an underestimation of the effect size.

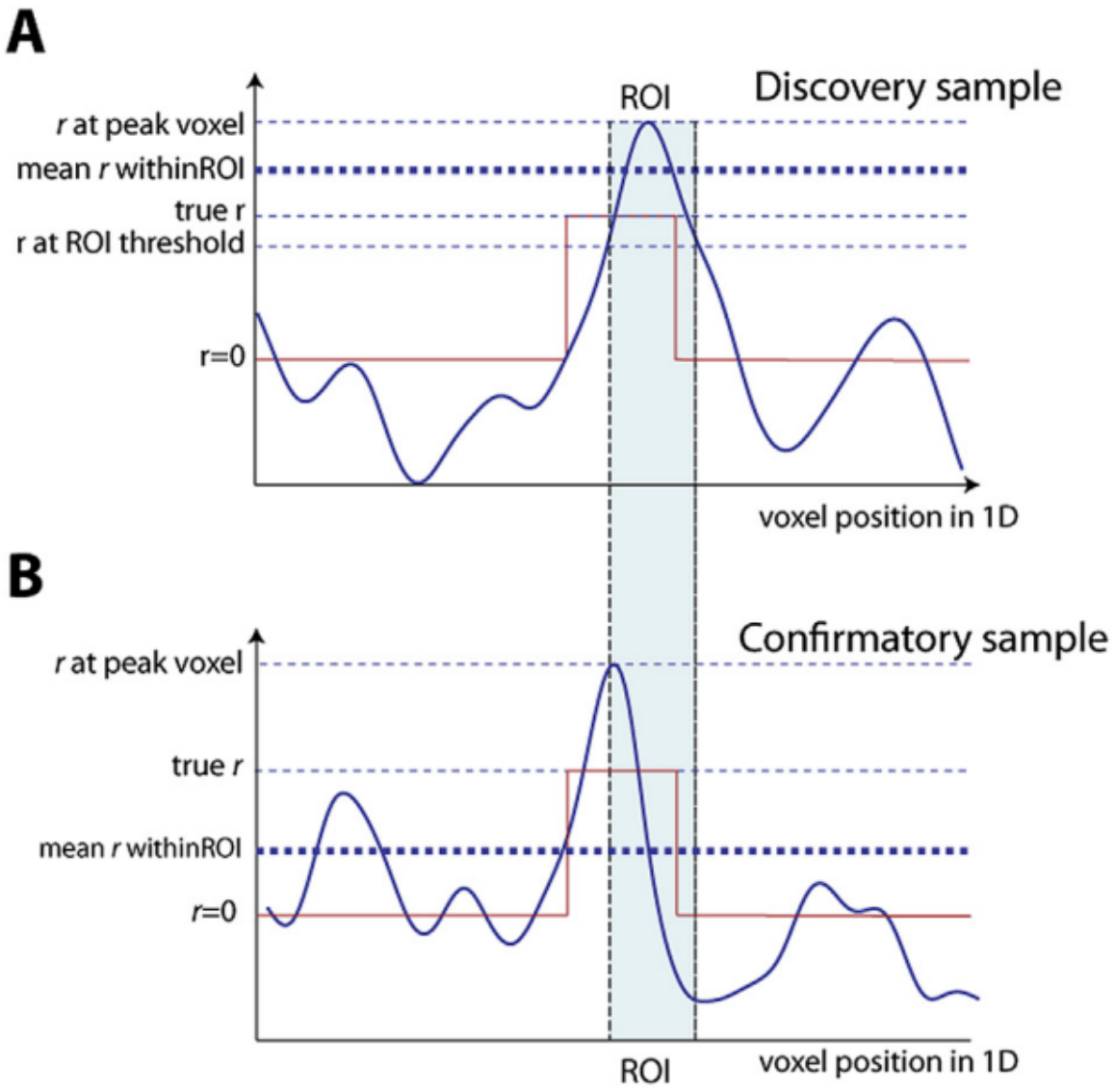


Fig. 2 - Taken from Kanai (2015) with permission, this figure illustrates the issues of over- and underestimation of effect sizes, and uncertainty in effect location. A) An example of a discovery sample. The exploratory nature of the discovery results in an overestimation of the effect size, and uncertainty in terms of the effect location. B) Due to the uncertainty in the effect location and our rigid use of ROIs, the effect size is attenuated.

Note that there are two types of uncertainty in both the discovery sample as well as the confirmatory sample. In the discovery sample, there is often an overestimation of the

effect size due to the exploratory nature of the analysis (Kriegeskorte et al., 2010). In addition, there is uncertainty regarding the exact location of the effect, again due to the exploratory nature of the analysis. In the confirmatory sample, the spatial uncertainty introduced by the discovery sample will often result in an attenuation of the effect size. Our ROI method cannot reduce this spatial uncertainty in the confirmatory sample, as we did not allow the effect to be present in any other area. Future replication attempts should find ways of taking into account the spatial uncertainty introduced by the discovery, while still remaining confirmatory in nature.

These issues certainly have the potential to impair the ability of a replication attempt to detect a true effect. However, they also point out problems of first discoveries: overestimation of effect sizes, and spatial uncertainty in effect locations.

We feel that these problems further emphasize the need for more replication research and meta-analyses. If we were to combine the results from both the discovery sample (Figure 2A) and the confirmatory sample (Figure 2B) we would get a more accurate overall perspective on the true effect size and spatial location of the effect. The conclusion drawn from this figure should not be that our replication underestimates effect sizes, but that replication plays an important part in the scientific process of updating knowledge and determining true effect size and location.

