Table of Contents

Proceedings of the Master's Programme Cognitive Neuroscience of the Radboud University

Contact Information:
Journal CNS
Radboud University
Postbus 9104
6500 HE Nijmegen
The Netherlands
nijmegencns@gmail.com

From the Editors-in-Chief

Dear Reader,

In your hands you are holding the second issue of the 12th volume of the *Proceedings of the Master's Programme Cognitive Neuroscience*. We are proud to present a selection of the most interesting research done by alumni of our programme. The group of graduates that submit their theses to the second issue is comparatively smaller than for the first issue each year. The journal team has therefore decided to publish four articles this time to guarantee high quality content throughout the issue.

The Master's Programme Cognitive Neuroscience has always been about uniting different disciplines with the common goal of expanding our knowledge of the brain and its capabilities. This is well-reflected in our cover: We work on understanding how we make sense of the incoming information when reading a book in our Language and Communications specialization. We program complex software on microchips to simulate the interaction of neurons in our Brain Networks and Neuronal Communications specialization. We investigate how our perception can be tricked by optical illusions in the Perception, Action and Control specialization. We navigate through mazes to unravel how our brain manages to not get lost in the buzzing world around us in our Plasticity and Memory specialization. These examples are only the tip of the iceberg of what we contribute to science every day.

This issue concludes our time as Editors-in-Chief and we wish our team the best of luck moving forward. All that is left to say now is to thank you for your continued interest. We hope you enjoy reading the interesting articles!

Nijmegen, July 2017

Kim Fricke & Yvonne Visser

Editors-in-Chief

Neural Mechanisms of Action-Selective and Stimulus-Selective Stopping

Ruben van den Bosch[1,2]
Supervisors: Bram Zandbelt[1,2] , Roshan Cools[1,2]

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands
[2]Radboud University Medical Centre Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands

The past decade has seen a surge of interest in selective stopping. Researchers studying selective stopping have relied on the independent race model of simple stopping. Further- more, they have investigated selective stopping with a heterogeneous set of tasks, including action-selective and stimulus-selective stop tasks. Action-selective stop tasks probe control of specific actions and stimulus-selective stop tasks examine control triggered by specific stimuli. However, it remains unclear whether the independent race model extends to selective stopping and whether selective stopping is a homogeneous or heterogeneous construct. Here, we addressed these important gaps. We tested whether selective stopping performance is in agreement with predictions of the independent race model, using a Bayesian hypothesis testing approach based on the Bayes factor. We performed these tests at the group- and individual-level. We then compared action- and stimulus-selective stopping in terms of performance and brain activation, using functional magnetic resonance imaging.

We found violations of the predictions of the independent race model in 91% of the individuals in action-selective stopping and 74% of the individuals in stimulus-selective stop- ping. These individual violations were almost completely masked by the group performance. Furthermore, performance did not differ between the two selective stopping types and there appeared to be no differences in inhibition-related brain activity.

These results suggest that the independent race model does not generally extend to selective stopping and that action-selective and stimulus-selective stopping form a homogeneous construct.

Keywords: cognitive control, race model, response inhibition, selective stopping, functional MRI

Corresponding author: Ruben van den Bosch; E-mail: r.vandenbosch@donders.ru.nl

The ability to inhibit intended actions in response to environmental changes is an essential act of control in daily life and central to adaptive human behaviour. A drastic form of cognitive control is complete inhibition.

For decades, this has been studied with the simple stop-signal paradigm (Verbruggen & Logan, 2008). In the simple stop-signal task subjects make quick responses to go-stimuli but try to cancel the response when an infrequent stop-signal occurs after a variable delay (stop-signal delay; t_d). The trials on which such a stop-signal occurs can be divided based on whether stopping succeeds (stop-inhibit trials) or fails (stop-respond trials). Stopping performance on this task is characterised by three main findings: (1) the probability of responding (P_r) given a stop-signal increases with t_d ; (2) stop-respond response time (RT) is shorter than no-signal RT; (3) stop-respond RT increases with t_d .

The independent race model provides a theoretical framework from which these findings can be understood (Logan & Cowan, 1984; Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014). This model explains stopping performance as the outcome of a race between a GO process that executes the response and a STOP process that cancels it. If the GO process finishes before the STOP process, the response is executed; if the STOP process finishes before the GO process, the response is canceled. Under these assumptions the model predicts exactly the pattern of findings observed in the standard stop-signal task. Besides a theory of simple stopping, the independent race model provides methods to estimate the latency of the covert STOP process, known as the stop-signal reaction time (SSRT). The SSRT can be estimated from the proportion of stop-respond trials and the distribution of no-signal RTs.

The independent race model has stimulated extensive use of the stop-signal task in various fields of research. This has greatly furthered our understanding of, for example, the lifespan development of control (Coxon, Impe, Wenderoth, & Swinnen, 2012; Van de Laar, Van den Wildenberg, Van Boxtel, & Van der Molen, 2011), response inhibition itself (Logan, 1994; Verbruggen & Logan, 2008) and clinical and neurological disorders, including attention deficit hyperactivity disorder (ADHD; Dimoska, Johnstone, Barry & Clarke, 2003; Janssen, Heslenfeld, Van Mourik, Logan, & Oosterlaan, 2015), Schizophrenia (Thakkar, Schall, Boucher, Logan, & Park, 2011; Zandbelt, Buuren, Kahn, & Vink, 2011) and Parkinson's Disease (Gauggel, Rieger, & Feghoff, 2004; Van de Wildenberg et al., 2006).

Neuroscience studies have demonstrated that simple stopping manifests in the motor system and also involves areas in the frontal lobe and basal ganglia. Neurophysiology studies in animals have shown that eye movement-related activity of neurons in the frontal eye field (Hanes, Patterson, & Schall, 1998) and superior colliculus (Paré & Hanes, 2003) as well as limb movement-related activity of neurons in the dorsal premotor cortex (Mirabella, Pani, & Ferraina, 2011) and basal ganglia nuclei (Schmidt, Leventhal, Mallet, Chen, & Berke, 2013) decays in response to the stop-signal within the time required to cancel the movement (Schall & Boucher, 2007). Transcranial magnetic stimulation studies in humans have shown similar dynamics for primary motor cortex excitability (Coxon, Stinear, & Byblow, 2006; Van de Wildenberg et al., 2009). Human imaging, stimulation, and lesion studies suggest that simple stopping relies on the inferior frontal cortex (Aron, Fletcher, Bullmore, Sahakian, & Robbins, 2003; Aron & Poldrack, 2006; Chambers et al., 2006; Verbruggen, Aron, Stevens, & Chambers, 2010), pre-supplementary motor area (Chen, Muggleton, Tzeng, Hung, & Juan, 2009; Li, Huang, Constable, & Sinha, 2006), and basal ganglia structures such as the striatum (Zandbelt, Bloemendaal, Hoogendam, Kahn, & Vink, 2012; Zandbelt & Vink, 2010) and subthalamic nucleus (Aron & Poldrack, 2006; Jahfari et al., 2011; Van de Wildenberg et al., 2006).

Notwithstanding the deep insights this research has yielded, simple stopping is limited as a model of cognitive control in daily life and psychiatric disorders. Simple stopping evokes control that is non-selective; after the stop-signal subjects have to stop all their planned actions. In reality, most situations require selective stopping, a more flexible form of control. It comprises control that is targeted at specific actions or triggered by specific stimuli and has been studied with selective stopping tasks. Selective stopping research has been suggested to not only have greater ecological validity, but also greater clinical relevance (Aron, 2011). However, two main factors currently complicate the interpretation of data from selective stopping research.

Firstly, although selective stopping research has relied on the independent race model, it is uncertain whether the model extends from simple stopping to selective stopping. One reason is that studies to date have reported tests of the predictions of the independent race model incompletely or not at all. Another reason is that the available data on tests of the predictions provide mixed evidence. For example, although at least one selective stopping

study showed that all three predictions of the race model held (Smittenaar, Guitart-Masip, Lutti, & Dolan, 2013), many others reported violations of at least one of the predictions (Bissett & Logan, 2014; De Jong, Coles, & Logan, 1995; Dimoska, Johnstone, & Barry, 2006; Van de Laar, Van den Wildenberg, Van Boxtel, & Van der Molen, 2010; Verbruggen & Logan, 2015) in as many as 61% of participants (Bissett & Logan, 2014). A final reason is that tests of the race model's predictions are often performed for the group as a whole rather than for each individual separately, which may mask violations occurring in a subset of individuals. To illustrate, one selective stopping study reported that one of the predictions (stop-respond RT should be shorter than no-signal RT) held for the group as a whole, but also reported that the very same prediction was violated in 32% of the participants (Sebastian et al., 2015).

Together, this state of affairs is unsatisfactory, because if it turns out that the independent race model does not extend generally from simple stopping to selective stopping, then SSRT estimates reported by previous selective stopping studies may be invalid and conclusions derived from them may be flawed. To address this problem, a systematic investigation of predictions of the independent race model across different forms of selective stopping and performed at the individual level is necessary. This will help clarify how often violations of predictions of the independent race model occur.

Secondly, it is unclear whether selective stopping is a homogeneous or a heterogeneous construct. As pointed out by Bissett and Logan (2014), selective stopping research has used a heterogeneous set of tasks, yet all of them are called selective stopping, as if selective stopping is a homogeneous construct. In some tasks (e.g., Aron & Verbruggen, 2008; Coxon et al., 2016; Coxon, Stinear, & Byblow, 2009; Majid, Cai, Corey- Bloom, & Aron, 2013; Smittenaar et al., 2013) participants are instructed to stop certain actions (e.g., a left-hand response), while continuing others (e.g., a right-hand response). We call this action-selective (AS) stopping (note that Bissett and Logan (2014) have called this unconditional motor-selective stopping). In other tasks (e.g., Bissett & Logan, 2014; Dimoska et al., 2006; Van de Laar et al., 2010; Ruiter, Oosterlaan, Veltman, Van den Brink, & Goudriaan, 2012; Sebastian et al., 2015; Sharp et al., 2010; Verbruggen & Logan, 2015) participants are instructed to stop to certain signals, while ignoring others. We call this stimulus-selective (SS) stopping. It remains unclear whether AS and SS stopping tap into the same form of control, as these tasks have never been compared directly.

On the one hand, several findings seem to suggest that they do involve the same form of control, including SSRTs that are in the same range, response strategies that show striking resemblances (Bissett & Logan, 2014; Macdonald, Stinear, & Byblow, 2012), and activation in seemingly similar brain regions, such as ventrolateral frontal cortex, dorsal frontal cortex, and basal ganglia (Coxon et al., 2016, 2009; Majid et al., 2013; Ruiter et al., 2012; Sebastian et al., 2015; Sharp et al., 2010; Smittenaar et al., 2013). On the other hand, violations of the independent race model have mainly been reported for SS stopping (Bissett & Logan, 2014; Dimoska et al., 2006; Van de Laar et al., 2010; Verbruggen & Logan, 2015) and rarely for action- selective stopping (De Jong et al., 1995). Moreover, AS stopping may involve at least two cognitive steps before the inhibition of a response is initiated (discriminating the signal and selecting the action to be canceled) and SS stopping may involve only one (discriminating the signal). Consequently, AS stopping may rely more heavily on motor-related brain regions, such as the pre-supplementary motor area (pre-SMA), the supplementary motor area (SMA), the dorsal premotor area (PMd), and possibly the basal ganglia. To address this problem, a direct comparison of AS and SS stopping tasks in terms of behavioural and neural measures of stopping is necessary.

In the present functional magnetic resonance imaging (fMRI) experiment, we used a task with AS and SS stopping intermixed, allowing us to compare these forms of selective control both behaviourally and neurobiologically. We tackle the problems described above by addressing two research questions:

1. Does the independent race model extend to selective stopping?
2. Is selective stopping a homogeneous or heterogeneous construct?

If the independent race model does not extend to selective stopping, we would expect to find that stopping performance would violate any of the three qualitative predictions of the race model. Alternatively, if the independent race model does extend to selective stopping, we would expect to find that stopping performance is in line with all three of the model's qualitative predictions.

If selective stopping is a homogeneous construct, then we would expect that AS and SS stopping would not differ in terms of stopping performance and brain activation measures of selective stopping. Alternatively, if selective stopping

is a heterogeneous construct, then we would expect that AS and SS stopping would differ in terms of stopping performance and brain activation measures of selective stopping.

Methods

Pre-registration

This project has been pre-registered at the Open Science Framework (www.osf.io) on May 30, 2016. In the pre-registration, all the methods, procedures, outcome measures and confirmatory analyses are described in detail. Additional, not pre-registered, analyses were exploratory, rather than confirmatory, and are indicated as such in the text below. The document is available on request and it will be made publicly accessible upon publication of this research. At the time of pre-registration, four datasets had already been collected, but not analysed.

There were four deviations from the pre-registration. Firstly, 24 subjects were included for this thesis, rather than the pre-registered 30. Before publication of this project as a research article, additional subjects will be scanned until thirty subjects are included in the analyses. Secondly, we used a Bayesian repeated-measures ANOVA instead of Bayesian logistic regression in the analysis of the effect of t_d on P_r (see 'Tests of independent race model predictions' section). Thirdly, in the analysis of the effect of t_d on stop-respond RT we added a restriction to the Bayesian repeated-measures ANOVA model that was not in the pre-registration (see 'Tests of independent race model predictions' section). Fourthly, in the fMRI analysis, only six motion regressors were used instead of the pre-registered twenty-four see 'Functional MRI analyses' section).

Participants

Twenty-five healthy participants volunteered for this study. One participant was excluded after the practice session (see 'Practice session' section), bringing the number of participants to twenty-four (mean age 23.8 years, range 18-33; 17 females). All participants had normal or corrected to normal vision and did report no history of neurological or psychiatric illness or claustrophobia. Written informed consent was obtained from all participants. The study procedures were in accordance with the Declaration of Helsinki and have been approved by the local Institutional Review Board (Committee on Research Involving Human Subjects Arnhem-Nijmegen, registered under CMO2014/088).

Task

We used a mixed AS and SS stop task (Fig. 1). All stimuli were presented in the centre of the screen on a grey background. The fixation stimulus was a white cross, subtending 3° along its vertical axis. The primary (go) stimulus was a white Hiragana character, subtending 6° along its vertical axis. On each trial it was chosen from a set of two. The secondary stimulus was a playing card suit symbol subtending 3° along the vertical axis and presented on top of the primary stimulus at 80% opacity. It was chosen from a set of four: an orange diamond, a cyan heart, a yellow spade, or a purple club.

Each trial started with the presentation of the fixation stimulus for 200 ms. The fixation cross was immediately followed by the primary go stimulus, which remained on the screen for 1200 ms, regardless of response time. Following go stimulus offset, feedback was presented for 500 ms. The next trial started after a further 200 ms, during which a blank screen was shown.

The primary task was to respond to the identity of the Hiragana character. Both characters required a bimanual response. This kept the primary task the same throughout the experiment in order to keep AS and SS stop trials as similar as possible. One character was mapped onto the two upper keys of a response pad; the other character was mapped onto the two lower keys. The character-to-key mapping was counterbalanced across participants. Participants pressed the two upper keys with their left and right middle fingers and the lower keys with their left and right index fingers. Participants were instructed to respond as accurate, fast, and simultaneously as possible.

On 40% of the trials, one of the four secondary stimuli (signals) was presented. The other 60% percent were no-signal (NS) trials. Three of the four signals indicated that an adjustment of the response to the go-stimulus was required. There were two versions of the task, counterbalancing the stimulus-to-signal mapping across participants. One stimulus (orange diamond/purple club) acted as the stop-left (SL) signal; it instructed participants to stop their left-hand response, but not their right-hand response. A second stimulus acted as the stop-right (SR) signal (cyan heart/yellow spade); it instructed participants to stop their right-hand response, but not their left-hand response. A third stimulus acted as the stop-both (SB) signal (yellow spade/cyan

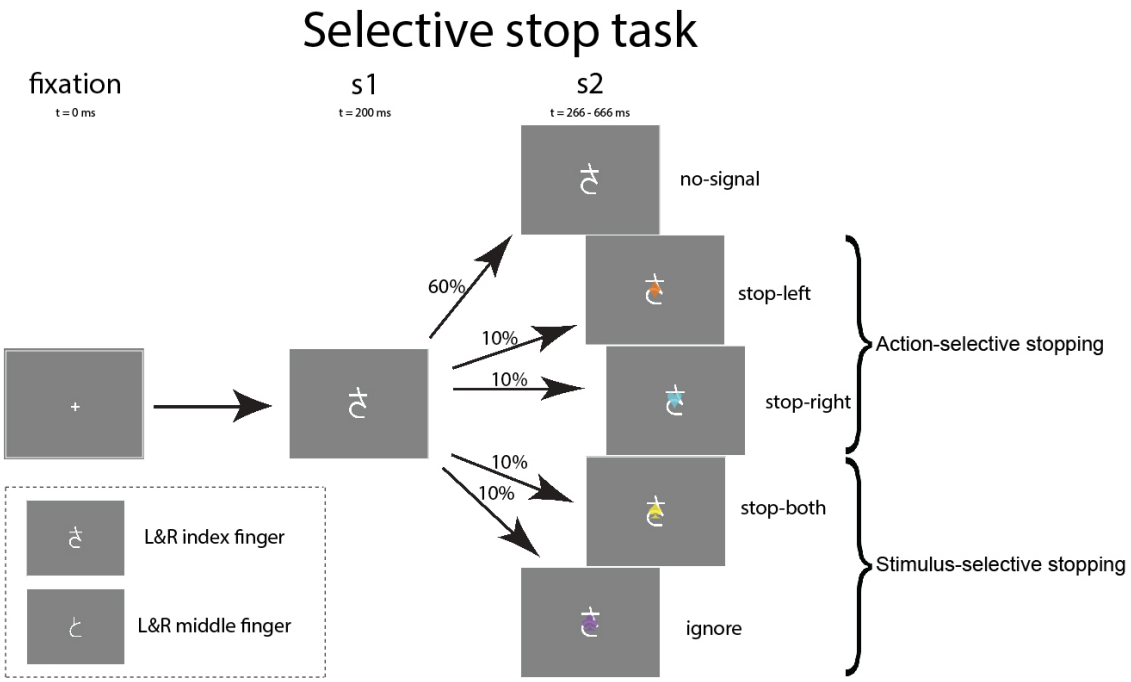


Fig. 1. Schematic of the selective stopping task. Each trial began with 200 ms fixation, followed by one of two go stimuli (s1). The main task was to quickly respond by pressing a button with either both index fingers or both middle fingers, depending on the identity of the go stimulus. On 40% of the trials the go stimulus was followed by one of four secondary stimuli (s2; a colored playing card suit symbol) after a delay of 66, 166, 266, 366 or 466 ms. Two served as action-selective stop-signals, indicating to stop the left- hand response but continue the right-hand response, or vice versa. The other two served as stimulus-selective stop-signals; one indicated to stop both responses, the other signal was to be ignored and the response should be carried out as normal.

heart); it instructed participants to stop both their left- and right-hand response. The fourth stimulus acted as the ignore (IG) signal (purple club/orange diamond); it had to be ignored and the response had to be continued as planned. The primary and secondary stimulus were separated by a stimulus onset asynchrony (i.e., the stop-signal delay, t_d). There were five t_d , each occurring with equal probability: 66 ms, 166 ms, 266 ms, 366 ms, and 466 ms. These values were based on pilot data.

Procedure

Practice session. The experiment began with both written and verbal instructions. Next, the task was practiced in three stages. In stage 1, participants performed one block of 50 no-signal trials to acquaint them with the go task. In stage 2, participants performed one AS stopping block (only SL or SR signals could appear) and one SS stopping block (only SB or IG signals could appear), to acquaint them with both the AS and SS stop task, while maintaining go task performance. The blocks consisted of 100 trials each and the order of the blocks was counterbalanced across participants. In stage 3, participants performed one block of 100 trials in which AS and SS stop trials were intermixed, to acquaint them with the task as it would be performed in the scanner. After each practice block participants were provided with feedback on their performance. The practice block had to be repeated if one of the criteria listed in Table 1 was not met. Each block could be repeated up to five times. If any of these criteria was not met after five subsequent repetitions, the experiment was terminated and the participant was excluded. One participant was excluded for this reason. Upon successful completion of the practice session, the fMRI session was scheduled for another day.

Functional MRI session. Participants performed two runs of 6 experimental blocks while being scanned with fMRI, using an event-related design. Each block consisted of 100 trials and ended with a 12.6-second feedback screen on the task performance (identical to the practice session). If one of the performance criteria listed in Table 1 was not met on five subsequent blocks, the participant

Table 1.

Performance criteria. NS = no-signal; SL = stop-left; SR = stop-right; SB = stop-both; IG = ignore

Outcome measure	Performance criteria
Mean NS RT	< 650 ms
NS choice performance (i.e., correct response to the go stimulus	>= 85% correct
Mean difference between left- and right hand RT on NS trials	< 50 ms
SL trials	$20\% < P(respond \mid\mid SLsignal) < 80\%$
SR trials	$20\% < P(respond \mid\mid SRsignal) < 80\%$
SB trials	$20\% < P(respond \mid\mid SBsignal) < 80\%$
IG trials	$P(respond \mid\mid IGsignal) >= 80\%$

would be excluded. One participant was excluded for this reason.

The trials were presented in a pre-determined pseudo-random sequence. In order to determine the optimal trial order, we generated 100,000 pseudo-random trial sequences and selected the two sequences (one for each fMRI run) with the highest detection efficiency for the contrast between AS and SS stopping and the lowest variance inflation factor, as determined with MATLAB-software for optimisation of fMRI designs (Wager & Nichols, 2003).

Data acquisition

Task performance data. The experiment was run in PsychoPy (version 1.83.04) in Windows 7 Enterprise OS, on a DELL PRECISION T3500 computer. Visual stimuli were projected on a screen positioned 75 cm from the subject and were viewed through a mirror mounted on the head coil. Responses were collected using two MR-compatible response pads (Current Designs, Inc; Philadelphia, PA, USA), one for each hand.

Neuroimaging data. The experiment was performed on a 3.0 T Siemens Magnetom Skyra MRI scanner (Siemens Medical Systems, Erlangen, Germany) at the Donders Institute. Images were acquired using a 32-channel head coil. During task performance, a total of 1214 images with blood-oxygen level-dependent (BOLD) contrast were acquired in 2 runs, using a whole-brain T2*-weighted gradient echo multi-echo echo planar imaging (EPI) sequence (34 slices per volume; transversal acquisition; repetition time = 2100 ms; echo times 8.5 ms, 19.3 ms, 30 ms, and 41 ms; field of view = 224 x 244 mm; flip angle = 90°; 64 x 64 matrix; 3.5 mm in-plane resolution; 3 mm slice thickness; 0.5 mm slice gap). The first 6 scans of each run were discarded to allow T1 saturation to reach equilibrium. Before the first task run, 30 images were acquired during resting-state, with the same pulse sequence, for determining optimal weighting of echo times for each voxel. Between the two task runs, a whole-brain structural image was acquired for within-subject registration purposes, using a T1-weighted magnetisation prepared, rapid-acquisition gradient echo sequence (192 sagittal slices; repetition time = 2300 ms; echo time = 3.03 ms; field of view = 256 x 256 mm; flip angle = 8°; 256 x 256 matrix; 1.0 mm in-plane resolution; 1.0 mm slice thickness).

Data analysis

Bayesian hypothesis testing. Behavioural data were analysed with a Bayesian hypothesis testing approach, based on the Bayes factor (Kass & Raftery, 1995). Bayesian hypothesis testing is comparative in nature and the Bayes factor quantifies the support that the data provide for one hypothesis (e.g., the null hypothesis, H_0) over another (e.g., the alternative hypothesis, H_1). This approach has several advantages over classical hypothesis testing based on the *p*-value. Most importantly, it allows for obtaining evidence both in favor and against H_0, rather than against H_0 only. This is particularly relevant in this study, because we investigate whether the independent race model does (H_1) or does not (H_0) extend to selective stopping and

whether AS and SS stopping is (H_0) or is not (H_1) a homogeneous construct. In addition, the support for one hypothesis over another is provided as a continuous measure (the Bayes factor) instead of a forced, all-or-none, decision.

The Bayes factor describes the relative probability of the data under competing hypotheses. In Bayesian hypothesis testing, the relative odds of the hypotheses themselves are evaluated:

$$\underbrace{\frac{P(H_0|Data)}{P(H_1|Data)}}_{posterior\ odds} = \underbrace{\frac{P(Data|H_0)}{P(Data|H_1)}}_{Bayes\ Factor\ (B_{01})} \times \underbrace{\frac{P(H_0)}{P(H_1)}}_{prior\ odds}$$

Here, the prior and posterior odds describe the beliefs about the hypotheses before and after observing the data, respectively. The primary measure of interest, however, is the Bayes factor that quantifies the change in odds from prior to posterior. In other words, the Bayes factor describes how the evidence from the data should change our beliefs. The Bayes factor prefers the hypothesis under which the observed data are most likely. To illustrate, if $B_{01} = 5$, the data are five times as likely to have occurred under H_0 than under H_1; if $B_{01} = 1$, the data provide equal support for H_0 and H_1. While the Bayes factor is easy to understand, it can be useful to summarise its value in words. For this purpose, we used a set of labels listed in Table 2, proposed by Wetzels and Wagenmakers (2012).

Table 2.

Bayes factor (B_{01}) categories, as proposed by Wetzels and Wagenmakers (2012).

B_{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 – 1/10	Very strong evidence for H_1
<1/100	Extreme evidence for H_1

Bayesian hypothesis testing requires specification of priors, which describe the distribu- tion of effect size. Prior distributions should be specified for both the null hypothesis and the alternative hypothesis. Here, we attempted to specify prior distributions that convey little information while maintaining desirable properties by placing priors on standardised effect sizes (δ) for H_0 and H_1, as well as on the variance (σ^2; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The standardised effect size is assumed to be 0 under H_0 and distributed according to a Cauchy distribution with scale parameter $\tau = 0.5$ under H_1. The prior for σ^2 is less important, because it is the same for both hypotheses and cancels out in the Bayes factor. Following Rouder, Morey, Speckman, and Province (2012), we assume that σ^2 follows an inverse chi-square distribution with one degree of freedom.

In this study, we used Bayesian t-tests (Rouder et al., 2009) and Bayesian repeated- measures ANOVAs (Rouder et al., 2012). These tests involve running separate test models, one including the independent variable as a factor and one null model in which the only factor is the between-subject variance. Model comparison determines which model is the most likely, given the data. This determines whether or not the independent variable has an effect on the dependent variable.

For the Bayesian analyses, we used the Bayes factor package in the software R (https://cran.r-project.org/web/packages/BayesFactor/index.html) and JASP (https://jasp- stats.org/).

Behavioural analyses. The primary outcome measures in the behavioural analyses were P_r, no-signal RT and stop- respond RT for the different stop-signals. Only stop-respond trials with a bimanual response were included, because stop-respond trials with a unimanual response do not necessarily reflect failures of stopping: they may occur when participants successfully stop both prepared responses, but then erroneously discriminate the signal (e.g., confusing stop-left for stop- right) and execute the wrong response. In addition, trials with RTs faster than 150 ms were considered anticipated and were excluded from the analyses.

Tests of independent race model predictions. To test whether the independent race model extends to selective stopping, we tested whether AS and SS stopping performance was in line with the three predictions of the model. We tested these predictions for AS and SS stopping separately, both at the individual level and group level.

To test the first prediction (P_r increases with t_d) at the individual level, we plotted the inhibition functions of each subject. At the group level, we analysed P_r with a Bayesian repeated-measures ANOVA, with t_d as a factor. For the second prediction (stop-respond RT is faster than no-signal RT), we analysed the RTs at the individual level with Bayesian independent t-tests, with trial outcome (no-signal or stop-respond trial) as a factor. At the group level, we analysed the RTs with two Bayesian paired t-tests, one for AS and one for SS stopping. We tested the third prediction (stop-respond RT increases with t_d) by analysing stop-respond RT with a Bayesian repeated-measures ANOVA, with t_d as a factor. For this analysis (both at the individual level and the group level), the factor t_d had three levels rather than five: short delay (66, 166, 266 ms), intermediate delay (366 ms) and long delay (466 ms). Pooling of stop-respond RTs over the first three delays was necessary, because of the uneven distribution of stop-respond trials; there are more stop-respond trials at longer t_d.

There were two deviations from the pre-registration in these analyses. Firstly, we did not use Bayesian logistic regression for the analyses of the effect of t_d on P_r, because Bayesian logistic regression had not yet been implemented in the Bayes factor package for R. Secondly, we added a restriction to the models in the Bayesian repeated-measures ANOVAs. We described in the pre-registration that the data support an effect of the independent variable (t_d here; selective stopping type in next section's analyses) if the winning ANOVA model contains the independent variable as a factor. However, this model only supports a main effect, not necessarily in the right direction. In order to find evidence for or against an increase in P_r or stop-respond RT with increasing td, we created an order-restricted model. The order-restricted model took 10,000 samples from the posterior distribution of the full model (the not-order-restricted model) and computed the frequency of the correct ordering (higher P_r and longer RTs at longer t_d). We also report the results of the full model.

Behavioural tests of differences between selective stopping types. There were three behavioural analyses (group level) to test whether selective stopping is a behaviourally homogeneous or heterogeneous construct. Firstly, we analysed the P_r with a Bayesian repeated-measures ANOVA, with selective stopping type (AS or SS) as a factor. Secondly, we analysed the stop-respond RT with a Bayesian repeated-measures ANOVA, with selective stopping type as a factor. Thirdly, we analysed the two SSRTs using a Bayesian paired t-test, with selective stopping type as a factor. For the third analysis, subjects whose performance was not in line with all the predictions of the independent race model for both AS and SS stopping were excluded.

Functional MRI analyses

Image data were preprocessed using Statistical Parametric Mapping 12 (SPM12) software running in MATLAB (Mathworks Inc., Natick, MA, USA). The T1-weighted anatomical scans were skull-stripped using the FSL Brain Extraction Tool (Smith, 2002). The multi- echo images were combined with the PAID method (Poser, Versluis, Hoogduin, & Norris, 2006) using custom-written MATLAB software. The anatomical images were co-registered to the mean functional images using the normalised mutual information criteria method (Studholme, Hill, & Hawkes, 1999). The anatomical and functional images were then normalised to the standard Montreal Neurological Institute 152 (MNI 152) template. The normalised functional images were spatially smoothed using a 6-mm full-width at half-maximum Gaussian kernel.

First-level statistical analysis involved a mass-univariate approach based on general lin- ear models in SPM12. Each subject's whole-brain BOLD data were modeled with a general linear model, including 15 event-related predictors. These were the five different trial types, all subdivided in three ways based on the trial outcome. NS and IG trials were divided into 'fast', 'slow' and 'other' response trials, wherein the 'other' category contained all the incorrect responses. SL, SR and SB trials were subdivided into 'stop-respond-bimanual', 'stop-inhibit' and 'stop-respond-other' response trials. For all regressors, except the three NS regressors, we included a demeaned parametric modulator coding for stimulus onset asynchrony be- tween the go-stimulus and the signal (i.e., t_d). In addition, to account for residual head motion effects, we included the six motion parameters from the realignment procedure in the statistical model. Note that we did not include the first and second order derivatives, as is described in the pre-registration. Taken together, we included a total of 33 regressors per run (i.e., 15 predictors + 12 parametric modulators + 6 motion parameters).

The regressors were created by convolving the delta functions coding for go-stimulus onset with a canonical hemodynamic response function. Time series statistical analysis was performed using restricted maximum likelihood. Low frequency

drifts were controlled using a discrete cosine transform with a cutoff of 128 seconds. Serial correlations in the fMRI signal were estimated using restricted maximum likelihood estimates of variance components using a first-order autoregressive model. The resulting non-sphericity was used to form maximum- likelihood estimates of the activations. Time series statistical analyses were performed using restricted maximum likelihood.

We specified first-level contrasts to isolate activation associated with AS stopping, SS stopping and the difference between them. The contrasts and their purposes are listed in Table 3. The contrasts control for the attentional capture of the stop-signal by subtracting activity on ignore trials from activity on stop trials. A salient signal also occurred on ignore trials, but no inhibition was required, so only attention-related but not inhibition- related activation is subtracted. For this subtraction, we used only ignore trials on which a fast response was made, because on slow ignore trials a STOP process may have been activated temporarily (Bissett & Logan, 2014). The contrasts also control for the difference in speed of the GO process between signal and no-signal trials by first subtracting activation on no-signal trials from both stop- and ignore-related activity. To control for activation associated with the execution of a unimanual response on AS stop trials, we used conjunction analyses (Nichols, Brett, Andersson, Wager, & Poline, 2005) in the second-level statistics to test for activations occurring in both the $S_{AS,left}$ and $S_{AS,right}$ contrasts and in both the $S_{AS,left} - S_{SS}$ and $S_{AS,right} - S_{SS}$ contrasts.

Second-level statistical analyses consisted of two region-of-interest (ROI) analyses and one whole-brain analysis on the first-level contrasts.

First, we analysed brain activation in predefined ROIs using a Bayesian hypothesis testing approach based on the Bayes factor. This allowed us to compare activation in the same way as we compared task performance and find evidence in favor of and against the null hypothesis. ROIs were defined as 6-mm spheres around local maxima in key regions of inhibitory control reported by previous fMRI studies of AS and SS stopping (Table 4): inferior frontal gyrus (IFG), inferior frontal junction (IFJ), striatum (Str), pre-supplementary motor area (pre-SMA), supplementary motor area (SMA), dorsal premotor area (PMd), and primary motor cortex (M1). For each region, we also defined a 6-mm sphere ROI in the other hemisphere, flipping the sign of the x-coordinate. From these ROIs we extracted the mean activation level (i.e. parameter estimate) in the first-level contrasts. We then used Bayesian one-sample t-tests (t-test value = 0) for identification of activation associated with AS stopping ($B_{01} < 1$ in both the $S_{AS,left}$ and $S_{AS,right}$ contrasts) and SS stopping (< 1 in the SSS contrast) and to identify potential differences in activation in the ROIs between AS and SS stopping ($B_{01} < 1$ in the $S_{AS,left} - S_{SS}$ and $S_{AS,right} - S_{SS}$ contrasts).

Second, we performed another ROI analysis, now using a classical hypothesis testing approach and more broadly defined ROIs, based on probabilistic anatomical atlases. The classical hypothesis testing approach allowed us to analyse the fMRI data in a more common framework. The more broadly defined ROIs enabled us to assess activation within key areas of inhibitory control that fell outside the spheres the first ROI analysis focused on. The broad, anatomical

Table 3.

Contrasts created at the first level and their purpose. The subtractions within parentheses control for differences in the speed of the GO process between signal and no-signal trials. The sub- tractions between parentheses control for the attentional capture of the stop-signal.

Contrast	Purpose
$S_{AS,left} = (SL_{stop\text{-}inhibit} - NS_{correct\text{-}slow}) - (IG_{correct\text{-}fast} - NS_{correct\text{-}fast})$	Isolate activation associated with AS stopping
$S_{AS,right} = (SR_{stop\text{-}inhibit} - NS_{correct\text{-}slow}) - (IG_{correct\text{-}fast} - NS_{correct\text{-}fast})$	Isolate activation associated with AS stopping
$S_{SS} = (SB_{stop\text{-}inhibit} - NS_{correct\text{-}slow}) - (IG_{correct\text{-}fast} - NS_{correct\text{-}fast})$	Isolate activation associated with SS stopping
$S_{AS,left} - S_{SS} = SL_{stop\text{-}inhibit} - SB_{stop\text{-}inhibit}$	Isolate differences between AS and SS stopping
$S_{AS,right} - S_{SS} = SR_{stop\text{-}inhibit} - SB_{stop\text{-}inhibit}$	Isolate differences between AS and SS stopping

ROIs are listed in Table 5. The probabilistic map of each of the subregions was thresholded at 25% and the resulting maps were combined into one binary mask. This mask was used for small-volume correction for multiple comparisons.

Within these broad anatomical ROIs we used a one-sample t-test in SPM12 on the SSS contrast for the identification of activation associated with SS stopping. We used conjunction analyses to identify activation associated with AS stopping ($S_{AS,left} \cap S_{AS,right}$) and to identify potential differences in activation in the ROIs between AS and SS stopping ($S_{AS,left} - S_{SS} \cap S_{AS,right} - S_{SS}$). We report activations that were significant at $p < .001$ uncorrected for multiple comparisons and that survived small-volume correction at $p < .05$ after family wise error (FWE) correction for multiple comparisons.

Third, to explore activation associated with AS and SS stopping outside key areas of inhibitory control, we performed whole-brain voxel-wise random effects analyses. Again, we used a conjunction analysis to isolate AS stopping-related activation ($S_{AS,left} \cap S_{AS,right}$), a one-sample t-test on the S_{SS} contrast for SS stopping-related activation and a conjunction analysis for differences between AS and SS stopping related activation ($S_{AS,left} - S_{SS} \cap S_{AS,right}$

Table 4.

ROIs defined on the basis of local maxima reported by previous fMRI studies of selective stopping

ROI	MNI [x,y,z] coordinates	Reference	Contrast in reference
IFG	52, 10, 6	Majid et al. (2013)	MaybeStopRight+Left (Stop > Go)
IFJ	45, 8, 25	Sebastian et al. (2015)	AttentionalCapture > Go
Str	9, 6, 0	Ruiter et al. (2012)	SuccessfulInhibition > Control
Pre-SMA	20, 6, 62	Sharp et al. (2010)	Stop > Continue
SMA	15, -2, 72	Coxon et al. (2016)	StopLeftGoRight > StopAll ∩ StopLeftGoRight > Go ∩ StopRightGoLeft > StopAll ∩ StopRightGoLeft > Go
PMd	28, -2, 65	Coxon et al. (2016)	StopLeftGoRight > StopAll ∩ StopLeftGoRight > Go ∩ StopRightGoLeft > StopAll ∩ StopRightGoLeft > Go
M1	-36, -24, 63	Ruiter et al. (2012)	SuccessfulInhibition > Control

Table 5.

ROIs based on probabilistic neuroanatomical atlases.

ROI	Subregions included	Reference
vlFC	ventral premotor, inferior frontal junction, 44v, 44d, 45, inferior frontal sulcus	F.-X. Neubert, Mars, Thomas, Sallet, & Rushworth (2014)
dFC	supplementary motor area, pre-supplementary motor area, and dorsal premotor area	J. Sallet et al. (2013)
Str	executive, rostral motor, caudal motor	Tziortzi et al. (2014)
M1	area 4a, area 4p	Geyer et al. (1996)

$- S_{SS}$). We report activations that were significant at $p < .001$ uncorrected for multiple comparisons and that survived cluster-level correction at $p < .05$ family wise error-corrected for multiple comparisons.

Results

Behaviour

We tested the three predictions of the independent race model at the group level and at the individual level, separately for AS and SS stopping. Subsequently, we tested to what extent stopping performance differed between AS and SS stopping. One participant failed to meet all pre-set performance criteria (Table 1), hence this dataset was excluded, resulting in a sample of 23 participants in the behavioural analyses.

Figure 2 summarises response times and response probabilities for the main trial types.

P_r increased with t_d. Figure 3 depicts individual and group mean inhibition functions (P_r over the five t_d) for AS and SS stopping. Clearly, P_r increased with t_d in all individuals. Indeed, the analyses at the group level confirmed that, both for AS and SS stopping, the data were much more likely under a model including t_d than a null model that did not include t_d as a factor (AS stopping, $B_{01} = 9.12e-54$; SS stopping, $B_{01} = 1.40e-55$). The data were also more likely under an order-restricted model, in which P_r increases with td, than the null model (AS stopping, $B_{01} = 8.06e-56$; SS stopping, $B_{01} = 1.19e-57$).

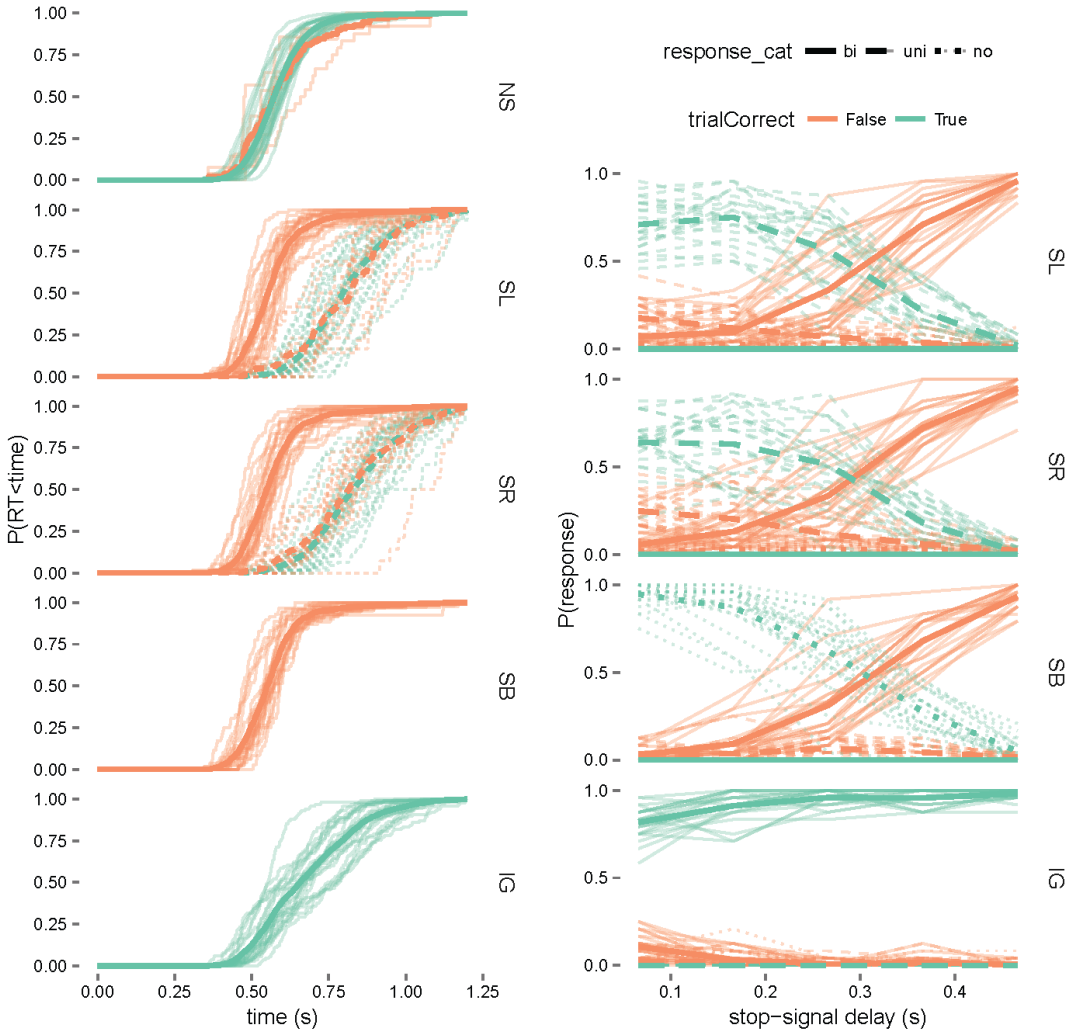


Fig. 2. Overall task performance. The left column of figures shows the cumulative distribution functions of RTs for the five trial types. The right column of figures shows the cumulative probabilities of a response over the signal delays (t_d). Faded lines represent individual subjects; bold lines represent the group mean. Solid lines represent bimanual responses, broken lines represent unimanual responses.