2.6. Concern 6: Review process/alternative analyses

Finally, Kanai (2015) wonders whether reviewers of a preregistered manuscript should be allowed to suggest alternative analyses when reviewing the final manuscript including the results. We feel that while reviewers are certainly allowed to suggest alternative analyses, authors should also be allowed to reject them. In a manuscript that is the result of a preregistered study, this rejection should not impact the decision to accept the manuscript for publication. Instead, this decision should rely on the authors' adherence to the preregistered protocol and the sensible interpretation of their findings. In order to facilitate further exploration of the data set, however, authors should make

their data publicly available. This way, the results of alternative analysis methods can still be published, albeit not in the original paper. Of course, we strongly recommend that a critical re-analysis of a replication data set is also conducted in a purely confirmatory fashion, including a preregistration document posted, for instance, on the Open Science Framework (https://osf.io/). Because the critical re-analysis is already informed by the data, the statistical results are already contaminated to some extent; preregistration prevents the (explicit or implicit) exploration of alternative methods of analysis until the desired result is found.

An intelligent reader will not make up his mind after reading a single paper. Different experiments show different results, and ideally a scientist will conduct research based on the general notions of a field of research, rather than on the outcome of a single study. Because of this, it is not necessary to include all possible alternative analyses in a single paper. Instead, data should be made publicly available, so that other researchers can conduct their own alternative analyses, and potentially publish their results. These alternative analyses should be preregistered or else labeled as exploratory, after which they can be further investigated, possibly by preregistered replication attempts. In this way, we can begin to elucidate the vastly complex patterns of conditions under which particular effects of interest have certain effect sizes and locations.

3. Impact and future directions

Both the results from our confirmatory replication study and the subsequent commentaries suggest that confirmatory replication studies deserve a more prominent role in the cognitive neurosciences. Future replications should optimize their methods in order to increase the accuracy of their replication attempt. Specific to SBB correlations and other neuroimaging findings, spatial uncertainty should be taken into account when performing a replication attempt. In order to mitigate the intrusion of QRPs, alternative analyses which take into account spatial uncertainty should also be preregistered. In order to prevent us from fooling ourselves and having our desires and wishes guide our

statistical reporting we should consistently and clearly indicate the difference between exploratory and confirmatory analyses in our research, and take caution when interpreting exploratory findings, until preregistered replications have confirmed those initial findings.

One way of taking spatial uncertainty into account is presented in Kanai (2015), point 4. In this section, the correlation between CFQ and GM in left SPL is replicated in our data set using a different, less conservative method. This method relies less heavily on the complete ROI identified by the initial finding, as it conducts a voxel-wise test within a restricted ROI based on the peak voxel coordinates of the original finding. By allowing for more freedom in spatial localization of the effect of interest, this method could be used in subsequent replication attempts to take into account the spatial uncertainty that is often introduced when a discovery is made. Kanai points out that this approach has limitations, such as the arbitrary size of the ROI in which the voxelwise tests are conducted, and the inability of this approach to quantify evidence in favor of the null hypothesis. Another way to take into account spatial uncertainty might be to perform a new, explorative voxel-wise test on the combined data from the original study and the replication attempt to identify the region in which both data sets show a significant correlation. This procedure minimizes the potential for the next replication attempt to underestimate the effect size.

Another promising endeavor is the adversarial collaboration (e.g., Matzke et al., 2015). In this approach, proponents and skeptics of a certain discovery decide to work together to design a confirmatory replication attempt of the discovery and agree on a common plan of analysis. Using Bayesian inference, the evidence may be monitored sequentially as the data accumulate, until the evidence is compelling in favor of either hypotheses. Adversarial collaborations can be multi-site endeavors, potentially resulting in much larger sample sizes than what can reasonably be obtained in single-lab studies.

Adversarial collaborations, Bayesian inference, preregistration, and methods for reducing spatial uncertainty together provide a promising starting point for future

replication attempts in the cognitive neurosciences in general, and in structural brain-behavior research in particular.

4. Conclusion

The findings from our replication attempts suggest that results in structural brain-behavior research might not be as reliable as previously thought. The subsequent commentaries of both Kanai (2015) and Muhlert and Ridgway (2015) propose factors that may have contributed to our inability to replicate certain effects. Additional research is needed to investigate these factors in order to provide an accurate account of these (and other) effects in terms of their size, location, and the specific conditions under which they apply. Replication is pivotal in the search for scientific truth. Our confirmatory replication and the subsequent commentaries represent an initial step towards a more reliable and replicable field of research. With this work we hope to stimulate other researchers to undertake similar replication attempts.

References:

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B.U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex, 66, 115-133.*

Kanai, R. (in press). 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.*

Kriegeskorte, N., Lindquist, M. A., Nichols, T. A., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. Journal of Cerebral Blood Flow & Metabolism, 30(9), 1551-1557.

Martinez, K., Madsen, S. K., Joshi, A. A., Joshi, S. H., Román, F. J., Villalon-Reina, J., et al. (2015). Reproducibility of brain-cognition relationships using three cortical surface-

based protocols: An exhaustive analysis based on cortical thickness. *Human Brain Mapping*.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144, e1-e15.*

Muhlert, N., & Ridgway, G.R. (in press). 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.*

Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (in press). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*.

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (in press). A power fallacy. *Behavior Research Methods.*

Chapter 4

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

Authors

Boekel, W.[1,2*], Forstmann, B.U.[1,2], Keuken, M.C.[1,2],

[1]Amsterdam Brain & Cognition Centre, University of Amsterdam, Amsterdam, the Netherlands, [2]Netherlands Institute for Neuroscience, an Institute of the Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands

Abstract

Recent efforts to replicate structural brain-behaviour (SBB) 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 (FA), mean diffusivity (MD), axial diffusivity (AD), and radial diffusivity (RD), 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.

Introduction

Many studies in the cognitive neurosciences aim to investigate the link between brain and behaviour. Recently, researchers have exploited significant advances in diffusion weighted imaging (DWI) to detect subtle differences in brain structure associated with differences in behavioural measures (e.g., Kanai and Rees, 2011) including the stop-signal task (Aron et al., 2007; Forstmann et al. 2012; Rae et al. 2015), conflict tasks such as the Simon task (Forstmann et al., 2008), and strategic decision-making tasks (Coxon et al. 2012; Forstmann et al., 2010; Mulder et al., 2013). In a recent study from

our group using a pre-registered confirmatory framework (Boekel et al., 2015a), we attempted to replicate studies which adopt this structural brain-behaviour (SBB) correlational approach. Confirmatory Bayesian statistical tests suggested that 8 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 Kanai, 2015, Muhlert and Ridgway, 2015; Boekel et al., 2015b).