Difference between stop-respond RT and no-signal RT. Figure 4A and C display the relationship between mean stop-respond RT versus the mean no-signal RT. Figure 4B and D show the boxplots of inividual RTs for these trials. At the group level, the stop-respond RT was faster than no-signal RT (M = 583 ms) in both AS stopping (M = 566 ms, B_{01} = 0.001) and SS stopping (M = 562 ms, B_{01} = 4.74e − 05). At the individual level, there was more than anecdotal evidence (B_{01} < 1/3) that stop-respond RT was faster than no-signal RT for nine subjects in AS stopping and for eight subjects in SS stopping.

Stop-respond RT did not increase with t_d. The mean stop-respond RT of each subject in the three t_d bins (short, intermediate, long) is displayed in Figure 5A.

At the group level, the data were much more likely under the full model, which included t_d as a factor, than the null model that did not include t_d , in AS stopping (B_{01} = 0.001). The data for SS stopping were only slightly more likely under the full model than under the null model (B_{01} = 0.735). However, the order-restricted model showed that stop-respond RT did not increase with increasing td, neither in AS stopping (B_{01} = *Inf*, meaning that none of the 10,000 samples of the posterior distribution of the full model had the correct ordering) nor in SS stopping (B_{01} = 21.89).

At the individual level, there was more than anecdotal evidence for the full model being more likely (B_{01} < 1/3) in four subjects in AS stopping and four subjects in SS stopping. The order-restricted model of increasing stop-respond RT with increasing t_d was supported with more than anecdotal evidence in two of those subjects in AS stopping and in all four of the subjects in SS stopping.

The distribution of the log10 transformed Bayes factors (log10 [B_{01}]) for the full model and order-restricted model are shown for in AS and SS stopping in Figure 5B.

Race model predictions taken together. In total, two subjects performed in line with all three predictions of the independent race model in AS stopping and six subjects performed in line with all three predictions in SS stopping (B_{01} < 1; negative log10[B_{01}] in Fig. 4 and 5). Thus, at least one of the predictions of the independent race model was violated in 91% of the individuals in AS stopping and in 74% of the individuals in SS stopping. There were no subjects that performed in line with all three predictions in both AS and SS stopping.

Little difference in behaviour between selective stopping types. Figure 6 provides a clear visual comparison of the inhibition functions and stop-respond RTs of AS and SS stopping.

The two selective stopping types only differed in the P_r. Both a full model and an order- restricted model that included selective stopping type as a factor were more likely than a null model without it as a factor. The effect was only small, however. Model comparison showed that the models including selective stopping type as a factor were 1.13 times more likely.

There was no effect of selective stopping type on the stop-respond RT. The full model without including selective stopping type as a factor was 3.56 times more likely than with it as a factor. The order-restricted models with and without selective stopping type as a factor both returned infinite Bayes factors, because there was no ordered effect of t_d on stop-respond RT at the group level.

Lastly, we intended to analyse the effect of selective stopping type on the SSRT. However, not one subject performed in line with all three predictions of the independent race model in both

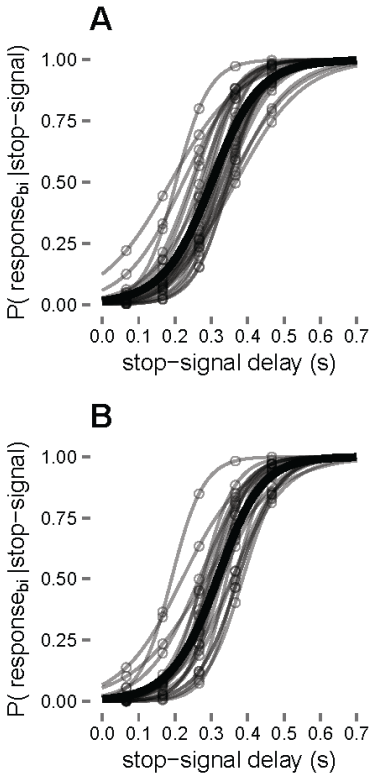


Fig. 3. Inhibition functions for the individual subjects and the group mean for AS (**A**) and SS stopping (**B**). Faded lines and open dots represent individual subjects; bold line represents the group mean.

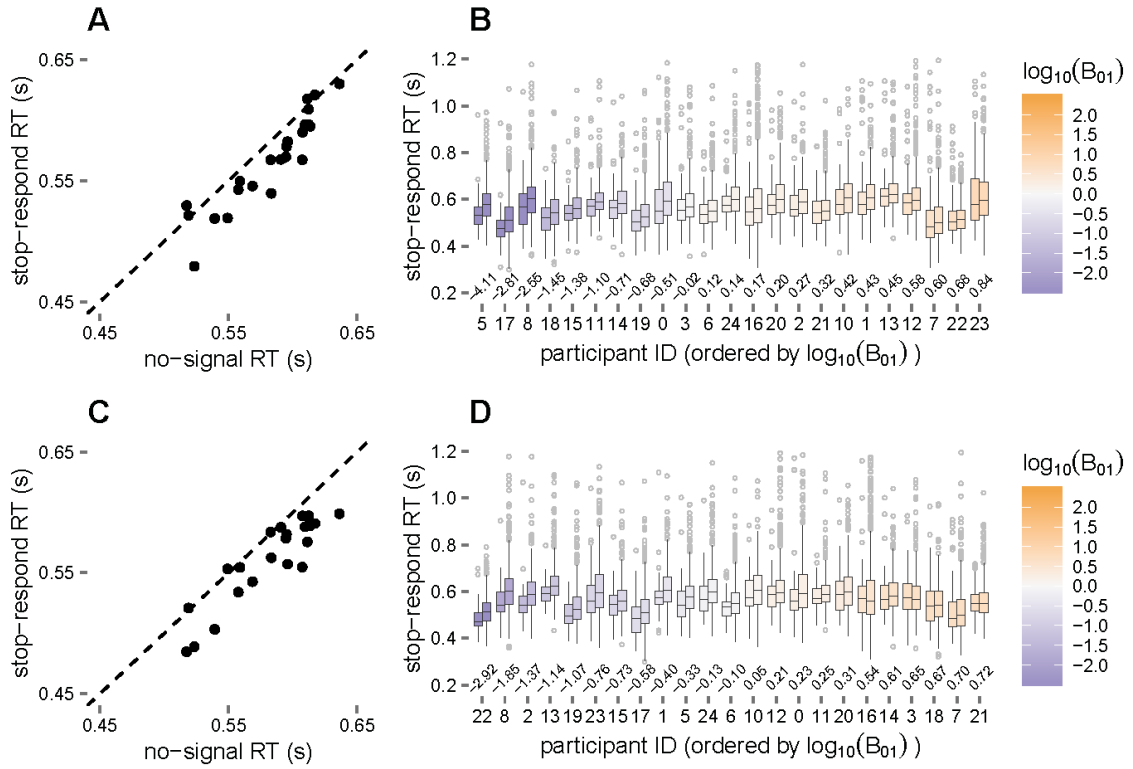


Fig. 4. Relation between stop-respond RT and no-signal RT. Panels **A** and **C** show the individual mean stop-respond against mean no-signal RT for AS and SS stopping, respectively. Panels **B** and **D** show boxplots of the distributions of stop-respond (left) and no-signal (right) RTs for AS and SS stopping, respectively. The values in panels **B** and **D** represent log10 Bayes factors, i.e. log10(B_{01}). Values bigger than 0.5 and smaller than -0.5 indicate more than anecdotal evidence for H_0 and H_1, respectively.

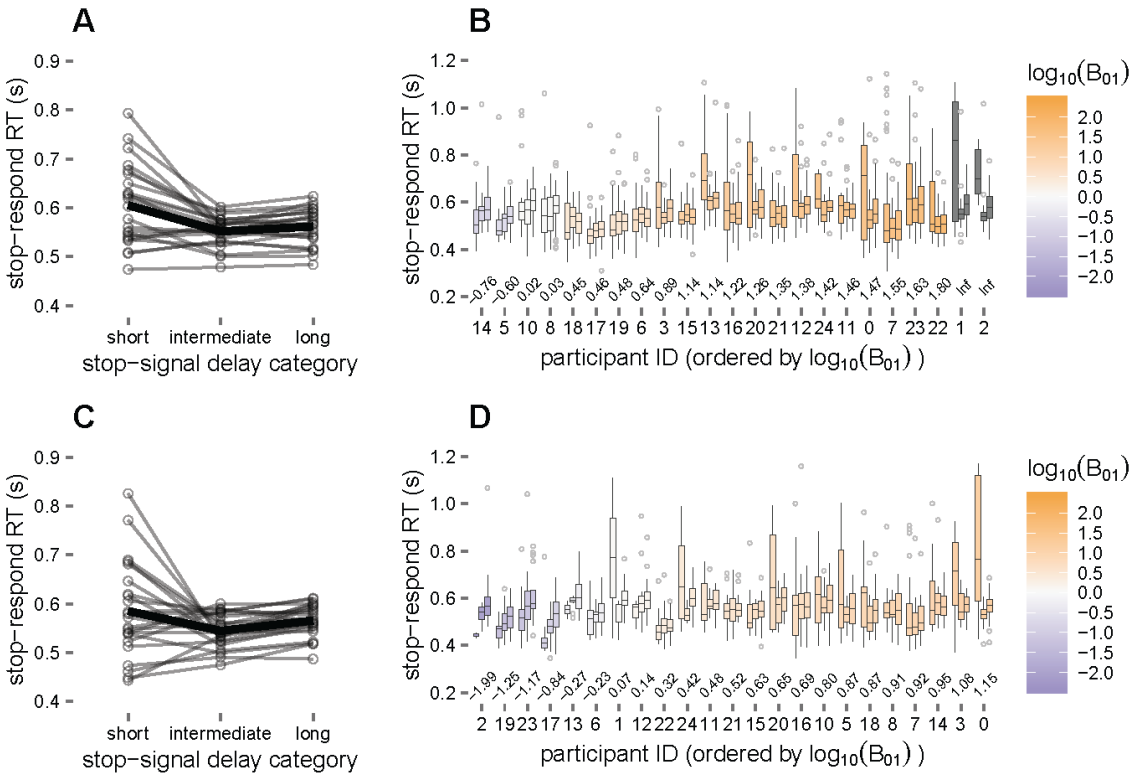


Fig. 5. Effect of stop-signal delay on stop-respond RT. Panels **A** and **C** show the individual mean stop-respond RT for the three td categories in AS and SS stopping, respectively. Panels **B** and **D** show boxplots of the distributions of stop-respond RTs for short (left), intermediate (middle) and long (right) delays in AS and SS stopping, respectively. Conventions as in Figure 4.

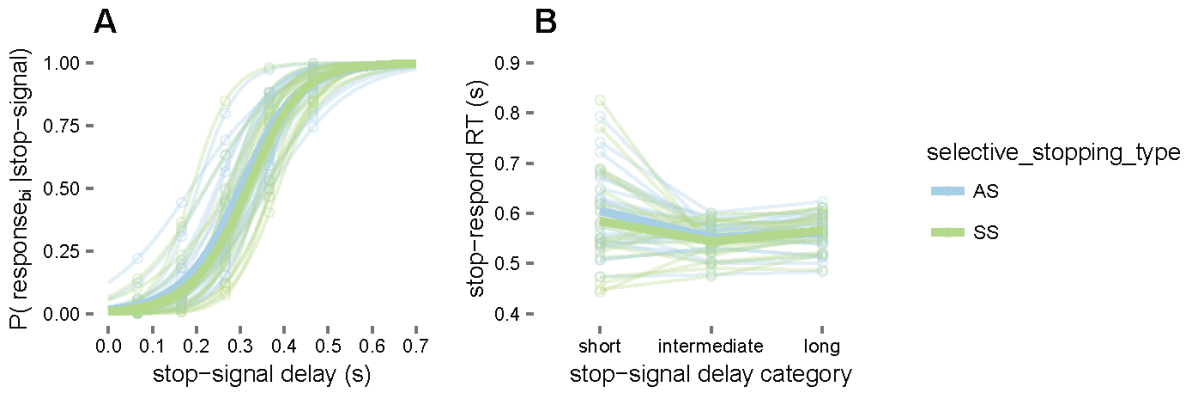


Fig. 6. Comparison of the inhibition functions (**A**) and stop-respond RTs in the three t_d categories (**B**) for AS and SS stopping.

AS and SS stopping. Thus, we could not estimate SSRT reliably, hence we could not analyse the effect of selective stopping type on the SSRT.

Functional Magnetic Resonance Imaging

We excluded two additional datasets from the fMRI analyses because of excessive head movement in the scanner, resulting in a final sample of 21 participants.

To identify brain activations related to stopping in simple stop-signal tasks, previous studies typically used the contrast stop > go. Figure 7 shows the whole-brain activation maps of the contrasts $SB_{inhibit} - NS_{slow}$, $SL_{inhibit} - NS_{slow}$, $SR_{inhibit} - NS_{slow}$ and $IGf_{ast} - NS_{fast}$.

These contrasts are comparable with the simple stop > go contrast, with the exception that they also control for the speed of the GO process. We found that AS and SS stopping, like simple stopping, activate a network of regions in the frontal and parietal lobe as well as the basal ganglia, suggesting that the task manipulation worked.

For the analyses below, we used contrasts that control for both the attentional capture of the salient signal and the speed of the GO process (Table 2). We applied these contrasts to isolate AS and SS stopping-related activations in predefined functional ROIs, broader anatomical ROIs and at the whole-brain level.

Functional ROI analyses. First, we tested which of the predefined ROIs were (de)activated in association with selective stopping, using Bayesian one-sample t-tests. Figure 8 shows boxplots of the contrast estimates in the ROIs, colour-coded for the evidence the data provide for the null versus the alternative hypothesis.

AS stop trials were associated with deactivation of the contra-lateral M1 (i.e. stop-right trials deactivated left M1 and vice versa). Two ROIs showed activation associated with AS stopping: left PMd and left SMA. For these ROIs the Bayes factor supported activation ($B_{01} < 1$) in both AS stopping contrasts. There was evidence for absence of AS stopping-related activation ($B_{01} > 1$) in the left and right IFG, left and right IFJ and the left and right striatum. The other ROIs were activated in one of the AS stopping contrasts but not the other, providing mixed evidence.

SS stop trials were associated with deactivation of both motor cortices. Four other ROIs showed deactivation associated with SS stopping: left IFJ, right PMd, left and right striatum. In the other ROIs there was evidence for no SS stopping-related activation.

There was a difference between activations associated with AS stopping and SS stopping in the left IFJ, left and right PMd, left pre-SMA, left SMA and left and right striatum. For these ROIs there was evidence for a difference ($B_{01} < 1$) in both contrasts subtracting SS stopping-related activations from AS stopping-related activations. There was evidence for no effect of selective stopping type in the right IFG and the right pre-SMA. The other three ROIs were activated in one of the contrasts but not the other, providing mixed evidence for an effect of selective stopping type.

The contrast estimates in the left PMd and left SMA were positive in the contrasts subtracting SS stopping-related activations from AS stopping-related activations and they were activated in both AS stopping contrasts. Thus, the left PMd and left SMA were activated during AS stopping and their activation was greater in AS stopping than in SS stopping.

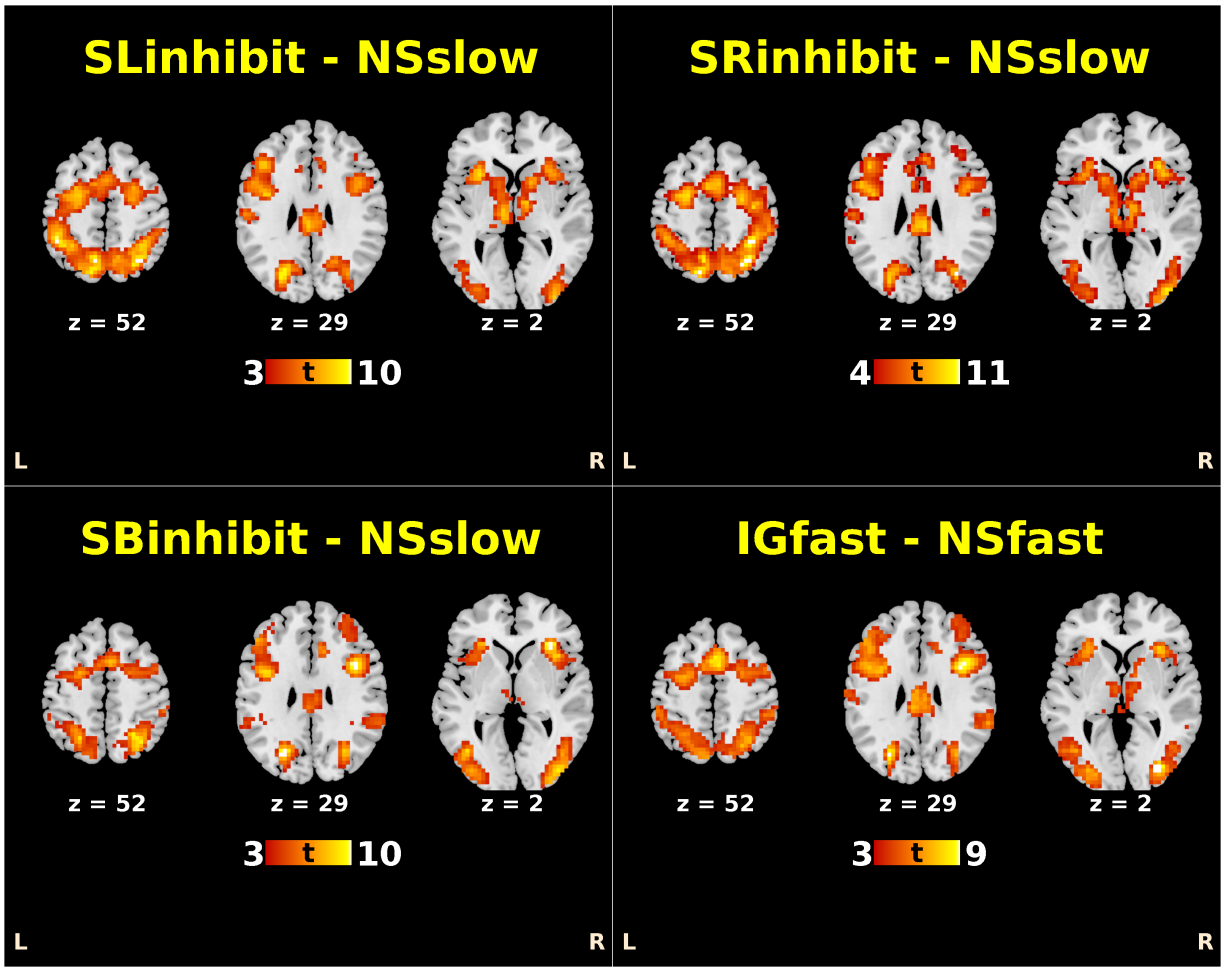


Fig. 7. Whole-brain activation on signal trials. One-sample t-tests of the contrasts $SB_{inhibit} - NS_{slow}$, $SL_{inhibit} - NS_{slow}$, $SR_{inhibit} - NS_{slow}$ and $IG_{f\,ast} - NS_{fast}$. The activations are significant at the cluster-level (cluster-defining threshold $p < .001$ uncorrected; cluster probability $p < .05$, family wise error-corrected for multiple comparisons) and are overlaid on a template brain in MNI space (neurological orientation).

Broad anatomical ROI analyses. Next, we investigated selective stopping-related brain activation in more broadly defined ROIs (Table 5), based on probabilistic anatomical atlases, using classical hypothesis testing. Local maxima of activations in the conjunctions of the AS stopping contrasts and the contrasts subtracting SS stopping-related activations from AS stopping-related activations are shown in Table 6 and 7. Figure 9 shows the masked and small-volume corrected activation maps.

The left and right PMd showed significant activation associated with AS stopping ($p < .001$ uncorrected and $p < .05$ FWE small-volume correction for multiple comparisons) and there was significantly greater activation in AS stopping than in SS stopping in the left and right PMd, the left PMv, and the left SMA ($S_{AS,left} - S_{SS} \cap S_{AS,right} - S_{SS}$). There were no significant activations in the SS stopping contrast in the anatomical ROIs.

Whole-brain analyses. We also applied the same contrast and conjunctions at the whole-brain level to identify activations associated with selective stopping outside the key inhibitory control areas.

Besides the earlier reported activations in the left and right PMd, the whole-brain analysis revealed that AS stopping was associated with activations in the left superior and inferior parietal lobule (Table 8, Fig. 10A). Furthermore, activations associated with AS stopping were greater than SS stopping-related activations in superior prefrontal areas, including the left and right PMd, the left precentral gyrus, superior and inferior parietal areas and the left thalamus and right cerebellum (Table 9, Fig. 10B). There were no significant activations in the SS stopping contrast at the whole-brain level.

Exploratory analyses

We performed two exploratory analyses. First, we compared the observed stop-respond RTs with the

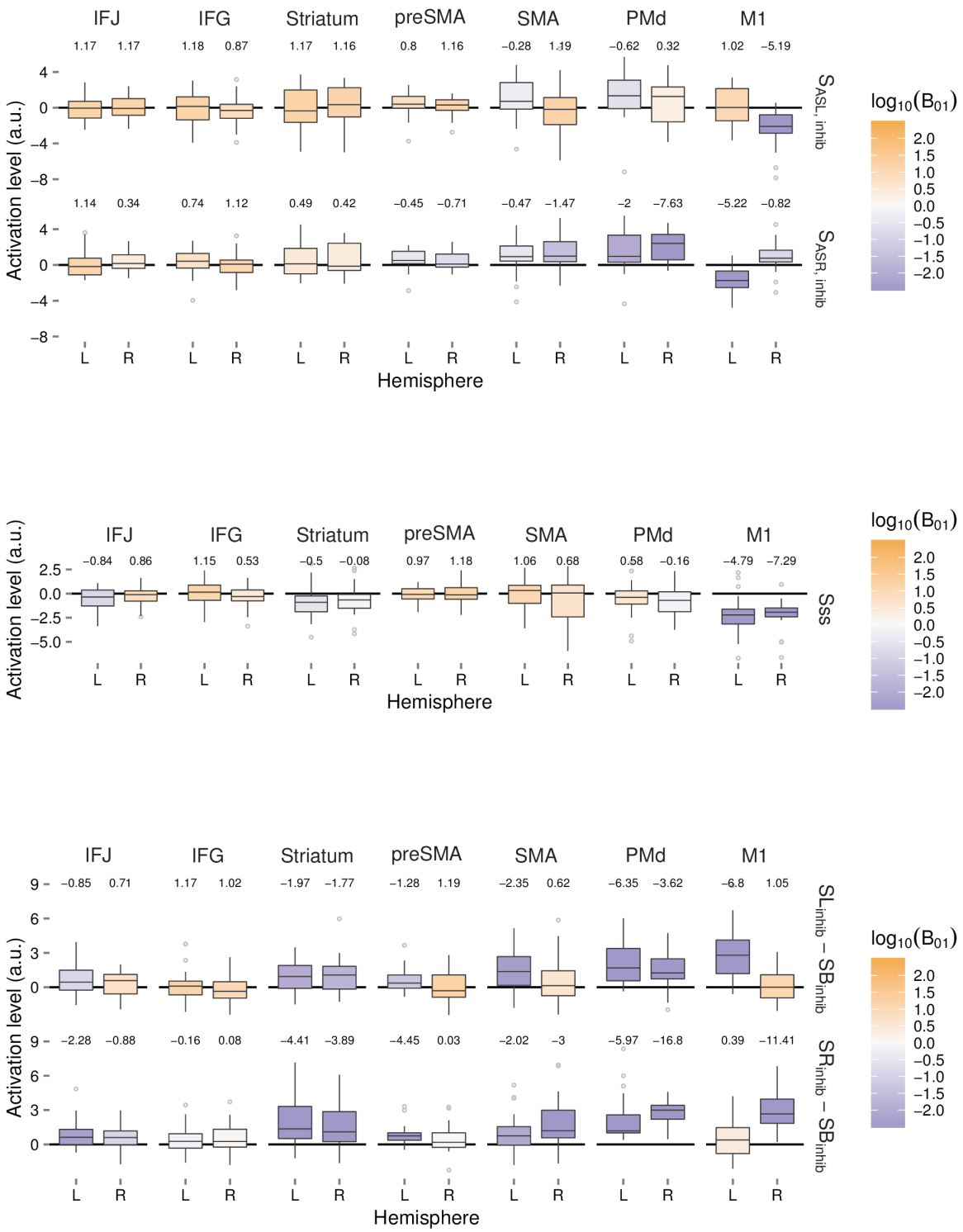


Fig. 8. Boxplots of the activation levels (a.u.) in the predefined functional ROIs. The upper panel shows the activation in the AS stopping contrasts: $S_{AS,left}$ and $S_{AS,right}$. The middle panel shows the activation in the SS stopping contrast: SSS. The lower panel shows the activation in the contrasts subtracting SS stopping-related activations from AS stopping-related activations: $S_{AS,left} - S_{SS}$ and $S_{AS,right} - S_{SS}$. The boxplots are color coded with the corresponding Bayes fit'tors. Conventions of the color coding and values as in Figure 4.

Table 6.

Local maxima of brain activations in the anatomical ROIs during AS stopping (*SAS,left* ∩ *SAS,right*) in MNI x-, y-, and z- coordinates with associated Z-score and small-volume corrected p- value at $p < .05$ FWE-corrected.

Region	Hemisphere	x	y	z	Voxel Z value	*pz*
PMd	L	-24	-4	56	4.76	.004
PMd	R	26	-7	52	4.62	.007

Table 7.

Local maxima of differences in brain activations in the anatomical ROIs between AS and SS stopping (*SAS,left – SSS* ∩ *SAS,right – SSS*) in MNI x-, y-, and z-coordinates with associated Z-score and small-volume corrected p-value at $p < .05$ FWE- corrected.

Region	Hemi-sphere	x	y	z	Voxel Z value	*pz*
PMd	L	-24	-7	56	5.90	.000
PMd	R	26	-7	52	5.57	.000
PMv	L	-55	7	28	5.41	.000
SMA	L	-2	0	52	4.54	.01

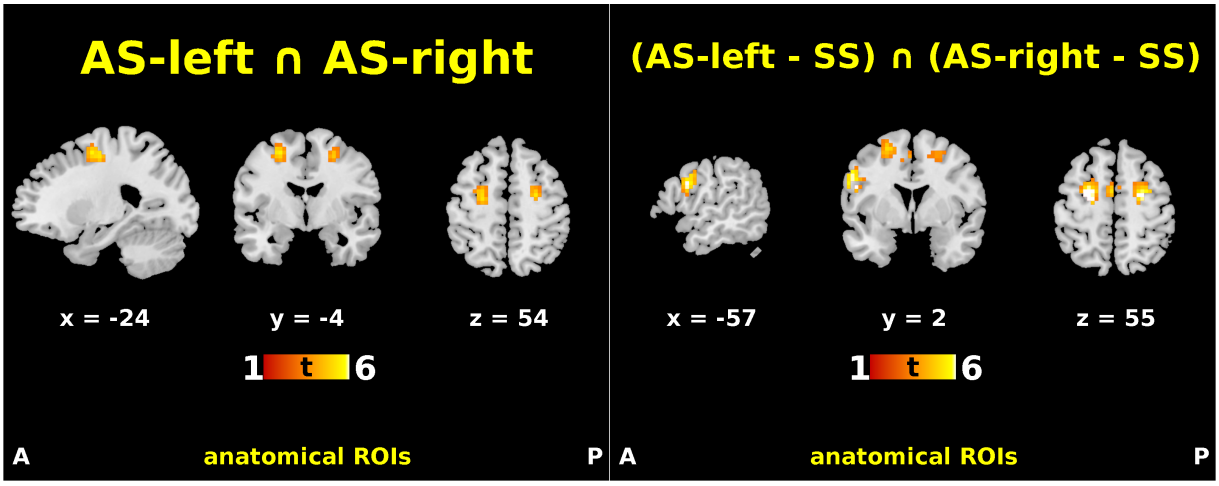


Fig. 9. Brain regions significantly activated in the broad anatomical ROIs in the conjunctions $S_{AS,left} \cap S_{AS,right}$ (left) and $S_{AS,left} - S_{SS} \cap S_{AS,right} - S_{SS}$ (right). Activations are significant at $p < .001$ uncorrected and $p < .05$ FWE small-volume corrected for multiple comparisons, and are overlaid on a template brain in MNI space (neurological orientation).

Table 8.

Local maxima of brain activations (whole-brain analysis) during AS stopping (*SAS, left* ∩ *SAS, right*) in MNI x-, y-, and z-coordinates with associated Z-score ($p < .001$ uncorrected) and cluster size in number of voxels (k; $p < .05$ FWE-corrected).

Region	Hemisphere	x	y	z	Voxel Z value	pz	Cluster size (k)	pk
Superior Parietal Lobule (Precuneus)	L	-10	-63	52	5.45	.000	66	.003
Superior Frontal Gyrus (PMd)	L	-24	-4	56	4.76	.000	55	.007
Superior Frontal Gyrus (PMd)	R	26	-7	52	4.62	.000	35	.045
Inferior Parietal Lobule	L	-41	-38	49	4.46	.000	80	.001

stop-respond RTs that the independent race model predicts (Fig. 11). We performed this analysis, because performance was in line with the second prediction of the independent race model at the group level, yet more than half of the individuals violated it. The model predicts that the mean stop-respond RT of an individual corresponds to that individual's mean of the fast bin of no-signal RTs. In our study, the mean stop-respond RT of most subjects was only slightly faster than their overall mean no-signal RT (fast and slow combined). The independent race model predicts a much larger difference (60 ms on average), as can be seen in Figure 11.

Second, we investigated why the SS stopping contrast yielded no significant activations, by examining the two contrasts from which S_{SS} is built up: $(SB_{stop-inhibit} - NS_{correct-slow})$ and $(IG_{correct-fast} - NS_{correct-fast})$. Subtracting the second subcontrast from the first was supposed to control for activations associated with the attentional capture of the salient stop-signal. However, examination of the two contrasts revealed that both activated the same brain regions (Fig. 12A), to the same degree (Fig. 12B). Thus, subtracting the two subcontrasts for the main S_{SS} contrast completely canceled out the activations.

Discussion

The past decade has seen a surge of interest in selective stopping. Researchers studying se- lective stopping have relied on the independent race model of simple stopping for estimation of the primary outcome measure of stopping, the stop-signal reaction time (SSRT). Further- more, they have investigated selective stopping with a heterogeneous set of tasks, including action-selective stop tasks probing control of specific actions and stimulus-selective stop tasks examining control triggered by specific stimuli. However, it remains unclear whether the independent race model extends to selective stopping and whether selective stopping is a homogeneous or heterogeneous construct. Here, we addressed these important gaps by testing whether selective stopping performance is in agreement with predictions of the independent race model, and by comparing action- and stimulus-selective stopping in terms of performance and brain activation.

We found violations of the predictions of the independent race model in almost all subjects in both AS and SS stopping, suggesting that the model does not apply to selective stopping. Our behavioural and neuroimaging results further suggest that AS and SS stopping were not different in terms of stopping.

Selective stopping involves a race, but not an independent race

We found striking differences between the results of the tests of the independent race model's predictions at the group level and at the individual level. For the group as a whole, selective stopping performance was in agreement with two predictions of the independent race model: the probability of responding increased with stop-signal delay and response times were on average faster on stop-respond than no-signal trials. These findings are in line with previous selective stopping studies that tested predictions of the independence race model (Aron & Verbruggen, 2008; Sebastian et al., 2015; e.g., Smittenaar et al., 2013). However, we found that a third prediction of the model was violated:

Table 9.

Local maxima of differences in brain activations (whole- brain analysis) between AS and SS stopping (*SAS,left – SSS* ∩ *SAS,right – SSS*) in MNI x-, y-, and z-coordinates with associated Z-score ($p < .001$ uncorrected) and cluster size in number of voxels (k; $p < .05$ FWE-corrected).

Region	Hemi-sphere	x	y	z	Voxel Z value	*pz*	Cluster size (k)	p_K
Inferior Parietal Lob-ule	L	-38	-38	42	6.42	.000	613	.000
Superior Frontal Gyrus (PMd)	L	-24	-10	60	6.02	.000	193	.000
Precentral Gyrus	L	-55	4	28	5.75	.000	83	.001
Superior Frontal Gyrus (PMd)	R	26	-7	52	5.57	.000	110	.000
Thalamus	L	-13	-21	10	5.00	.000	129	.000
Superior Parietal Lobule	R	15	-66	63	4.82	.000	73	.002
Cerebellum	R	18	-66	-21	4.60	.000	98	.000
Posterior-Medial Frontal	L	-2	0	52	4.54	.000	42	.023
Inferior Parietal Sul-cus	R	36	-35	38	4.47	.000	54	.008

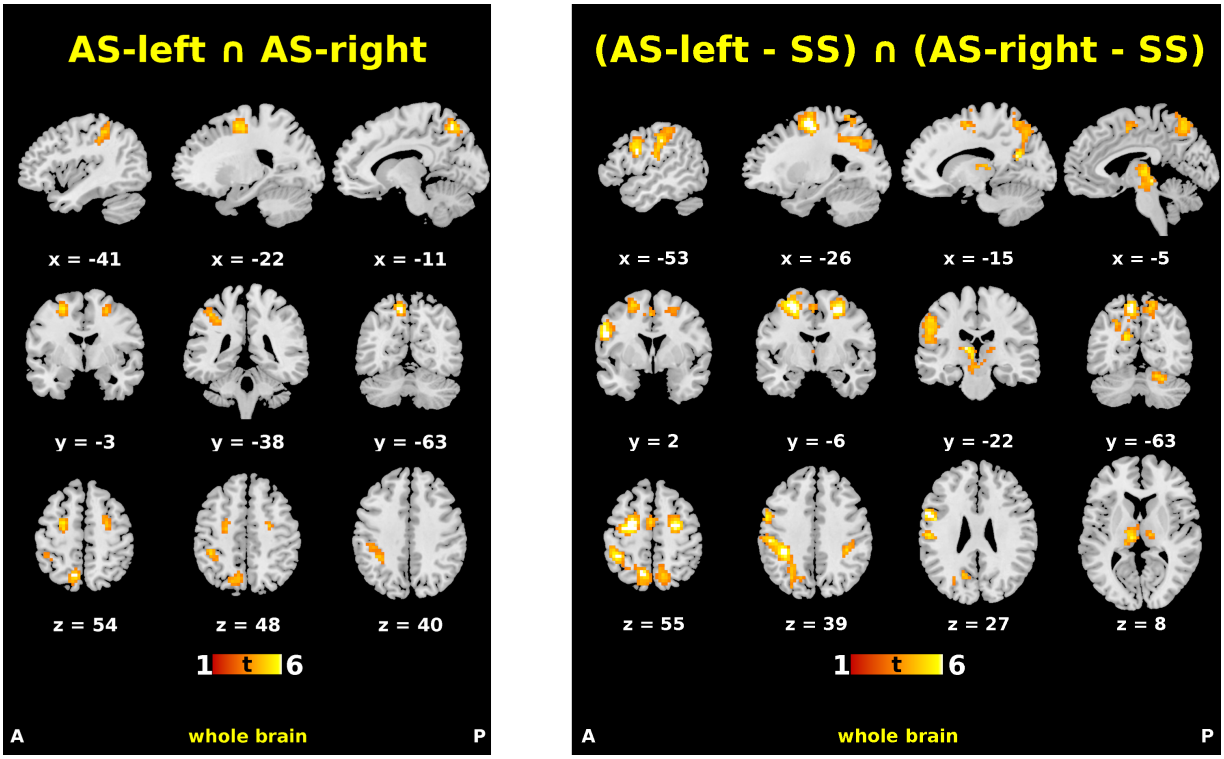


Fig. 10. Brain regions significantly activated at the whole-brain level in the conjunctions $S_{AS,left} \cap S_{AS,right}$ (left) and $S_{AS,left} - S_{SS} \cap S_{AS,right} - S_{SS}$ (right). The activations are significant at the cluster-level (cluster-defining threshold $p < .001$ uncorrected; cluster probability $p < .05$, family wise error-corrected for multiple comparisons) and are overlaid on a template brain in MNI space (neurological orientation).

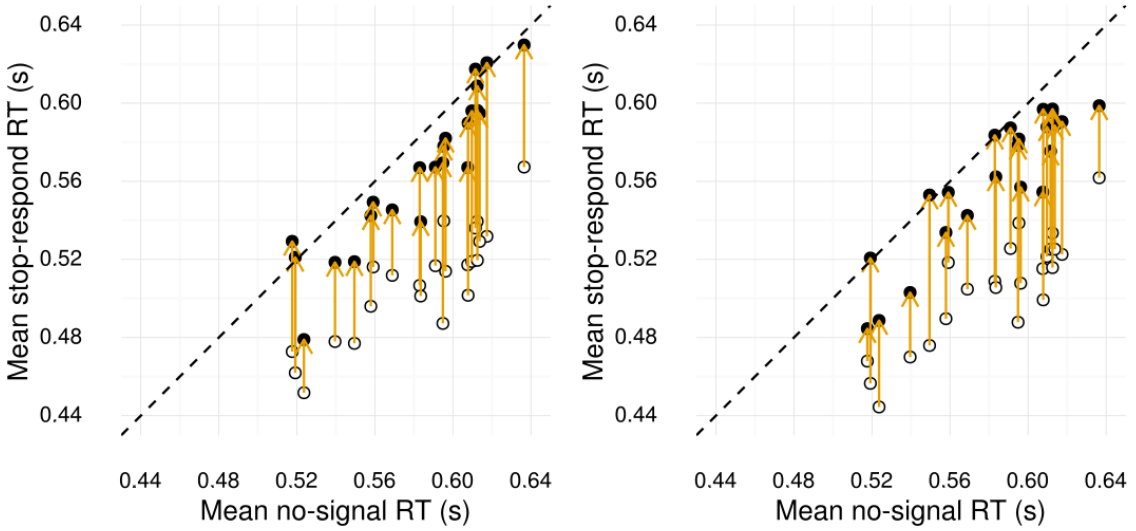


Fig. 11. Difference between the observed stop-respond RTs and stop-respond RTs predicted by the independent race model. Solid dots represent the observed mean RTs, open dots represent the stop-respond RTs predicted by the independent race model, given the observed no-signal RTs.

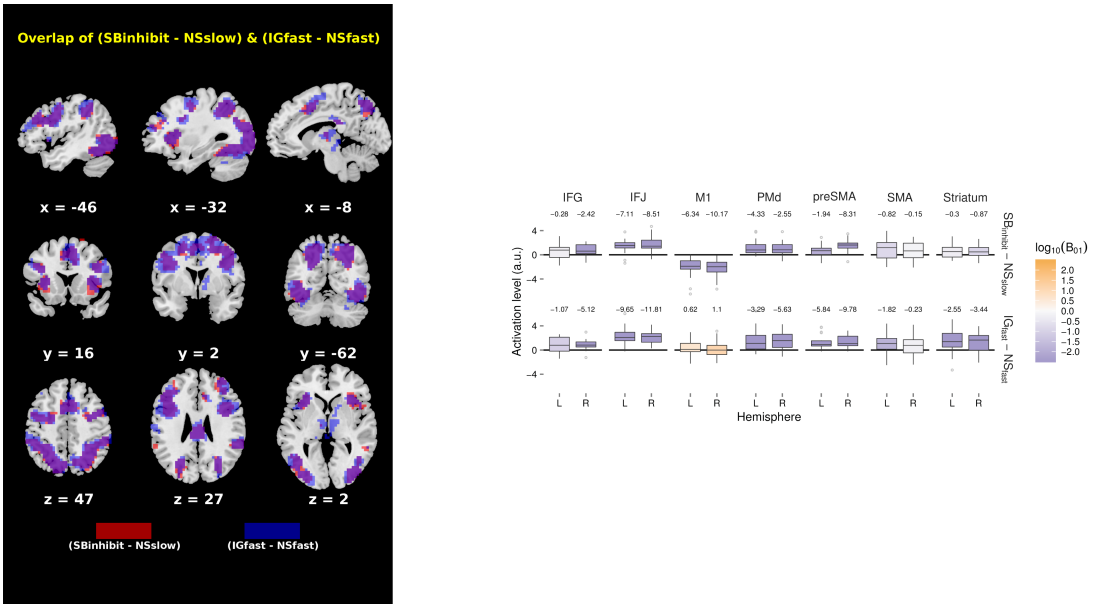


Fig 12. Activations in the S_{SS} subcontrasts $SB_{stop-inhibit} - NS_{correct-slow}$ and $IG_{correct-fast} - NS_{correct-fast}$. The left figure shows the overlap of the activations in the two contrasts in purple. A mask was created of each activation map and they are displayed on top of each other at 0.5 opacity. The right figure shows that activation levels (a.u.) within the functional ROis are similar in the two contrasts.

stop-respond RT did not increase with stop-signal delay. Although one study reported selective stopping performance in line with this prediction (Smittenaar et al., 2013), other studies did not test this assumption.

Violations were more dramatic at the individual level. Although, the prediction that the probability of responding increases with stop-signal delay held for all subjects, at least one of the other two predictions was violated in 96% of the subjects in AS stopping and 78% of the subjects in SS stopping. These results support previous work suggesting that stopping is indeed a race between a GO and a STOP process (Boucher, Palmeri, Logan, & Schall, 2007; Logan & Cowan, 1984; Mallet et al., 2016; Ramakrishnan, Sureshbabu, & Murthy, 2012; Schmidt et al., 2013), but that the assumption of independence between the two processes is violated in selective stopping (Bissett & Logan, 2014; De Jong et al., 1995).

Our results show that violations of the assumptions of the independent race model in individuals can be completely masked by the performance of the group as a whole. To illustrate, Figure 11 shows that even though the observed mean stop-respond RTs were much slower than the independent race model predicts, they were generally still faster than the mean no-signal RTs. Given the data of the group as a whole, the prediction that stop- respond RT is faster than no-signal RT was a thousand times more likely to be true than not in AS stopping, and even more in SS stopping. Nonetheless, testing this prediction for each individual unveiled that it was violated in about two thirds of the subjects in both AS and SS stopping. Estimated SSRTs for these subjects would be invalid, because the SSRT cannot be reliably estimated if this prediction is violated (Logan et al., 2014; Verbruggen & Logan, 2015).

Taken together, the results demonstrate that the independent race model's assumptions of independence often do not hold. The impact of these violations should not be underestimated, because nearly all studies of response inhibition rely on the independent race model, as they use SSRT as an outcome measure. Yet, only few studies report tests of the model's predictions and those that do report tests have performed them at the group level (e.g., Sebastian et al., 2015; Smittenaar et al., 2013). We urge users of stop tasks in general and selective stop tasks in particular to assess and report tests of the independent race model at the individual level and calculate estimates of SSRT if and only if individual datasets meet all qualitative predictions.

Action-selective and stimulus-selective stopping: more similar than different

Figure 6 shows that stopping performance was nearly identical for the two selective stopping types. Our results suggest that AS and SS stopping form a homogeneous construct, and neither involve selective stopping.