Previous investigations into the test-retest reliability of diffusion weighted imaging (DWI) have generally suggested stability (Vollmar et al., 2010, Buchanan et al., 2014, Owen et al., 2013, Jovicich et al., 2014, Madhyastha et al., 2014, Heiervang et al., 2006, Jansen et al., 2007, Pfefferbaum 2003, Wang et al., 2012, Fox et al., 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 fronto-occipital 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 behaviour.

For example, Mulder et al., (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 assump-

tion 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 artefacts in the DWI data (Drayer, 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 principle 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 methods section: “probabilistic tractography”), and tract volume is given in mm^3.

We investigate these measures in four white matter pathways: (i) the tract between the subthalamic nucleus (STN) and the inferior frontal cortex (IFC) shown to be involved in stopping behaviour (Aron et al., 2007, Aron et al., 2014), (ii) the inferior fronto-occipital fasciculus (IFOF) involved in the Simon task (Forstmann et al., 2008), (iii) the tract between the striatum (in our analysis comprising putamen and caudate, excluding the nucleus accumbens) and pre-supplementary motor area (pre-SMA), which has been implicated in the speed-accuracy tradeoff (Forstmann et al., 2010), and (iv) the tract between

STN and ventromedial prefrontal cortex (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).

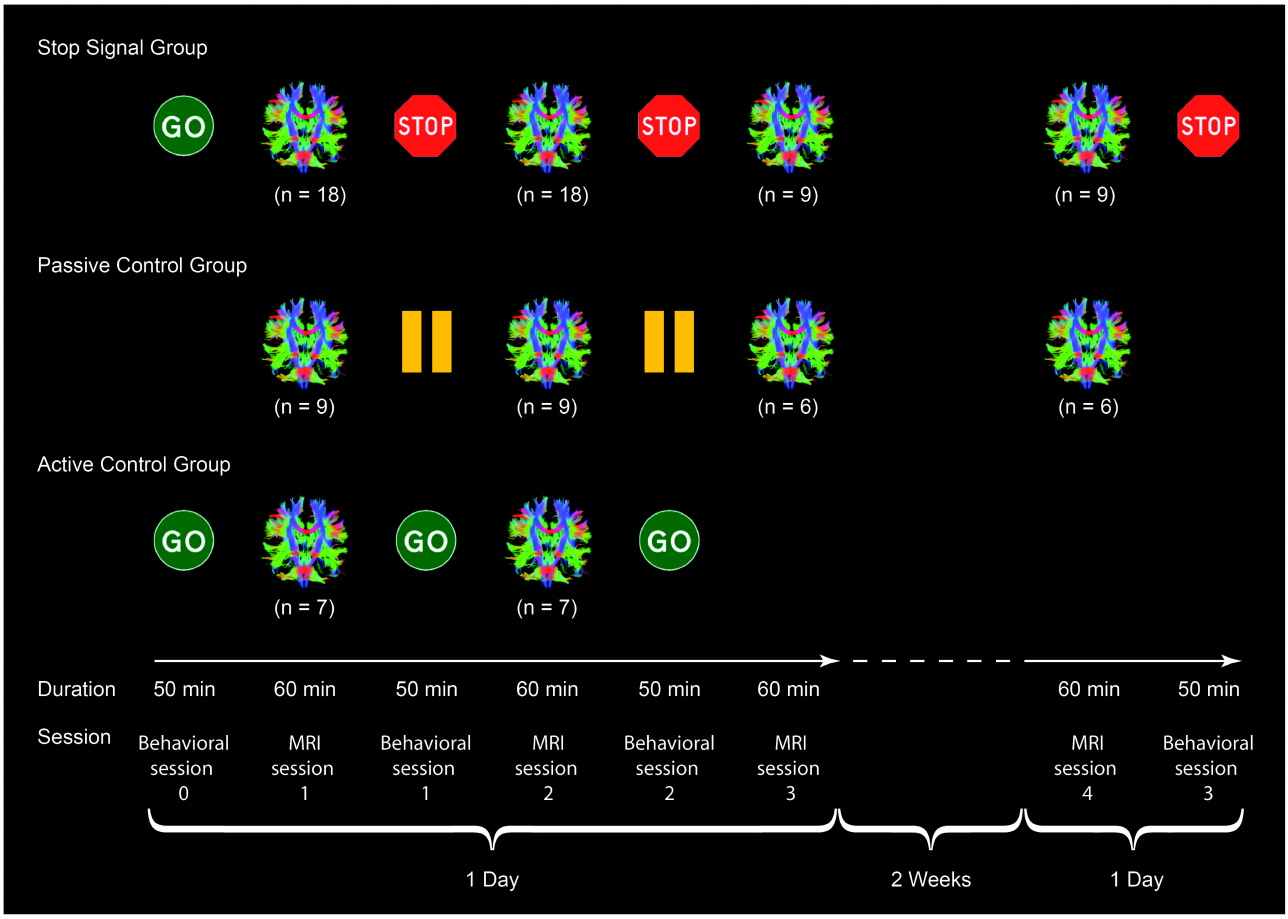
Here we investigate a structural DWI data-set including 34 participants who were scanned in two DWI sessions an hour apart. Out of these 34 participants, 15 were additionally scanned an hour after the second scan session, as well as in a two-week follow-up scan session. In comparison to the majority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

The reliability assessment of DWI measures in the above-mentioned 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 behaviour.

Methods

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 labelled the cognitive control network (Aron et al., 2007, Aron et al., 2014). This pilot study initially comprised data from 15 participants, who were divided into a stop group (9 participants performing the stop-signal task between scans), and a passive control group (6 participants who did not perform the stop-signal task between scans). Later, this data set was expanded to include an active control group of 7 participants, who performed a go task (i.e., the stop-signal task without stop-signals) between DWI scans. Moreover, 9 and 3 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 (s.d. 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 on behavioural stop-signal tasks between scans, and the active control group performed on 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 centre 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 190ms and was updated after every stop-trial by an addition or subtraction of 50ms, 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 RT distributions) software (version 2.0; Matzke et al., 2013). The MCMC sampling settings were number of chains: 3, number of samples: 20000, number of burn-in: 5000, and amount of thinning: 5. All incorrect RT's and RT shorter than 200ms 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 a third session on the same day, as well as in a two-week follow-up session.