The evidence is two-fold. First, there was distinct response slowing on both AS and SS signal trials. The continued response on AS stop trials and the continued responses on ignore trials were both slower than no-signal RTs (see Fig. 2). Such response slowing has been reported before in both AS stopping (e.g., Aron & Verbruggen, 2008; Cai, George, Verbruggen, Chambers, & Aron, 2012; Coxon et al., 2012) and SS stopping (e.g., Bissett & Logan, 2014; Sebastian et al., 2015; Sharp et al., 2010). If subjects had selectively stopped the responses, the continued and ignore RTs should not differ from no-signal RTs.

Second, there was no difference between brain activation on stop-both trials and ignore trials. An exploratory analysis into the SS stopping contrast revealed that stop-both and ignore trials activated the same brain regions to the same degree (see Fig. 12). That implies that there was inhibition-related activity not just on stop trials, but also on ignore trials. Taken together, these results suggest that, instead of stopping selectively, subjects globally inhibited all responses when a signal occurred, and subsequently selectively re-initiated, or released inhibition of, the correct response (Aron & Verbruggen, 2008; Bissett & Logan, 2014).

There was a difference between AS and SS stopping, however, in the comparison of their associated brain activations. The functional and anatomical ROI analyses showed that the left and right PMd, the left PMv and the left SMA were activated in AS stopping and more so than in SS stopping (see Fig. 7, Fig. 8 and Table 7). Thus, AS stopping seemed to rely more on brain regions associated with motor planning than SS stopping. We speculate that these activations reflect action reprogramming on AS stop trials, rather than a difference in response inhibition.

If subjects applied non-selective, global inhibition in both AS and SS stopping, then AS and SS stop trials did not differ in terms of stopping. The difference that then remains lies in the re-initiation or continuation of the correct response. In SS stopping, on ignore trials, that correct response is the same as the initial response to the go-stimulus, but on AS stop trials action reprogramming is required: instead of a bimanual response, now only a left-hand or only a right-hand response must be made. In the fMRI contrasts that we used, the activity on ignore trials was subtracted from the activity on stop trials. Since, there appeared to be inhibition-related activity on ignore trials, this activity was subtracted from the inhibition- related activity on the stop trials. Thus, the AS stopping activations that survived the subtraction may reflect the action reprogramming that is required on AS stop trials but not on SS stop trials, rather than a difference in response inhibition. Previous findings of areas associated with action reprogramming are in line with the current observations of the PMd, PMv and SMA in the ROI analyses (Buch, Mars, Boorman, & Rushworth, 2010; Chambers et al., 2007; Coxon et al., 2016; Mirabella et al., 2011) and the precentral, and superior and inferior parietal areas in the whole-brain analysis (Mars, Piekema, Coles, Hulstijn, & Toni, 2007).

Conclusion

Nearly all selective stopping research has relied on the independent race model of simple stopping. In addition, selective stopping has been investigated with a heterogeneous set of tasks, including action-selective and stimulus-selective stopping paradigms, implicitly assuming that selective stopping is a homogeneous construct. However, it has been unclear whether the independent race model extends to selective stopping and whether selective stopping is a homogeneous or heterogeneous construct.

Our findings suggest that selective stopping can be modeled as a race, but not as an independent race. We found violations of the independent race model's assumptions in nearly every subject in both action- and stimulus-selective stopping. These individual violations were almost completely masked by the performance at the group level. We therefore urge users of stop tasks in general and selective stop tasks in particular to assess and report tests of the independent race model at the individual level. The results further suggest that action-selective and stimulus-selective stopping form a homogeneous construct, as subjects appeared to stop non-selectively rather than selectively on both trial types.

References

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience, 6*(2), 115–116.

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: Role of the subthalamic nucleus. *The Journal of Neuroscience, 26*(9), 2424–2433.

Aron, A. R., & Verbruggen, F. (2008). Stop the presses dissociating a selective from a global mechanism for stopping. *Psychological Science, 19*(11), 1146-1153.

Aron, A. R. (2011). From reactive to proactive and selective control: Developing a richer model for stopping inappropriate responses. *Biological Psychiatry, 69*(12), e55–68.

Bissett, P. G., & Logan, G. D. (2014). Selective stopping? Maybe not. *Journal of Experimental Psychology: General, 143*(1), 455–472.

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review, 114*(2), 376–397.

Buch, E. R., Mars, R. B., Boorman, E. D., & Rushworth, M. F. S. (2010). A network centered on ventral premotor cortex exerts both facilitatory and inhibitory control over primary motor cortex during action reprogramming. *The Journal of Neuroscience, 30*(4), 1395–1401.

Cai, W., George, J. S., Verbruggen, F., Chambers, C. D., & Aron, A. R. (2012). The role of the right presupplementary motor area in stopping action: Two studies with event-related transcranial magnetic stimulation. *Journal of Neurophysiology, 108*(2), 380–389.

Chambers, C. D., Bellgrove, M. A., Stokes, M. G., Henderson, T. R., Garavan, H., Robertson, I. H., Mattingley, J. B. (2006). Executive "brake failure" following deactivation of human frontal lobe. *Journal of Cognitive Neuroscience, 18*(3), 444–455.

Chambers, C. D., Bellgrove, M. A., Gould, I. C., English, T., Garavan, H., McNaught, E., … Mattingley, J. B. (2007). Dissociable mechanisms of cognitive control in prefrontal and premotor cortex. *Journal of Neurophysiology, 98*(6), 3638–3647.

Chen, C. Y., Muggleton, N. G., Tzeng, O. J. L., Hung, D. L., & Juan, C. H. (2009). Control of prepotent responses by the superior medial frontal cortex. *NeuroImage, 44*(2), 537–545.

Coxon, J. P., Goble, D. J., Leunissen, I., Impe, A. V., Wenderoth, N., & Swinnen, S. P. (2016). Functional brain activation associated with inhibitory control deficits in older adults. *Cerebral Cortex, 26*(1), 12–22.

Coxon, J. P., Impe, A. V., Wenderoth, N., & Swinnen, S. P. (2012). Aging and inhibitory control of ac- tion: Cortico-subthalamic connection strength predicts stopping performance. *The Journal of Neuroscience, 32*(24), 8401–8412.

Coxon, J. P., Stinear, C. M., & Byblow, W. D. (2006). Intracortical inhibition during volitional inhibition of prepared action. *Journal of Neurophysiology, 95*(6), 3371–3383.

Coxon, J. P., Stinear, C. M., & Byblow, W. D. (2009). Stop and go: The neural basis of selective movement prevention. *Journal of Cognitive Neuroscience, 21*(6), 1193–1203.

De Jong, R., Coles, M. G., & Logan, G. D. (1995). Strategies and mechanisms in nonselective and selective inhibitory motor control. *Journal of Experimental Psychology. Human Perception and Performance, 21*(3), 498–511.

Dimoska, A., Johnstone, S. J., Barry, R. J., & Clarke, A. R. (2003). Inhibitory motor control in children with attention-deficit/hyperactivity disorder: Event-related potentials in the stop-signal paradigm. *Biological Psychiatry, 54*(12), 1345–1354

Dimoska, A., Johnstone, S. J., & Barry, R. J. (2006). The auditory-evoked n2 and p3 components in the stop-signal task: Indices of inhibition, response-conflict or error-detection? *Brain and Cognition, 62*(2), 98–112.

Gauggel, S., Rieger, M., & Feghoff, T. A. (2004). Inhibition of ongoing responses in patients with parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry, 75*(4), 539–544.

Hanes, D. P., Patterson, W. F., & Schall, J. D. (1998). Role of frontal eye fields in countermanding saccades: Visual, movement, and fixation activity. *Journal of*

Neurophysiology, 79(2), 817–834.

Jahfari, S., Waldorp, L., Van den Wildenberg, W. P. M., Scholte, H. S., Ridderinkhof, K. R., & Forstmann, B. U. (2011). Effective connectivity reveals important roles for both the hyperdirect (fronto- subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibi- tion. *The Journal of Neuroscience, 31*(18), 6891–6899.

Janssen, T. W. P., Heslenfeld, D. J., Van Mourik, R. , Logan, G. D., & Oosterlaan, J. (2015). Neural correlates of response inhibition in children with attention-deficit/ hyperactivity disorder: A controlled version of the stop-signal task. *Psychiatry Research, 233*(2), 278–284.

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

Li, C.-s. R., Huang, C., Constable, R. T., & Sinha, R. (2006). Imaging response inhibition in a stop-signal task: Neural correlates independent of signal monitoring and post-response processing. *The Journal of Neuroscience, 26*(1), 186–192.

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review, 91*(3), 295–327.

Logan, G. D. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In D. Dagenbach & T. H. Carr (Eds), *Inhibitory processes in attention, memory, and language* (pp.189–239). Academic Press.

Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review, 121*(1), 66–95.

Macdonald, H. J., Stinear, C. M., & Byblow, W. D. (2012). Uncoupling response inhibition. *Journal of Neurophysiology, 108*(5), 1492–1500.

Majid, D. S. A., Cai, W., Corey-Bloom, J., & Aron, A. R. (2013). Proactive selective response suppression is implemented via the basal ganglia. *Journal of Neuroscience, 33*(33), 13259–13269.

Mallet, N., Schmidt, R., Leventhal, D., Chen, F., Amer, N., Boraud, T., & Berke, J. D. (2016). Arkypallidal cells send a stop signal to striatum. *Neuron, 89*(2), 308–316.

Mars, R. B., Piekema, C., Coles, M. G. H., Hulstijn, W., & Toni, I. (2007). On the programming and reprogramming of actions. *Cerebral Cortex, 17*(12), 2972–2979.

Mirabella, G., Pani, P., & Ferraina, S. (2011). Neural correlates of cognitive control of reaching movements in the dorsal premotor cortex of rhesus monkeys. *Journal of Neurophysiology, 106*(3), 1454–1466.

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J.-B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage, 25*(3), 653–660.

Paré, M., & Hanes, D. P. (2003). Controlled movement processing: Superior colliculus activity associ- ated with countermanded saccades. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 23*(16), 6480–6489.

Poser, B. A., Versluis, M. J., Hoogduin, J. M., & Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. *Magnetic Resonance in Medicine, 55*(6), 1227–1235.

Ramakrishnan, A., Sureshbabu, R., & Murthy, A. (2012). Understanding how the brain changes its mind: Microstimulation in the macaque frontal eye field reveals how saccade plans are changed. *The Journal of Neuroscience, 32*(13), 4457–4472.

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.

Ruiter, M. B. de, Oosterlaan, J., Veltman, D. J., Van den Brink, W., & Goudriaan, A. E. (2012). Similar hyporesponsiveness of the dorsomedial prefrontal cortex in problem gamblers and heavy smokers during an inhibitory control task. *Drug and Alcohol Dependence, 121*(1), 81–89.

Sallet, J., Mars, R. B., Noonan, M. P., Neubert, F. X., Jbabdi, S., O'Reilly, J. X., & Rushworth, M. F. (2013). The organization of dorsal frontal cortex in humans and macaques. *Journal of Neuroscience, 33*(30), 12255–12274.

Schall, J. D., & Boucher, L. (2007). Executive control of gaze by the frontal lobes. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 396–412.

Schmidt, R., Leventhal, D. K., Mallet, N., Chen, F., & Berke, J. D. (2013). Canceling actions involves a race between basal ganglia pathways. *Nature Neuroscience, 16*(8), 1118–1124.

Sebastian, A., Jung, P., Neuhoff, J., Wibral, M., Fox, P. T., Lieb, K., … Mobascher, A. (2015). Dissociable attentional and inhibitory networks of dorsal and ventral areas of the right inferior frontal cortex: A combined task-specific and coordinate-based meta-analytic fMRI study. *Brain Structure & Function, 221*(3), 1635-1651.

Sharp, D. J., Bonnelle, V., De Boissezon, X., Beckmann, C. F., James, S. G., Patel, M. C., & Mehta, M. A. (2010). Distinct frontal systems for response inhibition, attentional capture, and error processing. *Proceedings of the National Academy of Sciences, 107*(13), 6106–6111.

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

Smittenaar, P., Guitart-Masip, M., Lutti, A., & Dolan, R. J. (2013). Preparing for selective inhibi- tion within frontostriatal loops. *Journal of Neuroscience, 33*(46), 18087–18097.

Studholme, C., Hill, D., & Hawkes, D. (1999). An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition, 32*(1), 71–86.

Thakkar, K. N., Schall, J. D., Boucher, L., Logan, G. D., & Park, S. (2011). Response inhibition and response monitoring in a saccadic countermanding task in schizophrenia. *Biological Psychiatry, 69*(1), 55–62.

Van de Laar, M. C., Van den Wildenberg, W. P. M., Van Boxtel, G. J. M., & Van der Molen, M. W. (2010). Processing of global and selective stop signals: Application of donders' subtraction method to stop-signal task performance. *Experimental Psychology, 57*(2), 149–159.

Van de Laar, M. C., Van den Wildenberg, W. P. M., Van Boxtel, G. J. M., & Van der Molen, M. W. (2011). Lifespan changes in global and selective stopping and performance adjustments. *Frontiers in Psychology, 2,* 357.

Van den Wildenberg, W. P. M., Van Boxtel, G. J. M., Van der Molen, M. W., Bosch, D. A., Speelman, J. D., & Brunia, C. H. M. (2006). Stimulation of the subthalamic region facilitates the selection and inhibition of motor responses in parkinson's disease. *Journal of Cognitive Neuroscience, 18*(4), 626–636.

Van den Wildenberg, W. P. M., Burle, B., Vidal, F., Van der Molen, M. W., Ridderinkhof, K. R., & Hasbroucq, T. (2009). Mechanisms and dynamics of cortical motor inhibition in the stop-signal paradigm: A TMS study. *Journal of Cognitive Neuroscience, 22*(2), 225–239.

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences, 12*(11), 418–424.

Verbruggen, F., & Logan, G. D. (2015). Evidence for capacity sharing when stopping. *Cognition, 142,* 81–95.

Verbruggen, F., Aron, A. R., Stevens, M. A., & Chambers, C. D. (2010). Theta burst stimulation dissociates attention and action updating in human inferior frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 107*(31), 13966–13971.

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *NeuroImage, 18*(2), 293–309.

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

Zandbelt, B. B., & Vink, M. (2010). On the role of the striatum in response inhibition. PLoS ONE, 5(11), e13848.

Zandbelt, B. B., Bloemendaal, M., Hoogendam, J. M., Kahn, R. S., & Vink, M. (2012). Transcranial magnetic stimulation and functional MRI reveal cortical and subcortical interactions during stop-signal response inhibition. *Journal of Cognitive Neuroscience, 25*(2), 157–174.

Zandbelt, B. B., Buuren, M. van, Kahn, R. S., & Vink, M. (2011). Reduced proactive inhibition in schizophrenia is related to corticostriatal dysfunction and poor working memory. *Biological Psychiatry, 70*(12), 1151–1158.

Differential Prosocial Behaviour Without Altered Physical Responses in Mirror Sensory Synesthesia

Kalliopi Ioumpa[1]
Supervisors: Rob van Lier[1], Tessa M. van Leeuwen[1], Sarah Graham[2]

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands
[2] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

In synesthesia specific sensory stimuli lead to unusual, additional experiences. Mirror-sensory synesthetes mirror the pain or touch that they observe in other people on their own body. It has not been examined whether physical, bodily responses accompany it. We hypothesised that synesthetes would show deviations in hormonal levels and physiological responses when viewing arousing pictures. We expected more extreme ratings of pleasantness and arousal of the pictures due to the synesthetic experience.
Previous studies have yielded contradictory evidence on whether mirror sensory synesthetes demonstrate enhanced empathic behaviour. Altruism is related to empathy, but it has not yet been examined in mirror sensory synesthesia. We hypothesised that synesthetes would show more empathic and altruistic behaviour and enhanced theory of mind, and that this would relate to the physiological and rating responses to arousing pictures. We diagnosed mirror-sensory synesthetes with an established touch-interference paradigm and asked them to rate pictures with positive, negative, and neutral context for valence and arousal while their heart rate, skin conductance, and pupil dilation were recorded. Cortisol levels were assessed. Altruism was tested with a one shot Dictator's Game where participants divided money between themselves and a second player. Questionnaires on empathy, theory of mind, personality traits, and pain perception were completed. Eighteen mirror-sensory synesthetes and 18 controls participated. Our results show that mirror sensory synesthetes are more altruistic and more strongly impacted by positive and negative images. The stronger the reported synesthesia, the stronger the effect on the pictures ratings. We did not find evidence for differential physical responses to arousing pictures. Synesthetic experience does not alter overall perception about pain but makes mirror-sensory-synesthetes develop personality characteristics similar to synesthetes of other types (enhanced extraversion and openness to new experience). Synesthetes scored higher only on some of the empathy measures, highlighting the need for further investigation on the hypothesis.

Keywords: mirror sensory synesthesia, empathy, altruism, theory of mind, stress

Corresponding author: Kalliopi Ioumpa; E-mail: ioumpa.k@gmail.com

Synesthesia is the phenomenon of specific sensory stimuli leading to unusual, additional experiences. One of the most common examples is grapheme-colour synesthesia, where, for instance, the black written letters of a text would have a specific colour for an individual (Hochel & Milán, 2008). Synesthesia has a prevalence of 4% in the general population and there are three main aspects characterizing it (Grossenbacher & Lovelace, 2001). First, the experiences are elicited by particular stimuli that would not evoke such experiences in most members of the population; the inducing stimuli can be perceptual or conceptual. Second, the concurrent experiences are automatic, involuntary and are extremely difficult to suppress. Finally, the nature of the synesthetic experience is similar to that of a conscious perceptual event. It has been found that synesthesia is more common among people with autism (Baron-Cohen et al., 2013; Neufeld et al., 2013) and that synesthetes (individuals with synesthesia) have several personality traits in common such as enhanced creativity, better memory for verbal material and openness to experience (Banissy et al., 2013). It is known that synesthesia is heritable and that there are genetic components involved (Hochel & Milán , 2008; Asher et al., 2009; Tomson et al., 2011). However, the type of synesthesia inherited from generation to generation does not necessarily need to be the same (Barnett et al., 2008). Specific chromosomal regions have been implicated and candidate genes suggested but there is no clear indication about the exact genetic cause (Gregersen et al., 2013; Asher et al., 2009; Tomson et al., 2011)

The Mirror Sensory type

In this study we focused on the subtype Mirror Sensory Synesthesia, which is characterised by the production of conscious experiences similar to another person's observed state. Synesthetes of this type, when seeing or imagining someone else being in pain or being touched, feel the sensation like it would be on their own body (Banissy & Ward, 2007). There are two variations of the condition that often co-occur, one for the experience of pain (Mirror Pain Synesthesia) and one for the experience of touch (Mirror Touch Synesthesia). In each case, synesthetes feel the observed sensation on their own body, localised at the same spot. For some synesthetes an observed touch on the left cheek triggers a synesthetic sensation on their left cheek (anatomical correspondence), but for others the synesthetic sensation is felt on the right cheek (as if they were looking in a mirror, a specular correspondence).

Mirror sensory synesthesia has been related with enhanced empathic behaviour (Banissy & Ward, 2007) and has a prevalence of 1.6% among the general population, making it one of the most common forms of synesthesia, along with grapheme-colour synesthesia (prevalence of 1.4%) (Fitzgibbon et al., 2012b). The mechanism suggested for its cause proposes increased activity in the tactile mirror system above a threshold for conscious tactile perception in synesthetes of this type, causing observed tactile perception to be consciously perceived (Holle, Banissy & Ward, 2013). The brain region of anterior insula (Blakemore, Bristow, Bird, Frith, & Ward, 2005) that is associated with self-processing, seems to play an important role along with primary and secondary somatosensory cortex, and left premotor cortex (Blakemore et al., 2005).

Empathy, altruism and theory of mind in mirror sensory synesthesia

Mirror sensory synesthetes have been shown to exhibit heightened empathy on the "empathic quotient" (Baron-Cohen & Wheelwright, 2004) questionnaire (Banissy & Ward, 2007; Goller, Richards, Novak, & Ward, 2013). Empathy involves feelings of sympathy and a desire to relieve another's suffering. After witnessing someone else's distress, unpleasant empathic arousal can motivate the observer to help the other in order to reduce his or her own distress and feel relieved as well (Waal, 2008). In mirror sensory synesthesia, heightened empathy would be expected and logical as individuals experience on their own body any unpleasant sensation observed on others. Thus they would be more sensitive to the misfortunes of others and would be more willing to relieve their suffering. However, a recent study by Baron-Cohen et al. (2016) revealed no differences in empathy between synesthetes (N = 46) and controls in the "empathic quotient" questionnaire (Baron-Cohen & Wheelwright, 2004). The authors suggested that enhanced empathic behaviour demonstrated by initial studies was due to their small sample sizes (N = 10 for Banissy & Ward and N = 23 for the Goller's et al. study). Among the measures used in Baron-Cohen's 2016 study were the "empathic quotient" and the theory of mind measure "reading the mind in the eyes test" (Baron-Cohen, Wheelwright, Hill, Raste & Plumb, 2001). Theory of mind is also relevant for mirror-sensory synesthesia. It refers to the ability to attribute mental states to other people

and make sense of their behaviour (Premack & Woodruff, 2010). Because of theory of mind we can understand that others have beliefs, desires, intentions, and perspectives that are different from one's own and can make predictions about them. It is considered as a form of social intelligence and overlaps with the term empathy as the latter can be seen as a special form of simulation (Baron-Cohen & Wheelwright, 2004). In order to shed more light on the empathy debate for mirror sensory synesthesia, in our study we asked our participants to complete the "empathic quotient", the "reading the mind in the eyes test" and additionally the "interpersonal reactivity index" (Davis, 1980) which also assesses empathic behaviour. In an attempt to quantify this phenomenological sensitivity to the environment we asked mirror sensory synesthetes to complete the Emotional Contagion Scale (Doherty et al., 1997).

Empathy is considered to be one of the main reasons why people adopt an altruistic behaviour (Waal, 2008). It is still debatable whether altruism could be considered as a trait co-inherited with other traits (Rushton, 1982). Heightened altruism in mirror-sensory synesthesia has not yet been established, but might be expected on the basis of the reported heightened empathy ratings. In this study we also wanted to explore this possibility.

Conclusions about human altruism can be drawn by observing human behaviour after natural disasters, in everyday life or through simulations of everyday situations. To stimulate empathic or altruistic behaviour in the laboratory, economical games based on game theory are being used (Kollock, 1998). Many versions of these games have been used. The basic idea involves one player dividing a sum of money between him- or herself and a second player who is either able to accept or reject the offer (Ultimatum game) (Croson, 1996) or just has a passive role (Dictator game) (Eckel & Grossman, 1996). A third player can be added and spend his or her own amount to punish players, making an unequal split (Fehr & Rockenbach, 2004). The behaviour of people who deny small offers in the ultimatum game, give big amounts of money in the Dictator game, or punish unequal sharing, could be interpreted as altruistic. The majority of participants offer a rather "fair" deal of money and greedy proposers are generally "punished". The observed differences do not have to do with the amount of money (Cameron, 1999). To assess altruistic behaviour in mirror sensory synesthesia we asked from our participants to play a one shot Dictator's game.

Physical reactions when experiencing mirror sensory synesthesia

Experiencing mirror sensory synesthesia could be unpleasant when synesthetes observe others in a painful situation. People respond to challenging situations with changes in their behaviour and autonomic and neuroendocrine parameters, aimed at recovering the disturbed homeostasis. The concept of stress refers to the physiological mechanisms responsible to maintain and restore the balance after such changes. One of the main stress-coping strategies involves alterations of the hypothalamic-pituitary-adrenal (HPA) axis release of glucocorticoids such as cortisol (Munck, Guyre & Holbrook, 1984). It has been shown that biochemical analysis of salivary cortisol is a good representative of the plasma or serum levels of cortisol (Nicolson, Storms, Ponds & Sulon, 1997). Physiological responses like heart rate, pupil dilation and skin conductance are also influenced and are a good indication of stress responses (Bradley, Miccoli, Escrig & Lang, 2008). In our study we aimed to test whether unpleasant synesthetic experience could elicit a physical stress response in mirror-sensory synesthetes. If that were the case, it would be worth investigating whether the strength of the bodily stress response would correlate with how empathic a synesthete would be. In order to test that, we asked our participants to rate pictures with arousing negative context while measuring their heart rate, pupil dilation and skin conductance. Cortisol levels were also assessed as an additional measure for stress. Pictures with arousing pleasant context were also shown so it could be checked if synesthetes would be more affected physically by them too. Our participants also completed the Zung self-report scale (Zung, 1965) for depressive behaviour. In this way we could examine whether their synesthesia causes depressive symptoms. If that would be the case, differential rates in empathy and in the rest of our measures could have been a result of this depressive behaviour.

Approach

In this study we characterised the empathic behaviour, altruistic traits, theory of mind, and stress responses of mirror sensory synesthetes. We hypothesised that people with mirror sensory synesthesia would be likely to score higher on tests assessing empathic-altruistic behaviour and theory of mind. Moreover, individuals with mirror sensory

synesthesia might develop higher cortisol levels, have larger physiological responses and be more affected while viewing pictures with arousing context. In order to address these questions, we assessed physiological responses (heart rate, pupil dilation and skin conductance) and cortisol levels while synesthetes were watching pictures with pleasant, unpleasant and neutral context. Mirror-sensory synesthetes were diagnosed with an established behavioural interference paradigm (Banissy & Ward, 2007). Questionnaires about empathy and theory of mind were completed and a one shot Dictator's game was played to assess altruistic behaviour. It has not been studied so far whether mirror-sensory synesthesia could result in an altered pain perception in synesthetes of this type. In order to address this question we asked mirror sensory synesthetes to complete the Situational Pain Questionnaire (Clark & Yang, 1983) for pain perception. In Banissy et al. study of 2013 about personality characteristics in synesthesia, there was no mirror sensory synesthetes taking part. Thus, we asked our participants to complete The Big Five Inventory (John & Srivastava, 1999) in order to test for any personality differences in the effect of this type of synesthesia. Exploratory analyses for all questionnaires were performed.

Methods

Participants

Synesthetes were recruited through synesthesia associations and announcements in the media. Control participants were recruited through the university platform for recruiting participants for research studies (SONA system) and printed announcements at various locations.

We interviewed the people who contacted us via emails, asking whether they experienced mirror sensory synesthesia for pain and/or touch, whether they had experienced it all their life, and whether the experiences were consistent and stable over time. We also asked about their age and overall health condition. Only synesthetes experiencing the observed pain or touch on their own body at the same location as in the observed situation were invited to our laboratory. Some individuals were excluded due to severe comorbid disorders. In our laboratory 36 individuals, namely 18 synesthetes (M_{age} = 44, SD = 14.72) and 18 controls (M_{age} = 42, SD = 14.40), all women, completed the tests. We did not have any gender exclusion criteria during participants' recruitment. The groups did not differ in age ($t(28)$ = 0.2, n.s.). All participants received information about the study prior to participating and gave informed written consent to the study. Ethics approval was obtained from the local Ethics Committee of the Faculty of Social Sciences (ECSS) of Radboud University Nijmegen.

Laboratory experiments

During their visit to the laboratory, participants completed two behavioural experiments and a one-shot Dictator's game. The total duration was 2 hours. During the first experiment (Experiment 1) subjects viewed pictures with a negative, positive or neutral context and pictures of emotional faces of either a negative or positive context while their heart rate, skin conductance and pupil dilation were recorded to monitor stress responses for each picture. After the presentation of each picture, they were asked to give a rating on how pleasant or arousing they found it. Because this experiment was hypothesized to affect the overall mood and stress response of the participants, saliva samples for cortisol assessment were collected before and after the experiment, and a mood questionnaire was completed before and after the experiment. There were maximally two time slots per day for the laboratory experiments, at 1.30pm and at 4pm, as cortisol levels are generally stable at this time of day with no significant changes due to circadian rhythm effects (Kudielka, Schommer, Hellhammer & Kirschbaum, 2004).

The purpose of the second experiment (Experiment 2) was to diagnose participants for the Mirror Touch synesthesia type. We used an established paradigm where the subjects had to report the real touch that they received from an electrical device while watching videos of other people and an object being touched in a congruent or incongruent manner (Banissy & Ward, 2007). Differences in the error rates and reaction times between synesthetes and controls allowed for the diagnosis of Mirror Touch Synesthetes.

Finally, participants completed an exit questionnaire with questions about how they felt during each part of the experiment, and played a one shot Dictator's game to assess altruistic behaviour. We now turn to each experiment in detail.

Experiment 1 - Reaction to arousing pictures

In this experiment, participants were asked to rate arousing pictures on a scale from 1 to 9 for arousal and valence using the Self-Assessment Manikin (SAM) scales (Bradley & Lang, 1994).

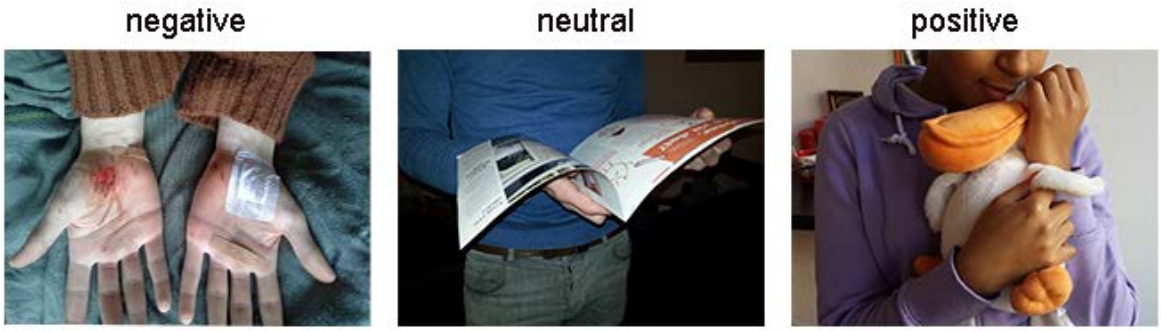


Fig. 1. Participants watched arousing pictures of negative, positive and neutral context. They were asked to rate each one of them for valence and arousal using a 9-point SAM scale. Physiological responses for each picture were also assessed.

At the same time, physiological responses (heart rate, pupil dilation and skin conductance) were recorded. Stress responses and mood alterations induced by this experiment were additionally assessed. Before leaving the laboratory, participants also rated the intensity of synesthetic pain and the synesthetic touch that they experienced during this part, on a 5-point Likert scale with (1) being "very slightly" and (5) "very much".

Stimuli. There were four stimulus categories: positive, negative, neutral and emotional pictures. The emotional condition included two subcategories of pictures with faces of positive and negative context. A negative picture would show injuries or painful events like a person having an injection or being cut with a knife. In the positive category pictures of people experiencing pleasant sensation of touch were included, such as holding hands or touching soft materials. Neutral pictures represented individuals performing everyday tasks like washing dishes or reading. Representative pictures for these three categories are illustrated in Figure 1. Finally, the emotional pictures included people with either negative (e.g. sad, angry, nervous) or positive (e.g. smiling, laughing) facial expressions.

Overall 90 pictures were presented, 20 of the negative, 20 of the positive, 20 of the neutral and 30 of the emotional condition (15 with pleasant and 15 with unpleasant mood states). The pictures were counterbalanced for landscape-portrait analogies and for the sex of actors across conditions, and it was ascertained that the luminance of the pictures did not significantly differ between conditions. Of the presented pictures, 54 were selected from the International Affective Picture System (IAPS) database for affective stimuli (Lang et al., 1999) and the rest were created by us.

For the stimuli that we created ourselves, valence and arousal ratings were not yet available, and we therefore asked 14 volunteers (ages: 21 to 64, 9 females) to rate them online using the 1-9 SAM scales for valence and arousal. Stimuli were presented using LimeSurvey software (https://www.limesurvey.org/). For the final stimulus selection, we included pictures with a valence rating (derived from the IAPS database or our pilot ratings) of $M_{negative}$ = 2.8, SD = 0.5 for the negative condition, $M_{positive}$ = 6.9, SD = 0.5 for the positive condition, $M_{neutral}$ = 5.5, SD = 0.6 for the neutral condition, $M_{emotion\_neg}$ = 3.7, SD = 1.2 for the emotional negative condition and $M_{emotion\_pos}$ = 6.7, SD = 1 for the emotional positive condition.

Subjective ratings. Each picture was initially presented alone on the screen for 6 seconds. After this time passed, the SAM scale for valence appeared below the picture. Participants had to type a number from 1 to 9 and press enter. After that, the SAM scale for arousal would appear and the participant had to complete it in the same way. If the overall duration of the trial at that point would be less than 20 seconds, a fixation cross would be presented until a 20 seconds length trial duration would be reached. In the case that after the second rating the trial would have already reached 20 seconds of duration the next trial would start right after the second rating. This interval of 20 seconds between the presentations of each picture was important for the physiological responses' measurement as skin conductance needs this amount of time to get back to baseline.

Before the actual experiment, participants had the chance to familiarize themselves with the task through 5 practice trials, and to ask questions. A chin rest was used and participants were instructed to move as little as possible in order to avoid motion artifacts in the physiological measurements.

Physiological responses. Heart rate, skin conductance and pupil dilation were recorded for the duration of the entire experiment. Heart rate (mV) and skin conductance (delta microsiemens) were measured with the BioPack Student Lab (https://www.biopac.com/education/) software using a sampling frequency of 1000 Hz. For the heart rate measurement, three electrodes were used. Two of them were placed on the participant wrists and one on the right ankle. For the skin conductance two electrodes were placed, one on the middle finger and one on the index finger. Gel was also applied at the touching points to facilitate the conductance of ions. For the pupil dilation assessment, the SensoriMotorInstruments (http://www.smivision.com/en.html) Red 3 eye tracker was used with a sampling rate of 500 Hz.

Heart rate during the presentation of the pictures (0-6 s) was determined for each condition and each participant. Data were analyzed using routines built in-house in Matlab 2013a (MathWorks). Raw ECG traces were epoched into segments of 6 seconds from picture onset until picture offset. Trials on which the heart signal exceeded 0.2 mV on average or sunk below -0.2 mV on average for more than two seconds during picture presentation were excluded from analysis. ECG data for the remaining trials were detrended and the peaks detected in order to calculate the number of beats per minute. On average, M = 63.4 (SD = 22.1) trials remained for each participant. Three synesthetes and one control were removed from analysis, as less than 40 trials each remained after cleaning. We statistically compared the heart rate (beats per minute) during the period of 0-6 seconds from picture onset for each experimental condition and between synesthetes and controls.

The peak of the skin conductance response during 8 seconds after each picture presentation was statistically compared across conditions and participants. Data were analyzed using routines built in-house in Matlab 2013a (MathWorks). The relative changes in skin conductance were computed as the difference between the amplitude of the peak and the value at the beginning of each trial. Trials on which the peak occurred during the initial 500ms of the trial (0-500ms), exceeded the baseline value by 10%, exceeded the value 4 or where there was a large jump in the value (>0.2) were excluded from analysis.

Relative pupil dilation was determined for each condition and each participant by dividing the pupil dilation at each moment during the stimulus period (0-6 s picture presentation) by the average pupil dilation during a 1 second pre-stimulus baseline (-1 until 0 s). Data were analyzed using routines built in-house in Matlab 2013a (MathWorks). Trials on which activity during the initial 200ms of the trial (0-200ms) dropped below or exceeded the baseline value by 10% were excluded from analysis. On average, M = 73 (SD = 13.7) trials remained for each participant. We statistically compared the pupil dilation during the period of 2-6 seconds after picture onset, when the pupil dilation had stabilised after the initial response to the change in luminance from the black fixation screen to each stimulus.

Cortisol measurements. Saliva samples for obtaining cortisol levels were collected at two time points during the laboratory visit and one more time at home. The first measurement at the laboratory took place right before the start of Experiment 1 and the second one at the first break of Experiment 2. As cortisol has its peak 20-30 minutes after the induction of stress we considered that as the best possible moment for the second measurement, since it was 25 minutes after Experiment 1.

Before arriving at the laboratory, participants had already received instructions for the cortisol sampling. In order to minimize differences in baseline cortisol levels, we instructed them not to brush their teeth, eat or drink anything but water for 1 hour before arriving, not to use any recreational drugs for 3 days and to refrain from drinking alcohol, exercising, and smoking for 12 hours as was instructed by the manufacturer's information (https://www.sarstedt.com/en/products/diagnostic/salivasputum/).

In order to gain a baseline measurement of the cortisol levels for each subject we asked them to perform an additional saliva collection at their home several days after their participation in the experiment. They received a saliva collection tube along with detailed instructions on how to use it during their visit to our laboratory. The saliva collection was done at the exact same time as the first (baseline) measurement at the laboratory and the saliva collection tube was mailed back to us via post. The same restrictions regarding food intake etc. were applied as during the lab visit.

Saliva was collected using the commercially available device of Salivette, Sarstedt. Participants first had to place the cotton swab from the tube in their mouth and chew gently on it for 1 min until it got humid. Then they placed it back in the Salivette tube. The samples were centrifuged at 1000 rpm for 5 minutes and then stored in a freezer at -20°C until analyzed. The samples were sent to Dresden LabService GmbH (http://www.labservice-dresden.de/) for cortisol determination in saliva.

Mood State. As we hypothesised that the arousing pictures in Experiment 1 might affect the participants' mood, the mood state of the participants was assessed using the Positive and Negative Affect Schedule (PANAS) questionnaire (Watson et al., 1988). Participants were asked to rate the extent to which they experienced each one of 20 emotions on a 5-point Likert Scale with (1) being "very slightly" and (5) "very much".

This measure was used at three time points: two times at the laboratory and one at the participant's home at the same time as the cortisol baseline saliva collection. In the laboratory, the first measurement took place right after the participant's arrival and the second after the end of the first experiment – so potential alterations in the participants' mood due to the presentation of arousing pictures could be detected.

Experiment 2 – Diagnosis for Mirror sensory synesthesia

In this experiment, participants were asked to report the location of actual touch (left, right, both, none) applied to their cheeks from an electric device while observing videos of another person or object being touched. This is an established paradigm for diagnosing mirror-sensory synesthesia (Banissy & Ward, 2007). Once again before leaving the laboratory participants also rated the intensity of synesthetic touch that they experienced during this part, on a 5-point Likert scale, with (1) being "very slightly" and (5) "very much".

Set up. Each video showed a boy, a girl, or an apple in one of the following conditions: touch on the right side, touch on the left side, touch on both sides, no touch. At the same time a tactile device applied real touch (left, right, both sides, or no touch) to the participant's face in a way that was perceived as congruent (observed touch same as felt) or incongruent (observed touch different from what was felt). Synesthetes, apart from the real touch, also felt the synesthetic touch on their face, which was the same as what was shown in the video. For potential synesthetes, congruency (specular or anatomical) was determined according to self-report after the presentation of pictures involving touching scenes. The tactile stimuli from the device were applied simultaneously with the ones in the videos. Participants had to report the location of the real touch by pressing one of the keyboard arrow keys (left for left felt touch, right for right felt touch, up for touch felt at both sides and down for no feeling of touch) as fast as possible. Their reaction times and error rates were recorded.

For synesthetes, during a congruent trial the touch was delivered to the same side of the face as the synesthetic experience. An example of an incongruent trial would be, receiving an actual touch on the right cheek but due to a synesthetic touch (elicited after the observed touch in a video) experiencing the feeling as being touched on the left cheek as well. In this case the correct answer would be 'right' and a 'mirror touch error' would be answering 'both'. An example of a no touch trial would be not receiving any actual touch from the device but observing touch in the video. This condition is hard for a synesthete experiencing the observed touch on his or her body whereas a control can easily reply 'none'.

Previous research has shown that synesthetes tend to get confused during the incongruent and no touch trials in this type of set-up. They make more errors and have longer reaction times than controls. They also seem to be faster at detecting congruent than incongruent touch in people and congruent touch in people than touch in objects (Banissy & Ward, 2007).

Before the actual experiment, participants had the chance to familiarize themselves with the task through 10 practice trials. After the end of each block, participants had a break where they could move freely and the white noise, which was played during the whole block to cancel the sound of the tactile device, was switched off. During the break after the first block, the second saliva measurement for cortisol level assessment took place.

Data were analyzed using routines built in-house in Matlab 2013a (MathWorks). Raw data were cleaned by rejecting RTs of incorrect trials and trials with RTs that were above or below 2 *SD* from the subject and condition mean. The number of outliers in the RTs was not different between the two groups ($F(1,30) = 0.29$, n.s.).

Stimuli. On average 81 congruent, 111 incongruent and 61 trials with no actual touch were used for each participant. For each one of the conditions 30% of the trials involved videos applying touch to a female actor, 30% of the trials touch to a male actor and 20% of the trials involved observed touch to an apple. The order of trials was randomised in three blocks. In our initial videos the touch was applied by a fingertip. After one synesthete reported feeling a sensation of touch at her own finger as well we made new videos

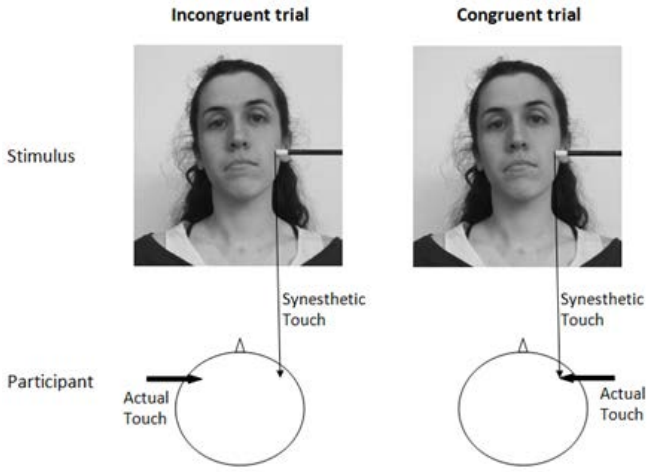


Fig. 2. Participants were asked to report the site of actual touch at their cheeks (left, right, both or none) while observing videos of a boy, a girl or an apple been touched. For mirror sensory synesthetes this task creates confusion because of the experience of synesthetic touch.

with the touch applied from a plastic stick similar to the tactile device to reduce the covariant factors. From all participants, 5 synesthetes completed the experiment with the initial fingertip videos and no differences were observed in their results.

The videos were presented on a 1700 CRT monitor with a refresh rate of 100 Hz and had an approximate duration of 4 sec. Participants could give their response while the video was still being played. The next video would only start after a response was recorded. There was a gap of 1,500 ms with a fixation cross before the start of the next trial.

Tactile device. The tactile stimuli were administered via an electrical device made in house. The touch was delivered via two plastic parts (one at each side) with round edges made in a way that they would resemble the feeling of a fingertip. Each one of the plastic parts was attached to a flexible plastic arm that would allow adjusting the device to the face. The plastic parts were attached to a surface supported by a microphone supporting rod. White noise was played via headphones during the whole experiment, so participants could not determine the location of the real touch due to mechanical noise of the device. In order to prevent the participants from moving (so the device would touch them at the same location during the entire experiment) a chin rest was used.

Dictator's game

After the end of, a one shot Dictator's game was played. Participants were given a big folder containing a sheet with instructions and two envelopes: one containing ten euros (in one euro coins) and one that was empty.