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.

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 x 188 x 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 minutes).

In each DWI scanning session, four repetitions of a multi-slice 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 x 224, TR = 7545 ms, TE = 86 ms, SENSE factor (AP) = 2, Bandwidth = 32.1 Hz/Px, acquisition time 5.30 minutes each). Diffusion weighting was isotropically distributed along 32 directions (b-value = 1000 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 non-brain tissue was removed using FMRIB's Brain Extraction Tool (BET; Smith, 2002) to create a brain-mask which was used in subsequent analyses. DTIFIT (Behrens et al., 2003) was applied to fit a tensor model at each voxel of the data (Smith, Jenkinson, Woolrich, & Beckmann, 2004) to derive FA, MD, AD, and RD measures for further analyses.

ROI definition

We extracted the striatum (STR) and STN ROIs from the probabilistic atlas from Keuken et al. (2014). The pre-SMA ROI was drawn in MNI space by using the coordinates reported by Johansen-Berg (2004). The IFC ROI was extracted from the Harvard/Oxford atlas included in FSL (Desikan et al. 2006). The vmPFC ROI was kindly 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 using FLIRT (12 degrees of freedom (dof), correlation ratio, tri-linear interpolation). The registered MNI template was then registered to individual b0 images. We subsequently nonlinearly optimised these transformations using the symmetric image normalization method, which is part of the Advance Normalization Tools (ANTs) (Avants et al. 2008). Using the resulting transformation matrices and warpfields, the ROIs were then transformed into individual space. Figure 2 provides an overview of the resulting ROIs that were subsequently used in probabilistic tractography.

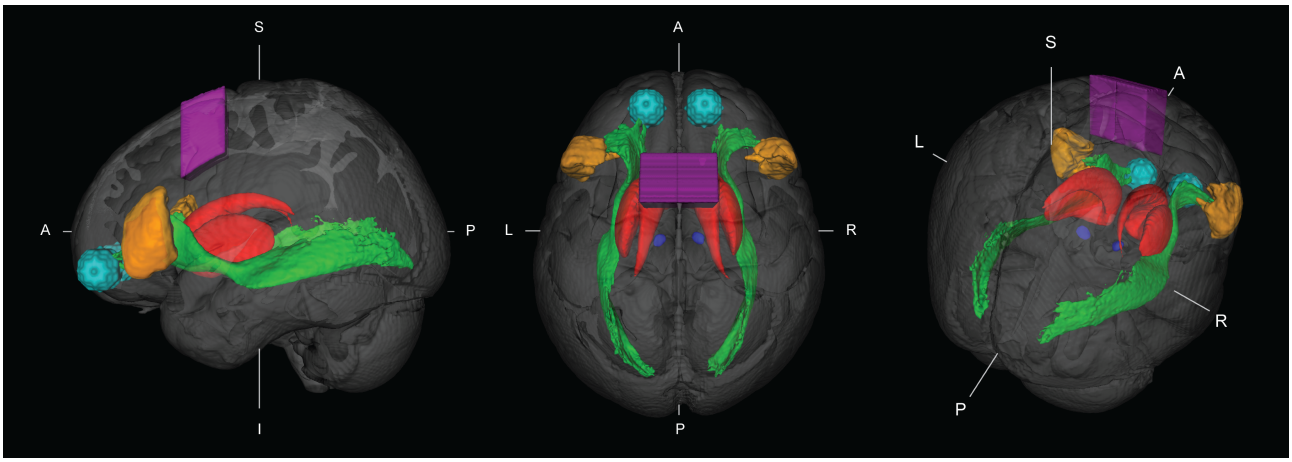


Figure 2. Masks used in probabilistic tractography. Cyan: ventromedial prefrontal cortex (vmPFC), green: inferior fronto-occipital fasciculus (IFOF), ochre: inferior frontal cortex (IFC), purple: pre-supplementary motor area (PreSMA), red: striatum (STR), blue: subthalamic nucleus (STN).

Probabilistic tractography

BedpostX (Behrens et al., 2003) was applied to the preprocessed DWI data to estimate voxel-wise diffusion parameter distributions. Estimation of tract strengths was conducted using probabilistic tractography (Behrens et al., 2003). 50000 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 which 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 i) originated from the seed-mask, ii) did not cross over to the contralateral hemisphere, and iii) 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 non-white 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 in FSL (TBSS; 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 non-linear registration to the FMRIB58_FA standard-space image. Affine registrations were then used to align images into 1x1x1 mm MNI152-space, and finally skeletonised. Subsequently, the mean skeletonised FA image was thresholded at FA of 0.2 (In supplementary tables S1-4 we include an additional analysis using an 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 skeletonised 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 skeletonised. This was done to further shrink and equalize our masks. We subsequently extracted average DWI measures from these tracts and performed an intra-class correlation (ICC) analysis augmented by Bayesian statistical tests to assess stability. We will henceforth refer to this multistep process as our shrinking operator.

Intra-class correlation

The consistency between the different scan session was estimated using the ICC correlation as implemented in the R Package irr (version 0.84, Garner et al. 2012). The ICC is a descriptive statistic that describes the similarity between measurements. We will adopt the labels provided by Cicchetti (1994) where values 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 measure ANOVAs with subjects as a random factor and paired t-tests using the BayesFactor toolbox (version 0.9.12-2; Morey et al., 2015) in R (version 3.0.2; R Foundation for Statistical Computing, http://www.R-project.org). T-tests were run between sessions 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-a-vis the composite alternative hypothesis H_1 (change). For instance, $BF_{10} = 3$ indicates that the observed data are 3 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 5 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 et al. (2015a; 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.

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

Bayes factor BF_{01}			Interpretation
	<	1/100	Extreme evidence for H_1

Results

We start by presenting behavioural findings which suggest that 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 non-white 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 supplementary tables S1-4.