The instructions informed the participant that he or she was to be Player A in a game. Participants always had the role of Player A in the game but they were informed that the role was assigned by a random draw. Having the role of Player A meant that the participant had the chance to either keep the whole amount of money in the envelope or give some of it to Player B. The only information that participants had about Player B was that he or she was also someone participating in a research experiment at the same institute and that his or her role had also been assessed randomly. The participants were reassured that their response would be completely anonymous as neither the experimenters nor Player B would learn about their identity or about the amount that they had donated.

After making their decision, participants had to place the money to be given in one of the envelopes and put it in a box that was placed in the laboratory for this reason. They could keep the rest of the amount in the second envelope and take it with them.

Participants did not know anything about the existence of this part at the beginning of the experiment and were only told about it at the moment that the Dictator's game would take place. They were instructed to open the folder and follow the instructions. The experimenter left the room during this part so the participant would be alone.

Questionnaire completion

Participants received an email with a link that led them to several online questionnaires, created by LimeSurvey software (https://www.limesurvey.org/). The overall duration for completion was 1 hour and participants could save their responses at

any time and resume later. For Dutch participants, the Dutch version of each questionnaire was used while English versions were used for everyone else. Subjects were asked about their synesthetic experience, history of neuropsychiatric disorders, empathic behaviour, theory of mind, personality characteristics, sensitivity to other people´s states, pain perception and stress coping. The questionnaires were completed in the same order as the one presented below.

Personal information - Synesthesia. Questions about age, gender, education level, history of psychiatric, neurological, or endocrine disease and current use of psychoactive drugs or corticosteroids were completed. Participants were also asked about their synesthetic experience (van Leeuwen et al., 2010). Questions regarding the different types of synesthesia, the age of onset and the strength of synesthetic experience over the course of time were completed.

Empathy assessment. In order to assess empathic behaviour we used the Empathy Quotient (Baron-Cohen et al., 2004) and the Interpersonal Reactivity Index (Davis et al., 1980) questionnaires. The Empathy Quotient (EQ) consists of 60 questions and it is designed to measure empathy in adults. From these questions, 40 are clinically relevant and 20 are there to distract participants. There are three main subscales: for cognitive empathy, for emotional reactivity and social skills. Each statement has to be rated on the scale of: strongly disagree, slightly disagree, slightly agree, strongly agree. The minimum score is 0 and corresponds to least empathetic behaviour possible and the maximum score is 80 and corresponds to the most empathetic behaviour possible.

The Interpersonal Reactivity Index (IRI) consists of 28 statements and has four main scales assessing cognitive and affective aspects of empathy: Empathic Concern (EC), Personal Distress (PD), Fantasy (FS) and Perspective Taking (PT). The EC measures feelings of sympathy and compassion for others in distress, the PD self-oriented feelings of anxiety and distress in response to tense interpersonal situations, FS scale measures the tendency to project oneself into fictional situations and finally the PT scale measures the tendency to adopt the psychological point of view of others. Each item is rated on a scale ranging from "does not describe me well" to "describes me very well". For each one of the subscales, a minimum score of 0 and maximum score of 28 is possible. For PD, FS and EC subscales higher scores indicate enhanced empathy. For the PD subscale higher scores are translated to self-oriented emotional reactivity.

Theory of mind. The Reading the Mind in the Eyes (Baron-Cohen et al., 2001) is an advanced theory of mind test where the participants' ability to put themselves into the mental state of others can be examined. Participants were presented with 25 photographs of the eye-region of the face of different actors of both sexes, and were asked to choose one out of four words that was best describing what the individual in the photograph was thinking or feeling. Higher scores demonstrate enhanced theory of mind.

Personality characteristics. The Big Five Inventory (BFI) (John & Srivastava , 1999) has 44 items and is designed to measure the components of the Big Five personality traits which are Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Individuals have to indicate the extent to which each one of the personality traits mentioned describe their own characteristics using ratings from 1 to 5 with (1) for "disagree strongly" and (5) for "agree strongly".

Sensitivity to other people´s state. The Emotional Contagion (EC) Scale (Doherty et al., 1997) is a questionnaire with 15 statements made to measure how much an individual is influenced from the emotion and affective behaviour of others. The extent to which someone is mimicking the five basic emotions of happiness, sadness, anger, love and fear can be determined. Participants are asked to rate each one of the statements on how well it applies to them from 1 to 4 with (1) being "never" and (4) "always".

Pain perception. The Situational Pain Questionnaire (SPQ) (Clark & Yang, 1983) is used to evaluate how participants estimate their own sensitivity to pain. The extent to which a person is able to differentiate painful scenarios from neutral can be determined. The questionnaire consists of 30 statements describing 15 events that are considered to be painful and 15 that are considered to be non-painful. Subjects have to rate the events using a scale from 1 to 10 for (1) being "not noticeable" and (10) "worst possible pain".

Depressive behaviour. Depressive behaviour was assessed with the Zung self-report scale (Zung, 1965). The questionnaire consists of 20 items with affective, psychological and somatic symptoms that

have been related with depression. Each question can be rated in a scale from 1 to 4 with (1) for "a little of the time" and (4) for "most of the time". Scores can range from 20 to 80 with higher scores indicating more severe levels of depression.

Results

Experiment 1 - Reaction to arousing pictures

Of the 18 synesthetes and 18 controls that were invited to complete the laboratory tests, one control did not finish the experiments. She stopped during the first experiment due to inability to stay still and as the amount of data was too low for inclusion, this participant was excluded from all analyses. Furthermore, heart rate and skin conductance data were lost for one control due to technical failure and pupil dilation data were lost for one synesthete due to a technical failure.

Subjective ratings. Subjective ratings for the arousing pictures were computed from 16 synesthetes and 14 controls. Apart from the one control who did not finish the experiment, three controls were excluded as they did not make correct use of the rating scales. Two synesthetes were excluded as they reported having autistic traits.

In a repeated measures ANOVA with the within-subject factor Condition (negative, positive, neutral picture context) and the between-subject factor Group (synesthetes, controls) a significant interaction between Condition x Group was found for both valence $F(2,56) = 10.2$, $p < .001$ and (in a separate repeated measures ANOVA) for arousal ($F(2,56) = 6.4$, $p < .01$). Regarding valence, there was an effect of group for each stimulus condition (all $p < .05$). For arousal, the interaction was driven by the negative condition. The results are plotted in Figure 3.

As expected synesthetes were more influenced by the context of negative and positive pictures rating

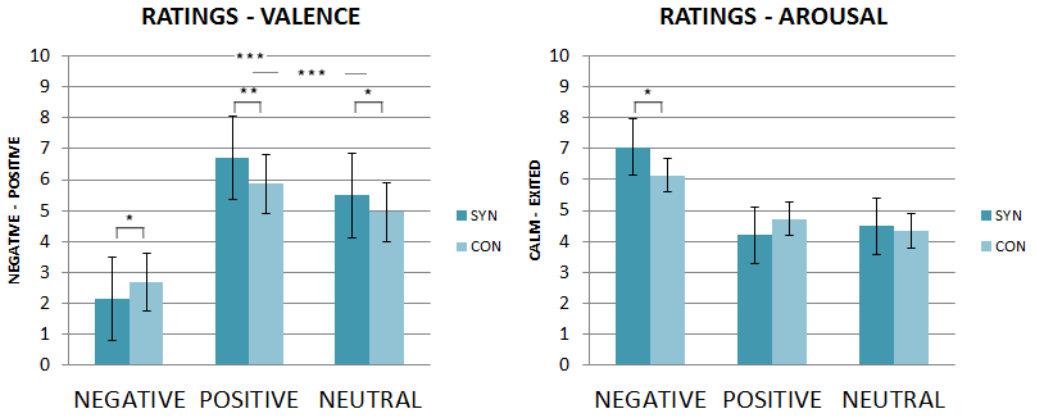


Fig. 3. Subjective ratings for arousing pictures of negative, positive and neutral context for valence (left) and arousal (right).

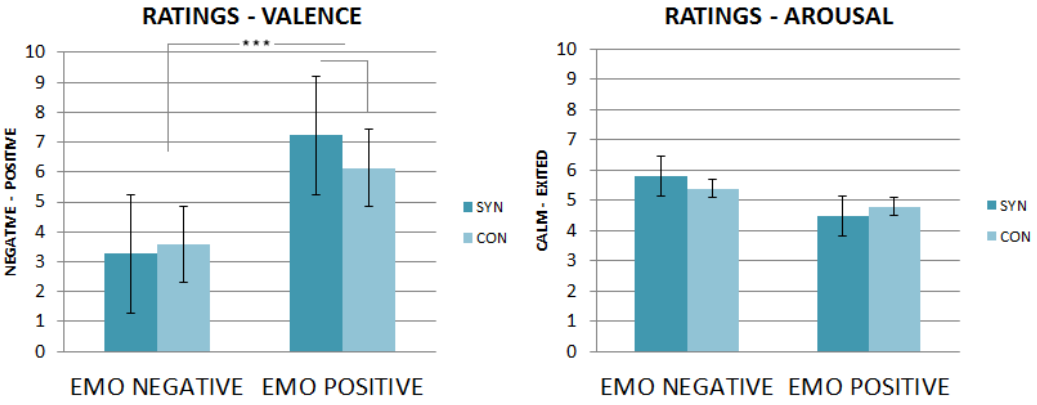


Fig. 4. Subjective ratings for arousing pictures of emotional faces of negative, positive and neutral context for valence (left) and arousal (right).

them as more unpleasant and pleasant respectively, than control participants did. According with our hypothesis, synesthetes also rated unpleasant pictures as more arousing and neutral ones equally arousing as control participants. We had not predicted that synesthetes would rate positive pictures as more pleasant than controls. Additionally, we were expecting that the positive pictures would have been rated as more relaxing and calming by synesthetes compared to controls, which was not the case.

Regarding the emotional faces conditions, a repeated measures ANOVA with the within-subject factor Condition (positive, negative emotional faces) and the between-subject factor Group (synesthetes, controls) revealed a significant interaction between Condition x Group ($F(1,28) = 17$, $p < .001$) for the valence ratings. A one way ANOVA showed that the interaction was driven by the positive emotional faces condition for valence $F(1,28) = 8.85$, $p < .001$. No effects were found for arousal. The results can be found in Figure 4.

These results show that synesthetes found the positive emotional faces more pleasant than controls did as we expected, but did not find the negative emotional faces more negative, as we had hypothesised. We also did not observe any differences for arousal between synesthetes and controls as we were expecting.

Physiological responses.

Pupil dilation. Data from 16 synesthetes and 17 controls were analyzed in this part as two synesthetes for whom less than 40 trials remained after cleaning were removed from analysis. In Figure 5 the results are plotted for the synesthetes (left) and controls (right). It can clearly be seen that the pupil dilation in the negative condition is larger than in the positive and neutral conditions. In a repeated measures ANOVA with the within-subject factor Condition (negative, positive, neutral picture context) and the between-subject factor Group (synesthetes, controls) a significant effect of Condition was indeed found ($F(2,62) = 9.63$, $p < .001$). No interaction between Condition x Group ($F(2,62) < 1$, n.s.) and no Group effect were found ($F(1,31) < 1$, n.s.). Post-hoc paired sample t-tests in which the data from the different conditions were compared showed that the positive and neutral condition pupil dilation did not significantly differ from each other ($t(32) = -0.97$, n.s.) while the pupil dilation in the positive condition and the neutral condition both significantly differed from the negative condition ($t(32) = -4.33$, $p < 0.001$ and $t(32) = 2.97$, $p < .01$, respectively).

Summarizing, we see the expected effect of the picture context with negative pictures inducing increased pupil dilation compared to positive and neutral ones, while there is no difference in pupil dilation responses between the groups.

For the emotional faces conditions, the results were very similar. The results are plotted in Figure 6. A repeated measures ANOVA with the within-subjects factor Condition (positive, negative emotional faces) and the between-subject factor Group (synesthetes, controls) revealed a significant effect of Condition ($F(1,31) = 17.6$, $p < .001$), but no interaction between Condition x Group ($F(1,31) = 1.37$, n.s.) and no Group effect was found ($F(1,31) < 1$, n.s.).

Heart rate. There were 15 controls and 15 synesthetes that had the required amount of artifact-free trials and remained in the analysis. In Figure 7, the average heart rate across the picture presentation period (0-6 s) is plotted for each condition and each group. A repeated measures ANOVA was run with

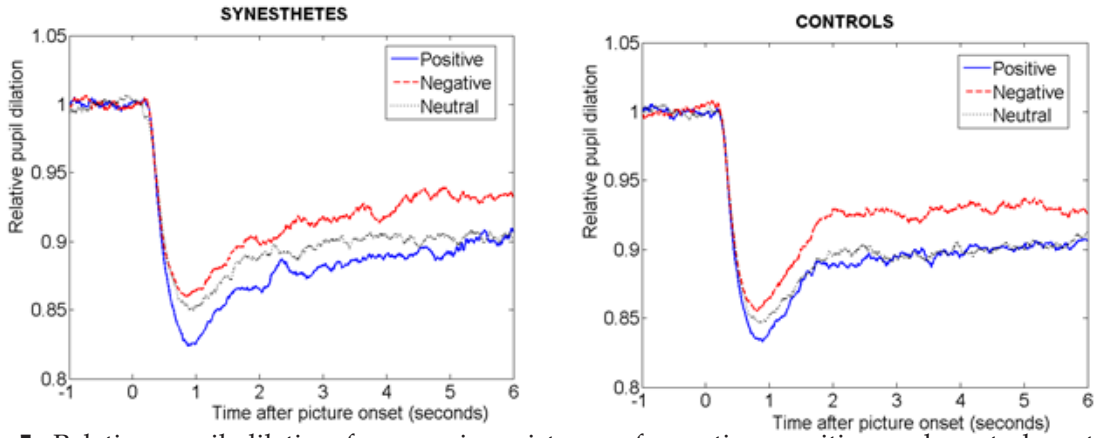


Fig. 5. Relative pupil dilation for arousing pictures of negative, positive and neutral context for synesthetes (left) and controls (right).

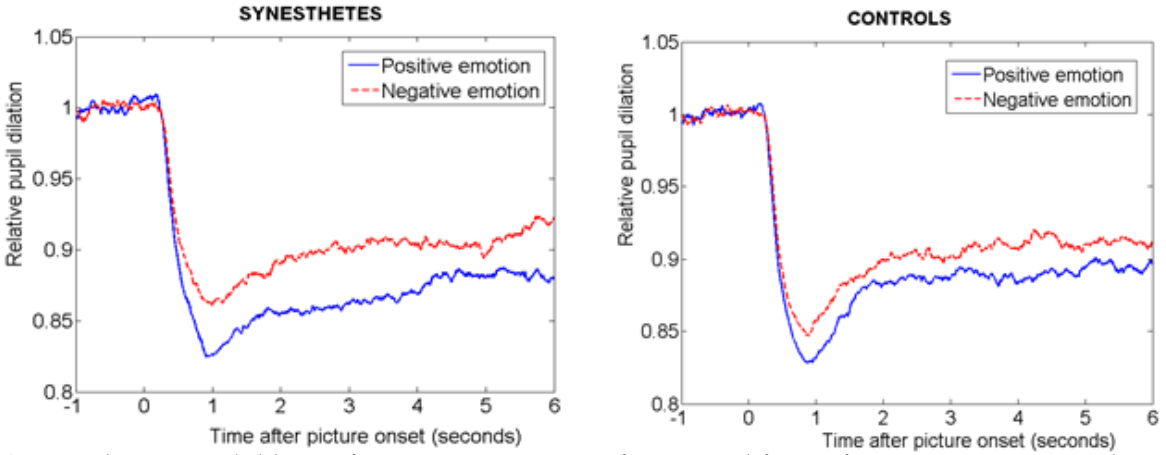


Fig. 6. Relative pupil dilation for arousing pictures of emotional faces of negative, positive and neutral context for synesthetes (left) and controls (right).

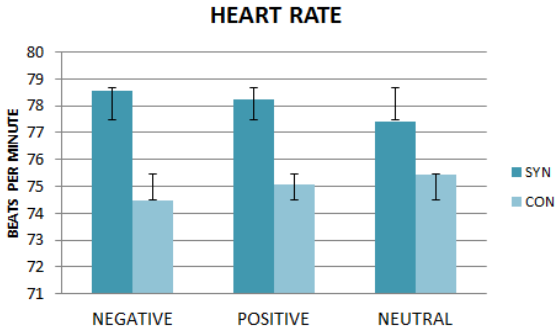


Fig. 7. Heart rate for arousing pictures of negative, positive and neutral context for synesthetes and controls.

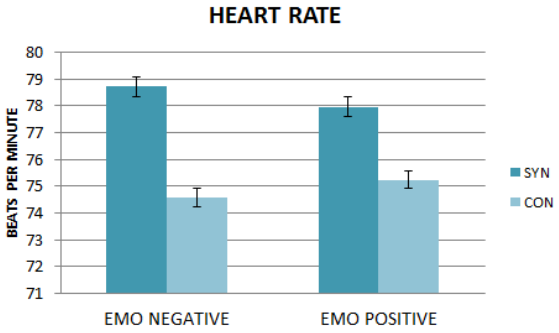


Fig. 8. Heart rate for arousing pictures of emotional faces of negative, positive and neutral context for synesthetes and controls.

the factors Condition (negative, positive, neutral picture context) and the between-subject factor Group (synesthetes, controls). No significant effects were found (Condition: $F(2,56) < 1$, n.s. and Group $F(1,28) = 1.09$, n.s.) but the interaction between Group and Condition was marginally significant at $F(2,56) = 2.27$, $p = .11$.

We can see in the plot as well that the interaction between Group and Condition is close to significance.

Synesthetes and controls seem to show a different pattern across conditions with synesthetes' heart rate going down moving from negative to neutral condition and the one of controls going up.

For the emotional faces manipulation, the results were highly similar. A repeated measures ANOVA was run with the factors Condition (negative, positive emotional faces) and the between-subject factor Group (synesthetes, controls). No significant effects were found (Condition: $F(1,28) < 1$, n.s. and Group $F(1,28) = 1.18$, n.s.) but the interaction between Group and Condition was marginally significant at $F(1,28) = 2.64$, $p = .12$. The results can be found at Figure 8.

Skin conductance. The galvanic skin response data did not yield any usable results. After cleaning of the raw data there were skin conductance data from only 7 synesthetes and 3 controls left with a sufficient number of trials ($N > 40$) for statistical analyses. We did not consider this amount of subjects sufficient for analysis.

Cortisol measurements. Cortisol results were obtained for 18 synesthetes and 16 controls. One control did not complete the cortisol measurements and for another control the data from two measurements were missing.

Baseline saliva cortisol levels, measured at the subjects' home, did not differ between synesthetes and controls ($t(32) = 1.20$, $p = .24$). For the cortisol values at Time 1 (before the experiment) and Time 2

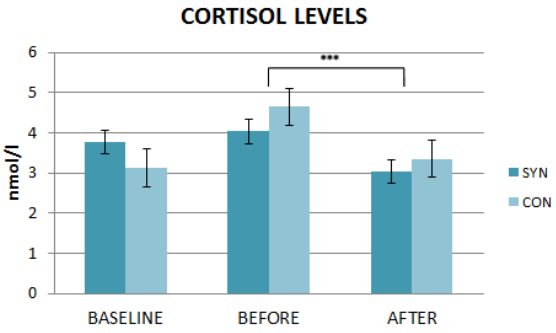


Fig. 9. Cortisol response induced from looking at arousing pictures of negative, positive and neutral context and of emotional faces of negative, positive and neutral context for synesthetes and controls.

(after the experiment) a repeated measures ANOVA was run with the within-subject factor Timepoint (Time 1, Time 2) and the between-groups factor Group (synesthetes, controls). A significant effect of Timepoint was observed ($F(1,32) = 12.7$, $p < .001$) that can be seen in Figure 9. No interaction between Timepoint and Group was observed ($F(1,32) < 1$, n.s.) and no Group effect ($F(1,32) < 1$, n.s.).

The cortisol values after the experiment are lower than before the experiment for all participants. This is unexpected, because the arousing pictures should have increased and not decreased the cortisol levels according to our hypothesis. It seems that participants were stressed in the beginning of the experiment when coming to the laboratory and more calm after the end of the first part. We can also observe that the change in salivary cortisol levels over time did not differ between the synesthetes and the controls, which is contradictive to our expectations as well. The baseline measurement at home did not differ from the one in the lab across participants and no baseline differences were observed between two groups, indicating that synesthetes as a group do not have higher overall levels and that the measurements on the day of the experiment at the laboratory were representative.

Mood state. In a repeated measures ANOVA with the within-subject factor Mood (difference of negative mood score before and after Experiment 1, difference of positive mood score before and after Experiment 1) and the between-subject factor Group (synesthetes, controls) no significant interaction between Mood x Group was found ($F(1,32) = 1.37$, n.s). This is not in line with our hypothesis as we were expecting that synesthetes' mood would be more affected compared to the one of controls after the presentation of pictures with arousing context.

In a separate repeated measures ANOVA with the within-subject factor Mood (difference of negative mood score before Experiment 1 and during the baseline measurement, difference of positive mood score before Experiment 1 and during the baseline measurement) and the between-subject factor Group (synesthetes, controls) no significant interaction between Mood x Group was found ($F(1,29) = 1.9$, n.s). This finding is expected and indicated that the synesthetes and controls did not differ significantly in their stress levels on the day that we performed our experiments in the laboratory. Thus any differences observed were only due to the experimental handlings and the tests performed.

Experiment 2 – Diagnosis for Mirror sensory synesthesia

Data from 15 synesthetes and 17 controls were analyzed in this part. Data from one synesthete were lost and two synesthetes were excluded for not following the instructions correctly. We analyzed both the error rates and the reaction times across the different experimental conditions.