Behaviour

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

Go-trial response times

To assess the between-group behavioural 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 behavioural 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 evi-

dence 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 two-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 hypotheses.

Table 2. Behavioural results split by groups and sessions. Session 0 is the go-practice session performed by participants to familiarize them with the directional decision component of the task.

Group	Session	Go RT	Accuracy	SSRT	P(StopFai)
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)	-	-

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

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
STN - IFC			BF_{01}	ICC	BF_{01}	ICC	BF_{01}	ICC

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
	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

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 measures (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 for 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 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; BF01 > 10) in favor of the absence of a difference over the four sessions included in this test.

Table 4. Reliability of DTI measures in the STN-vmPFC tract thresholded at 10% of robust range. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; VOL: volume; TS: tract strength; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF_{01}: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
STN - vmPFC			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

STR-PreSMA

We delineated a tract between STR and PreSMA 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 measures (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 for 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 all 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. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; VOL: volume; TS: tract strength; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF_{01}: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
Str - PreSMA			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

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
	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

IFOF

Finally, we investigated the reliability of DTI measures in the IFOF. We delineated the IFOF based on a registration method (see methods) 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 for 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 associ-

ated ANOVAs are all higher than 3, suggesting that the evidence is moderately (or in some cases strongly; BF01 > 10) in favor of the absence of a difference over the four sessions included in this test.

Table 6. Reliability of DTI measures in the IFOF thresholded at 10% of robust range. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF_{01}: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
IFOF			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

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, mis-registrations could have resulted in the IFOF ROI overlapping with non-white matter tissue. To illustrate Figure 3 depicts the bilateral IFOF masks in standard

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

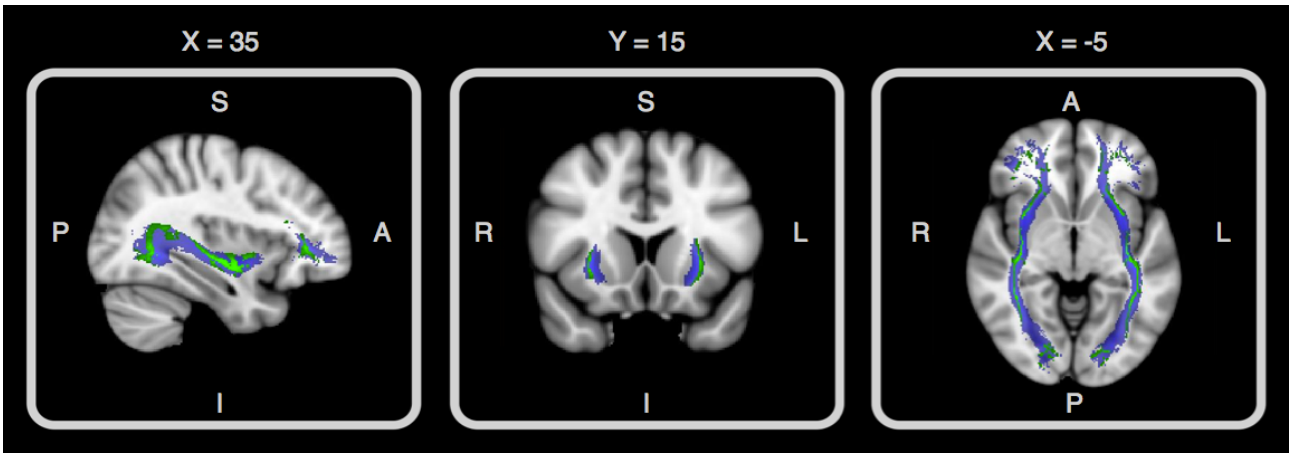


Figure 3. Bilateral IFOF in MNI space. The blue IFOF tract is wide and still shows overlap with non-white matter regions. The green mask represents the IFOF after applying the shrinkage operator.

We decided that more generally, additional shrinkage of all our tracts might mitigate the unwanted influence of non-white matter voxels, thereby further increasing reliability. We performed TBSS (see methods section “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 skeletonised them (i.e., we masked individual tracts with the group white-matter skeleton; see Figure 3 for a visualization of a skeletonised 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 skeletonisation procedure alongside the between-session 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 supplementary tables S1-4.

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 Tables 7 through 10.

Table 7. Reliability of DTI measures in the STN-IFC tract after the shrinking procedure. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF01: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
STN - IFC			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

Table 8. Reliability of DTI measures in the STN-vmPFC tract after the shrinking procedure. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF01: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
STN - vmPFC			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

Table 9. Reliability of DTI measures in the STR-PreSMA tract after the shrinking procedure. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF01: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
STR - PreSMA			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

Tract	Measure	Hemi	T-test S1-2		T-test S1-4		ANOVA	
		R	5.33	0.97	3.69	0.99	7.63	0.99

Table 10. Reliability of DTI measures in the IFOF after the shrinking procedure. FA: fractional anisotropy; MD: mean diffusivity; AD: axial diffusivity; RD: radial diffusivity; ICC: intra-class correlation coefficient; BF01: Bayes factors representing relative evidence in favor of the null-hypothesis.

Tract	Measure	Hemi	S1-2		S1-4		ANOVA	
IFOF			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

We expected to see higher stability in our tracts because the shrinking procedure should solve the problem of overlap between our tracts and non-white 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 non-white matter voxels from our tracts.

Discussion

We set out to assess the test-retest reliability of DTI measures in four tracts which have recently been the subject of structural brain-behaviour (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 ICC's augmented by a Bayesian statistical framework.

Our analyses showed general stability in our initial tracts which 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 grey-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) behaviour 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 to this comparison only provide anecdotal evidence in favor of stabili-

ty. 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 less 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 participants 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 pre-existing 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 (Celik et al., 2015, Wang et al., 2012, Vollmar et al., 2010, Buchanan et al., 2014). 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 which 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 and 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.

Acknowledgements

The work was supported by a Vidi grant from the Dutch Organization for Scientific Research (NWO) (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.

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., Matthews, P. M., Brady, J. M., Smith, S. M. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magn Reson Med, 50(5): 1077-88. doi:10.1002/mrm.10609.

Bennet, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? ANN N Y Acad Sci. 1191:133-155. doi:10.1111/j.1749-6632.2010.05446.x.

Boekel, W., Forstmann, B. U., & Wagenmakers, E.-J. (2015). Challenges in replication brain-behaviour correlations: Rejoinder to Kanai (2015) and Muhlert and Ridgway (2015). Cortex, in press. 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. (2015). Effect of imaging parameters on the accuracy of apparent diffusion coefficient and optimization strategies. Diagn Interv Radiol. 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. http://doi.org/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.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. Cortex, 49:609-610. doi:10.1016/j.cortex.2012.12.016.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & 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):968-80. 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: Palgrave MacMillan.

Drayer, B., Burger, P., Darwin, R., Riederer, S., Herfkens, R., & Johnson, G. A. (1986). MRI of brain iron. AJR, 147:103-110. doi:0361-803x/86/1471-0103.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., Bogacz, R., & Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. PNAS, 107(36):15916-15920. doi:10.1073/pnas.1004932107.

Forstmann, B. U., Keuken, M. C., Jahfari, S., Bazin, P.-L., Neumann, J., Schaefer, A., Anwander, A., & Turner, R. (2012). Cortico-subthalamic white matter tract strength pre-

dicts interindividual efficacy in stopping a motor response. NeuroImage, 60(1):370-375. doi:10.1016/j.neuroimage.2011.12.044.

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.

Fox, R. J., Sakaie, K., Lee, J.-C., Debbis, J. P., Lio, Y., Arnold, D. L., Melhem, E. R., Smith, C. H., Philips, M. D., Lowe, M., & Fisher, E. (2012). A Validation Study of Multicenter Diffusion weighted imaging: Reliability of Fractional Anisotropy and Diffusivity Values. AJNR Am J Neuroradiol, 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.

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., Calabresi, P. A., Pekar, J. J., van Zijl, P. C., & 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.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends in Cognitive Sciences, 18(5):235-241. doi:10.1016/j.tics. 2014.02.010.

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. Proc. Intl. Soc. Mag. Reson. Med., 15:790.

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., Smith, S. M., Higham, D. J., & 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., Wiltfang, J., Roccatagliata, L., Picco, A., Nobili, F., Blin, O., Bombois, S., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J. P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargalló, N., Ferretti, A., Caulo, M., Aiello, M., Ragucci, M., Soricelli, A., Salvadori, N., Tarducci, R., Floridi, P., Tsolaki, M., Constantinidis, M., Drevelegas, A., Rossini, P. M., Marra, C., Otto, J., Reiss-Zimmermann, M., Hoffman, K. T., Galuzzi, S., Frisoni, G. B., & 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., & 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.

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, in press. doi:10.1016/j.cortex.2015.02.020.

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., Sier, R., van der Putten, E. J., Schäfer, Turner, R., & 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., Gilmore, A. W., McDermott, K. B., Nelson, S. M., Dosenbach, N. U. F., Schlaggar, B. L., Mumford, J. A., Poldrack, R. A., & 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: Cambridge University Press.

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. http://doi.org/10.3389/fpsyg.2013.00918

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. Hum Brain Mapp, 35(9):4544-55. doi:10.1002/hbm.22493.

Morey, R. D., Rouder, J. N. Jamil, T. (2015) Computation of Bayes Factors for Common Designs [Computer software]. http://bayesfactorpcl.r-forge.r-project.org/.

Muhlert, N., & Ridgway, G. R. (2015). 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, in press. doi: 10.1016/j.cortex.2015.02.019.

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.

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.

Owen, J. P., Ziv, E., Bukshpun, P., Pojman, N., Wakahiro, M., Berman, J. I., Roberrts, T. P. L., Friedman, E. J., Sherr, E. H., & 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.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. NeuroImage, 40(2):409-414. doi: 10.1016/j.neuroimage.2007.11.048.

Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., Haselgrove, C., Helmer, K. G., Keator, D. B., Marcus, D. S., Poldrack, R. A., Schwartz, Y., Ashburner, J., & Kennedy, D. N. (2012). Data sharing in neuroimaging research. Frontiers in neuroinformatics, 6(9). doi:10.3389/fninf.2012.00009.

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., Ciccarelli, O., Cader, M. Z., Matthews, P. M., & Behrens, T. E. (2006). Tract-based spatial statistics:voxelwise analysis of multi-participant diffusion data. NeuroImage, 31(4):1487-505. doi:10.1016/j.neuroimage.2006.02.024

Smith, S. M., Jenkinson, M., Woolrich, M. W., & Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H. Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage, 23(Suppl 1):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., Duncan, J. S., Richardson, M. P., & 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.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. NeuroImage, 92:381-397. doi:10.1016/j.neuroimage.2014.01.060.

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.

General discussion

This thesis aims to raise awareness of the importance of replication in the cognitive neurosciences. We identified some problems in the literature in terms of reliability of research findings in chapter 1 and 2. In chapter 2 and 3 we provided reflections and discussions regarding our findings. Finally, in chapter 4 we provided an example of the kind of study which can be done to try to identify factors which might lead to decreases in reliability.

More specifically, in Chapter 1 we found that we could not replicate a promising finding of transfer between video-game practice and perceptual decision making. This failure to replicate was eventually made more notable in light of other non-replications in experimental psychology which had recently caused a controversy regarding replicability and reliability (Boekel et al., 2015; Klein et al., 2014; Aarts et al., 2016). In Chapter 2 we performed another replication of SBB findings and used a pre-registration protocol to prevent QRPs and facilitate transparency. The findings of this study sparked discussions in the literature presented in Chapter 3. Finally, Chapter 4 represents an example of how to address some of the issues raised by failed replications and discussions surrounding them.

Overall this thesis emphasises that replication and pre-registration are effective remedies for well-known issues of conventional academic culture; such as QRPs, the file drawer problem, and the lack of statistical power (e.g., low sample sizes). Below I provide examples of recent replication and pre-registration efforts.

Replication

A notable example of a large-scale replication effort is the reproducibility project in psychology, carried out by the open science framework (Aarts et al., 2015). This project started in 2011 and aimed to replicate 100 experimental and correlational studies published in the 2008 issues of three prominent journals in experimental psychology: *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *Journal of Personality and Social Psychology*, and *Psychological Science*. This replication effort focused on direct replication, in the sense that methods and conditions in the replication attempts were made as similar as possible to the methods and conditions of the original to-be-replicated experiments. This aim towards direct replication was facilitated by collaborations between original authors and replicators, the former providing study materials, as well as reviewing the preregistration protocol constructed by the replicators. In total, 270 authors contributed to the 100 replications performed in this project. The overall results of this project showed an attenuation of effect size in the replication compared to the original finding, which is in line with other replication efforts (Boekel et al., 2015; Klein et al., 2014, Wagenmakers et al., in press). However, there was also variation in the sense that some of the replication effect sizes confirmed the original findings, and some did not. The authors argued that no single replication attempt can confirm nor deny the existence of a previously shown effect, and encouraged additional replication efforts. In addition, a similar replication effort in the field of cancer biology is ongoing (Errington, 2014). Large-scale replication efforts, through their impact, raise awareness and appreciation of replication. This could lead not only to more large-scale replication projects, but might also inspire smaller groups to replicate their own work.

Pre-registration

Pre-registration arrived in cognitive neuroscience in 2013 in Cortex as a new publishing initiative: "Registered Reports" (Chambers, 2013). According to this format, authors submit a pre-registration protocol containing an introduction, hypotheses, experimental procedures, analysis pipeline, and a statistical power analysis. Crucially, the pre-registration of hypotheses and (statistical) methods can prevent some QRPs. It is important to note that this does not mean that pre-registered methods can never be deviated from. In the case of deviation from pre-registered analyses, alternate analyses can be labeled 'exploratory', which facilitates further hypothesising and provides a basis for subsequent research. The pre-registered experiment is peer-reviewed prior to data acquisition. Upon acceptance, researchers conduct the experiment according to the plan outlined in their pre-registration protocol. The resulting paper is peer-reviewed, and judged on its adherence to the preregistered protocol and the authors' sensible interpretation of their findings, regardless of the findings themselves. This also means that in this format, null-findings and significant findings are treated equally, creating another incentive for researchers to engage in this format.

By preventing the influence of QRPs, pre-registration can play a promising role in the coming changes towards a more reliable cognitive neuroscience. In 2013, three journals offered the pre-registration format; *Cortex*, *Attention, Perception & Psychophysics*, and *Perspectives on Psychological Science*. Three years later, more journals have joined, or are in the process of adopting the pre-registration format: *Cognition and Emotion*, *Drug and Alcohol Dependence*, *European Journal of Neuroscience*, *AIMS Neuroscience*, *Cognitive Research*, *Experimental psychology*, *Human Movement Science*, *International Journal of Psychophysiology*, *Royal Society Open Science*, and *Stress & Health*. A list of journals offering pre-registration formats can be viewed at https://osf.io/8mpji/wiki/home/. This list is kept up to date and includes editorials and guidelines for specific journals.

By providing pre-registration formats, these journals are incentivising researchers to pre-register their experiments. In addition there is an increased incentive for replica-

tions, because the pre-registration format combines well with replications, especially when done on a large scale (Aarts et al., 2015; Errington et al., 2014). These trends will serve to enhance transparency, reliability, and replicability of cognitive neuroscience research findings.

Conclusion

In this thesis I focused on the importance of replication and preregistration in cognitive neuroscience, and show how they can be used as tools to increase reliability of research findings. I discussed how these tools are already being used, in efforts such as the recently finalised reproducibility project in psychology (Aarts et al., 2015) and the ongoing similar effort in cancer biology (Errington et al., 2014), as well as the increasing number of journals adopting the pre-registration format. These efforts (and others sure to come) lead to a more realistic view of the current reliability of conventional scientific practice, thereby facilitating the continued self-improvement process of the scientific endeavour.

On a final note, it appears that the discussed efforts emerged because a large enough number of researchers were inspired to increase the reliability and reproducibility of scientific output. Their working together, and their combined resources allowed the large-scaleness of these projects. It is not unreasonable to argue that; if a larger number of researchers become similarly inspired to increase the reproducibility of scientific output and subsequently work together on large-scale collaborative projects, the scientific endeavour at large may gradually become more reliable. How might this process occur? What can an individual scientist do to remedy this ‘crisis of confidence’? More precisely, given the established academic culture, which measures can be taken by a single bachelor, master, PhD student, post-doc, or professor, to contribute to this shift towards a more reliable science? In the widely popular science tv series “Cosmos: A Spacetime Odyssey”, astrophysicist Neil deGrasse Tyson argues that five simple rules have lead humanity to its scientific discoveries and technological achievements:

Only a few centuries ago, a mere second in cosmic time, we knew nothing of where or when we were. Oblivious to the rest of the cosmos, we inhabited a kind of prison, a tiny universe bounded by a nutshell.

How did we escape from the prison? It was the work of generations of searchers who took five simple rules to heart.

***(1) Question authority**. No idea is true just because someone says so, including me.*

***(2) Think for yourself**. **Question yourself**. Don't believe anything just because you want to. Believing something doesn't make it so.*

***(3) Test ideas by the evidence gained from observation and experiment**. If a favorite idea fails a well-designed test, it's wrong. Get over it.*

***(4) Follow the evidence wherever it leads**. If you have no evidence, reserve judgment.*

And perhaps the most important rule of all...

***(5) Remember: you could be wrong**. Even the best scientists have been wrong about some things. Newton, Einstein, and every other great scientist in history -- they all made mistakes. Of course they did. They were human.*

Science is a way to keep from fooling ourselves, and each other.
(Braga et al, 2014; Episode 13)

References introduction and discussion

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P.R., Attwood A., Axt, J. Babel, M. et al. (2015). Estimating the reproducibility of psychological science. Science 349(6251).

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. Social cognitive and affective neuroscience, 4(4), 417-422.

Boekel, W., Forstmann, B. U., & Keuken, M. C. (in press). A test-retest reliability analysis of diffusion measures of white matter tracts relevant for cognitive control. Psychophysiology.

Boekel, W., Wagenmakers, E. J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. Cortex, 66, 115-133.

Boekel, W., Forstmann, B. U., & Wagenmakers, E. J. (2016). Challenges in replicating brain-behavior correlations: Rejoinder to Kanai (2015) and Muhlert and Ridgway (2015). cortex, 74, 348-352.

Braga, B., Cannold, M., Clark, J., Dolleman, E., Druxman, A., Druyan, A., Hanich, L., Holtzman, S., Kirr, S., MacFarlane, S., Micucci, J.J., Robertson, P., Bryson, C.S., McKinnon, M., Barry, K., Berry, D., Butler, A., Courtney, K.M., Oreck, S., Sweatman, C., & Vallow, K. (Producer), Braga, B., Druyan, A., Pope, B., & Dart, K. (Directors). 2014. *Cosmos: A Spacetime Odysse*y [TV Mini-Series]. United States: Cosmos Studios, Fuzzy Door Productions, Santa Fe Studios.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365e376.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. Cortex, 49(3), 609e610.

Cohen, J. (1994). The Earth Is Round ($p < .05$). American Psychologist, 49(12), 997-1003.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. Elife, 3, e04333.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012, August 27). Correcting the Past: Failures to Replicate Psi. Journal of Personality and Social Psychology. Advance online publication. doi: 10.1037/a0029709

Green, C. S., & Bavelier, D. (2012). Learning, attentional control, and action video games. Current biology, 22(6), R197-R206.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PloS Med, 2(8).

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011) Data replication & reproducibility. Again, and again, and again .... Introduction. Science 334(6060):1225.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science, 0956797611430953.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., … & Cemalcilar, Z. (2014). Investigating variation in replicability. Social Psychology.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research how often do they really occur?. Perspectives on Psychological Science, 7(6), 537-542.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E. J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. Journal of Experimental Psychology: General, 144(1), e1.

Mumford, J. A. (2012). A power calculation guide for fMRI studies. Social cognitive and affective neuroscience, 7(6), 738-742.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence?. Perspectives on Psychological Science, 7(6), 528-530.

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. Nature neuroscience, 17(11), 1510-1517.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638e641.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 0956797611417632.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., ... & Tetzlaff, J. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. Bmj, 343, d4002.

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. Journal of Experimental Psychology: General, 143(5), 1794.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems ofp values. Psychonomic bulletin & review, 14(5), 779-804.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupiáñez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Özdoğru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., Zwaan, R. A. (in press). Registered Replication Report: Strack, Martin, & Stepper (1988). Perspectives on Psychological Science.

Wagenmakers, E. J., & Forstman, B. U. (2014). Rewarding high-power replication research. Cortex, 51(10).

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American Psychologist, 61(7), 726.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PloS one, 6(11), e26828.

Research funding

The work was supported by a Vidi grant from the Dutch Organization for Scientific Research (NWO) awarded to Prof. Dr. Birte U. Forstmann.

Summary English: "On the importance of replicating research findings"

This thesis aims to raise awareness of the importance of replication in the cognitive neurosciences. We identified some problems in the literature in terms of reliability of research findings in chapter 1 and 2. In chapter 2 and 3 we provided reflections and discussions regarding our findings. Finally, in chapter 4 we provided an example of the kind of study which can be done to try to identify factors which might lead to decreases in reliability. This research has lead to some recommendations in terms of how future research might investigate and remedy the crisis of confidence currently experienced in cognitive neuroscience. Specifically, we propose that the pre-registered replication provides researchers with an excellent tool to simultaneously investigate reliability, while also preventing common Questionable Research Practices (QRPs) in a transparent way. This is done by making publicly available all methods and analysis plans of the replication prior to data acquisition. The subsequent execution of the replication project can be performed with very little opportunities for QRPs to interfere. Recently there have been some developments in terms of large-scale pre-registered replication efforts (Aarts et al., 2015, Errington et al., 2014), which leads me to optimistically suggest that the field is gradually increasing its reliability.

Samenvatting Nederlands: "On the importance of replicating research findings"

Dit proefschrift heeft als doel om licht te schijnen op de cruciale rol van replicatie in onderzoek binnen de cognitieve neurowetenschappen. In hoofdstuk 1 en 2 beschrijven we problemen in de literatuur met betrekking tot betrouwbaarheid van onderzoeksresultaten. In hoofdstuk 2 en 3 reflecteren en bediscussiëren we deze bevindingen. In hoofdstuk 4 beschrijven we ten slotte een voorbeeld van een onderzoek gericht op het uitzoeken van welke factoren invloed kunnen hebben op de eerder geobserveerde lage betrouwbaarheid van onderzoeksresultaten. Dit onderzoek heeft tot enkele aanbevelingen geleid voor toekomstig onderzoek. Om het aanzien van de betrouwbaar-

heid van onderzoeksresultaten te herstellen kunnen onderzoekers replicatie-studies pre-registreren voordat ze uitgevoerd worden. Door dit te doen worden alle onderzoekshypothesen en geplande analyses publiek toegankelijk, en kunnen “Questionable Research Practices” (QRPs) voorkomen worden. Recentelijk zijn er grote ge-pre-registreerde replicatie-onderzoeken uitgevoerd (Aarts et al, 2015, Errington et al., 2014) die als voorbeeld gezien kunnen worden voor toekomstige grootschalige onderzoeken. Door deze ontwikkelingen lijkt het er op dat het onderzoeksveld zichzelf aan het corrigeren is, waardoor de betrouwbaarheid van onderzoeksresultaten uiteindelijk hersteld kan worden.

General acknowledgements

I thank my supervisors and colleagues from the lab, as well as those from other labs and faculties, for all the help and advice they gave me during my PhD project. In addition I thank my family, specifically my parents and sisters, as well as all my friends, for helping me through these difficult times.