For the error rates, in a repeated measures ANOVA with the within-subject factor Condition (Congruent, Incongruent, No_touch) and the between-subjects factor Group (synesthetes, controls), a significant interaction of Condition x Group was found ($F(2,60) = 4.46$, $p < .05$), as well as a significant effect of Condition ($F(2,60) = 10.4$, $p < .001$) and a significant effect of Group ($F(1,30) = 11.73$, $p < .01$). Independent samples t-tests revealed that the interaction in the error rates was driven by group effects in the incongruent ($t(30) = 2.97$, $p < .05$) and no_touch ($t(30) = 3.04$, $p < .05$) conditions, while there were no differences in the error rates for the congruent condition ($t(30) < 1$, n.s.). In Figure 10, the results are summarised: it can be seen that synesthetes make more errors than controls in both the incongruent and no_touch conditions. This is expected, since if the synesthetes were affected by their synesthesia, confusion would lead to more errors on the incongruent and no_touch conditions, but not on the congruent condition.

For the reaction times, in a repeated measures ANOVA with the within-subject factor Condition (Congruent, Incongruent, No_touch) and the

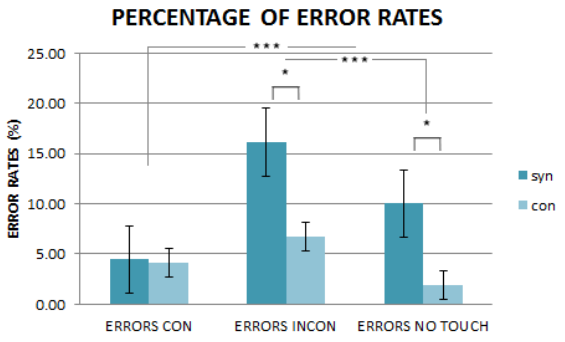


Fig. 10. Percentage of error rates for congruent, incongruent and no touch condition for synesthetes and controls.

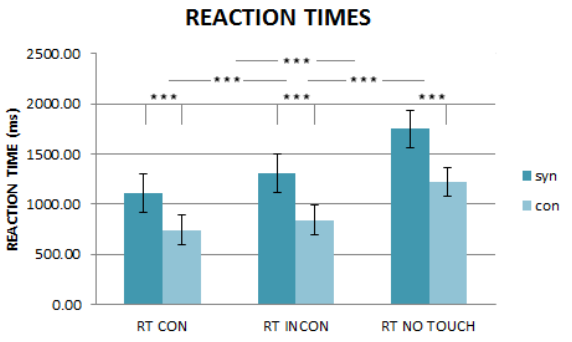


Fig. 11. Reaction times for congruent, incongruent and no touch condition for synesthetes and controls.

between-subjects factor Group (synesthetes, controls) there was no significant interaction of Condition x Group ($F(2,60) = 1.41$, $p = .25$), but there was a very clear effect of Condition ($F(2,60) = 72.6$, $p < .001$) and a significant effect of group ($F(1,30) = 13.4$, $p < .001$). The strong effect of Condition is mainly driven by longer reaction times in the no_touch condition, in which people did not receive a touch to the face but did observe touch in the video (see Fig. 11).

Longer reaction times in the no_touch condition were expected as participants waited until the end of the video waiting for any actual touch from the device. Synesthetes experienced synesthetic touch in this condition which was confusing when having to report what they felt as they had not experienced actual touch from the device. Synesthetes' overall delay that was observed in every condition could be due to an overall high level of confusion for synesthetes during this experiment.

Dictator's game

The results from the Dictator's game were computed from data of 17 synesthetes and 13 controls. Three controls and one synesthete were excluded, as they did not comprehend the instructions. We considered participants who delivered the envelope with the money for donation to ourselves and not to the box as instructed - while being alone in the room - as not having comprehended the instructions. In this way the condition of anonymity was not satisfied.

An independent samples t-test revealed a significant difference in the amount of money given between synesthetes and controls ($t(31) = 8{,}18$, $p < .05$) with synesthetes on average donating larger

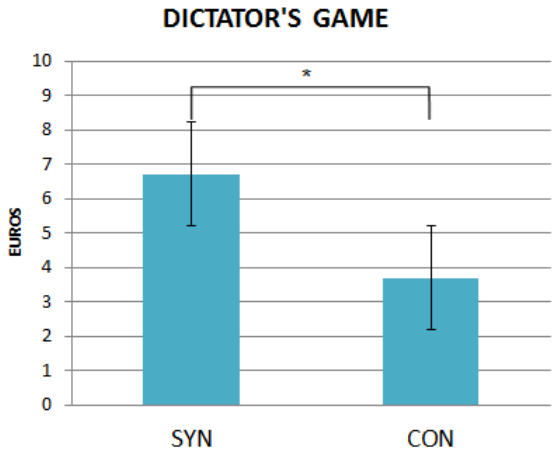


Fig. 12. Amount of money given in Dictator's game from synesthetes and controls.

amounts of money (6.71 vs 3.69 euros) as expected. The results are plotted in Figure 12. Age was significantly correlated ($r(35) = .49$, $p < .01$) with older participants giving larger amounts. Summarizing, as predicted, Dictator's game confirmed heightened altruism in synesthetes.

Questionnaire completion

The battery of online questionnaires on empathy and personal characteristics was completed by 11 synesthetes and seven controls who participated in the laboratory tests. Furthermore, synesthetes completed an exit questionnaire in the laboratory in which they reported on their experienced synesthesia during the experiments. Controls completed the exit questionnaire without the questions about synesthetic experience. First of all, the scores on the online questionnaires were compared between synesthetes and controls. Additionally, three sets of

correlation analyses were run. Apart from the first set of correlations that concerns only subjective rating that were completed during the laboratory tests, the rest of the correlations with all the online questionnaires should be regarded as exploratory.

The first set of correlations investigated whether there were correlations between the synesthetes' subjective ratings to arousing pictures (Experiment 1) and how strongly they rated their experience of synesthetic touch during Experiment 2 and their experience of synesthetic touch and pain during Experiment 1 in the exit questionnaire. The second set of correlation analyses compared the ratings of experiencing synesthetic touch in Experiment 2 and the ratings of experience of synesthetic touch and pain during Experiment 1 with the scores on the online questionnaires (for synesthetes only). Finally, we correlated the subjective ratings of the arousing pictures of Experiment 1 with the scores on the online questionnaires and this time all participants were included. For the sake of brevity only the statistics of the significant results will be reported.

Correlations between strength of synesthesia and valence/arousal ratings in Exp. 1

Participants who indicated to experience stronger synesthetic pain during Experiment 1 rated unpleasant images as more unpleasant ($r(11) = -.834$, $p = .001$), pleasant images as more pleasant $r(11) = .699$, $p = .017$, and rated unpleasant images as more arousing ($r(11) = .660$, $p = .027$). Synesthetes who reported experiencing stronger synesthetic touch (during Experiment 1) rated pleasant images as more exciting ($r(11) = -.611$, $p = .046$). Finally, synesthetes who indicated to experience stronger synesthesia for touch in Experiment 2 were less likely to report positive faces as unpleasant (marginal effect, $r(11) = -.580$, $p = .061$) but more likely to report positive faces as exciting (marginal effect, $r(11) = .532$, $p = .092$) during Experiment 1, although it should be noted that these correlations were not significant at the $p < .05$ level. These findings are expected as we had hypothesised that the stronger the synesthetic experience, the more affected by arousing pictures synesthetes would be.

Group Comparisons on Questionnaires

We conducted a series of one-way ANOVAs where we compared the Mean scores of the synesthetes (M_s) with the Mean scores of the Controls (M_c) for each completed questionnaire. We failed to find significant group differences on the Empathy Quotient (EQ) (and its subscales), The Reading the Mind in the Eyes experiment, the Situational Pain Questionnaire, and the Zung self-report depression scale. The analyses did reveal that synesthetes scored higher on the Empathetic Concern part of the Interpersonal Reactivity Index (IRI) ($M_s = 16.4$, $M_c = 11.7$; ($F(1,13) = 5.16$, $p = .041$) and on the total IRI score (IRI Total $M_s = 79.1$, $M_c = 61.0$; $F(1,11) = 6.42$, $p = .028$) but the other IRI subscales did not reveal significant group differences. On the Big Five Inventory (BFI) we found that synesthetes scored higher on extraversion ($M_s = 29.3$, $M_c = 23.6$; $F(1,13) = 9.09$, $p = .01$) and openness ($M_s = 41.7$, $M_c = 32.7$; $F(1,10) = 9.46$, $p = .012$). Finally, the synesthetes scored higher on almost all Emotional Contagion Scales (ECS) with the only exception being the sadness scale. We found the following statistics for the ECS scales: Happiness $M_s = 12.5$, $M_c = 10.0$; $F(1,13) = 4.97$, $p = .044$; Love: $M_s = 12.9$, $M_c = 9.6$; $F(1,13) = 6.20$, $p = .027$; Fear: $M_s = 12.1$, $M_c = 8.2$; $F(1,13) = 8.71$, $p = .011$; Anger: $M_s = 10.1$, $M_c = 7.2$; $F(1,13) = 8.74$, $p = .011$; Total EC score: $M_s = 58.4$, $M_c = 43.8$; $F(1,13) = 10.05$, $p = .007$. In line with our hypothesis synesthetes scored higher on ECS scales, at some parts of the IRI and the BFI scales for extraversion and openness (in agreement with what is already shown for synesthetes of other types (Banissy et al., 2013)). We were also expecting to observe different scores from controls for the EQ, Reading the Mind in the Eyes and in other subscales of the IRI, which was not the case.

Correlations between strength of synesthesia and online questionnaire scores

Participants who reported experiencing stronger synesthetic pain during Experiment 1 scored higher on the total IRI scale ($r(9) = .832$, $p = .005$) and participants experiencing stronger synesthetic touch (Experiment 1) scored higher on the ECS sadness subscale ($r(9) = .706$, $p = .033$) and the Zung depression scale ($r(9) = .795$, $p = .010$). The strength of experienced synesthetic touch during Experiment 2 did not reveal any correlations with any of the questionnaires scores. The observed correlations are expected as the stronger the synesthesia, the more empathic and sensitive to their environment we consider that synesthetes will be. We did not however find any significant correlations with other empathy measures as hypothesised.

Correlations between valence/ arousal ratings in Exp. 1 and online questionnaire scores

In these correlations, controls are included as well. Participants who rated unpleasant images as more unpleasant scored higher on the total IRI scale ($r(12) = -.663$, $p = .019$); the total ECS scale ($r(14) = -.533$, $p = .05$) and the Love subscale of the ECS ($r(14) = -.696$, $p = .006$). Participants who rated unpleasant images as more arousing also scored higher on the IRI fantasy subscale ($r(14) = .535$, $p = .049$); the BFI extraversion subscale ($r(13) = .617$, $p = .025$), and all of the ECS subscales except for happiness: Love: ($r(13) = .829$, $p = .000$); Fear: ($r(13) = .729$, $p = .005$); Anger ($r(13) = .660$, $p = .014$); Sadness: ($r(13) = .773$, $p = .002$); total ECS: ($r(13) = .802$, $p = .001$).

Participants who rated positive emotional faces as more pleasant also scored higher on the total IRI scale ($r(120) = .616$, $p = .033$), the pain perception scale ($r(14) = .535$, $p = .049$) and all the ECS scales except happiness and sadness. ECS Love: ($r(14) = .630$, $p = .016$); ECS fear: ($r(14) = .576$, $p = .031$); ECS anger: $r(14) = .546$, $p = .043$; ECS total: ($r(14) = .534$, $p = .049$). Subjects who rated positive faces as more exciting also scored higher on the emotional reactivity subscale of the EQ measurement. Finally, participants who rated negative faces as more arousing also scored higher on the fantasy subscale of the IRI ($r(14) = .604$, $p = .022$), the openness subscale of the BFI ($r(14) = .707$, $p = .007$), and three subscales of the ECS: Love: ($r(13) = .684$, $p = .010$); Anger: ($r(13) = .703$, $p = .007$), Sadness ($r(13) = .704$, $p = .007$).

It is expected that more empathic participants, as indicated from the IRI measure, would have more extreme ratings. It is also not surprising for individuals who are more sensitive to their environment, as indicated from the ECS measure, to have more extreme ratings as well. However we would also expect that more extreme ratings would also correlate with higher EQ and Reading the Mind in the Eyes test scores, which was not the case.

Discussion

In this study we characterised mirror sensory synesthetes with regard to their empathic and altruistic behaviour, theory of mind, personality characteristics and pain perception. We also examined whether experiencing this type of synesthesia is accompanied by physical responses of the body. In the case differential bodily responses would be observed for synesthetes, our question was whether any enhanced empathic or altruistic behaviour would accompany these effects.

We successfully diagnosed mirror-touch synesthesia with an established behavioural interference paradigm. Synesthetes did subjectively rate pictures with pleasant, unpleasant and neutral context more extreme than controls. These ratings were correlated with the strength of the synesthetic experience, as synesthetes subjectively reported it, demonstrating an effect of synesthesia in how much synesthetes are affected by other people's state. We did not find altered physiological responses (heart rate, pupil dilation and skin conductance) and cortisol levels for synesthetes during the viewing of arousing pictures. The Dictator's game revealed enhanced altruistic behaviour but a number of questionnaires about empathy, and the theory of mind test did not yield a significant effect. Mirror sensory synesthetes seem to have no alterations in their pain perception and demonstrate personality characteristics similar to the ones of synesthetes of other types.

The tactile experiment's manipulation – the diagnosis of mirror-touch synesthesia – was successful. Synesthetes got confused by the set up that included actual touch as well as synesthetic touch, making more errors than controls in exactly those conditions where their synesthesia was interfering, namely in the incongruent and no_touch conditions. Synesthetes were slower than controls for all stimulus conditions, but we did not observe any significant differential reaction times between synesthetes and controls for the different categories as it is reported in the literature (Banissy & Ward, 2007). A plausible reason for this may have been that participants were not pressured enough to answer very quickly. Thus, synesthetes may have taken more time overall as they found the experiment more difficult and confusing than controls. Had the factor of speed been emphasised more, they might have been faster in the congruent condition after all. For a number of synesthetes it was not very hard to distinguish between the synesthetic touch and the actual touch from the device, as the way they experienced these two forms of touch was very different. Thus, our paradigm did not work optimally for very sensitive synesthetes. In Baron-Cohen et al. (2016), it is reported that mirror sensory synesthetes also experience touch applied to objects on their own body. This is the first time that such an aim has been made. Further analyses of our data from the apple condition from Experiment 2 might contribute to finding out more about this scenario.

From the picture ratings in Experiment 1 we saw that synesthetes were more affected by arousing pictures than controls. After observing a picture with pleasant or unpleasant context synesthetes were getting more influenced which can be interpreted both as an increased sensitivity to other people's state and enhanced empathic behaviour. The fact that pictures ratings were correlated with the scores of IRI, EQ and ECS – measures of empathy – is an extra indication of this finding.

We can also see that the strength of the synesthetic experience (as it was subjectively rated) was highly correlated with the extent to which participants indicated to be affected by arousing pictures. Participants who indicated experiencing stronger synesthetic pain rated unpleasant images as more unpleasant, pleasant images as more pleasant and unpleasant images as more arousing. Also, synesthetes who reported experiencing stronger synesthetic touch rated pleasant images as more exciting. The synesthesia is thus very important for the ratings that were given and the way the participants experienced the pictures.

Turning to the physiological responses that were recorded during the pictures experiment, it is interesting to observe that for the pupil dilation, our manipulation worked, in the sense that there were different responses for each one of the stimulus categories in the expected direction. The participants' pupils were more dilated in the negative condition than in the positive and neutral conditions, indicating a response for the negative condition. We did not see an effect of group though, similar to the lack of a group effect in the other physiological measurements. It is a possibility that there is no physical response that accompanies the synesthetic experience. However, there are several reasons why our manipulation may not have been strong enough.

It could be the case that our pictures were not negative or positive enough to elicit physical responses and that we would have observed an effect if we would have used pictures with more extreme arousing context. This lack of affect can also be seen in the PANAS scores where synesthetes did not state a different emotional affect than control participants overall. Moreover as all pictures had a rather complicated background it can be that other aspects of the picture than the aspects that we wanted to emphasize have drawn the attention of the participants. This would mean that the subjective ratings would not necessarily only correspond to the synesthetic sensations elicited on participants' bodies. An answer to this question could be found by analyzing eye-movements and the location of fixation during the viewing of the picture. In sum, a larger sample size and more extreme pictures might have improved the set-up and yielded a group effect. We intend to perform further analyses separately for the physiological responses of the most extreme pictures from each condition as this might still reveal differential body responses in synesthetes and controls.

As we had hypothesised, the Dictator's game revealed enhanced altruistic behaviour for synesthetes. We observed a large effect with synesthetes donating significantly larger amounts of money than controls. It is the first time that the altruistic trait is been studied in mirror sensory synesthesia. A possible explanation is that because of being more susceptible to other people's state, synesthetes try to decrease (or avoid creating) any misfortunes around them as they are also directly affected themselves. An alternative interpretation involves the differential self-other distinction and representation of self-identity that has been found in mirror sensory synesthetes (Maister, Banissy & Tsakiris, 2013). Self-other merging can account for enhanced empathic and altruistic behaviour (Cialdini et al., 1997). Moreover, enhanced altruistic behaviour is demonstrated in in-group (enhanced self-other merging) compared to out-group circumstances (Güth, Ploner & Regner, 2009; Bernhard, Fehr & Fischbacher, 2006). Thus, increased altruism in mirror sensory synesthetes may be the outcome of a blurred self-other image.

Several concerns should be kept in mind when interpreting the results of the Dictator's game. Some synesthetes and controls donated more than the expected amount of money, which was 5 euros. This could be interpreted as either an actual extreme altruistic act or as a denial to play the game. A possible cause for the last explanation could be that participants were feeling insulted from this unexpected offer of money. Performing the analyses excluding the participants who donated more than 5 euros still resulted in a significant group difference ($t(29)=8{,}18$, $p<.001$), showing that the group effect is robust. The effect of age that we found could either be explained by the fact that older participants had a smaller need of money or because of altruistic behavior has been enhanced with age (Engel, 2011).We assume that the Dictator's game indeed measured the trait of altruism and that our set up was convincing for the participants. We cannot exclude the possibility that other feelings (e.g. feelings of insult) or lack of trust that the participants' response would indeed be anonymous contributed to their decision about the distribution

of money. Moreover it could be the case that our way of recruiting synesthetes played a role, as it was more personal than the one for controls. Before their visit to our laboratory synesthetes were extensively interviewed via emails.

Synesthetes scored higher on the Empathetic Concern part of the IRI, on the total IRI score, which can be interpreted as an indication for enhanced empathic behaviour. Interestingly, synesthetes scored higher on almost all Emotional Contagion scales with the only exception being the sadness scale, showing that they are indeed affected more by their environment and the different emotions around them. However, synesthetes did not score in a different way than controls in any other of the empathy measures or in the reading the mind in the eyes test which is in agreement with Baron-Cohen et al. results of 2016, but in disagreement with Banissy & Ward study of 2007. We believe that more studies are needed for a concrete conclusion on empathic behaviour in mirror sensory synesthesia. Larger samples and use of alternative measures of empathy might be crucial in shedding more light on the topic.

The results of the BigFive personality questionnaire revealed enhanced extraversion and openness to new experience in mirror sensory synesthetes which are traits that have been observed in the past for synesthetes with other types of synesthesia (Banissy et al., 2013). Another interesting observation is that experiencing mirror sensory synesthesia seems not to alter individuals' overall perception of pain, as it was measured in the Situational Pain Questionnaire. As the questionnaires were not completed from all the participants that were tested in our laboratory we cannot exclude the possibility of a type 1 error due to our small sample size.

Conclusion

In this study we have shown for the first time that mirror sensory synesthesia is accompanied by enhanced altruistic behaviour. Mirror sensory synesthetes indicated to be more strongly impacted by positive and negative images than control participants. Their synesthetic experience does not alter their pain perception but results in them developing personality characteristics that synesthetes of other types show. Further studies and additional analyses are needed in order to make solid conclusions about empathic behaviour and body responses in mirror sensory synesthesia.

Acknowledgements

We would like to thank Gerard van Oijen and Hubert Voogd for their precious help on constructing the tactile device and arranging the technical set-up. We are also grateful to the synesthesia associations and researchers who helped with recruitment and to all of our participants for their time and motivation.

References

Asher, J. E., Lamb, J. A., Brocklebank, D., Cazier, J.-B., Maestrini, E., Addis, L., Monaco, A. P. (2009). A whole-genome scan and fine-mapping linkage study of auditory-visual synesthesia reveals evidence of linkage to chromosomes 2q24, 5q33, 6p12, and 12p12. *American Journal of Human Genetics*, *84*(2), 279–285.

Banissy, M. J., Garrido, L., Kusnir, F., Duchaine, B., Walsh, V., & Ward, J. (2011). Superior facial expression, but not identity recognition, in mirror-touch synesthesia. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(5), 1820–1824.

Banissy, M. J., Holle, H., Cassell, J., Annett, L., Tsakanikos, E., Walsh, V., Ward, J. (2013). Personality traits in people with synaesthesia: Do synaesthetes have an atypical personality profile? *Personality and Individual Differences*, *54*(7), 828–831.

Banissy, M. J., & Ward, J. (2007). Mirror-touch synesthesia is linked with empathy. *Nature Neuroscience*, *10*(7), 815–816.

Barnett, K. J., Finucane, C., Asher, J. E., Bargary, G., Corvin, A. P., Newell, F. N., & Mitchell, K. J. (2008). Familial patterns and the origins of individual differences in synaesthesia. *Cognition*, *106*(2), 871–893.

Baron-Cohen, S., Robson, E., Lai, M.-C., & Allison, C. (2016). Mirror-Touch Synaesthesia Is Not Associated with Heightened Empathy, and Can Occur with Autism. *PloS One, 11*(8), e0160543.

Baron-Cohen, S., Johnson, D., Asher, J., Wheelwright, S., Fisher, S. E., Gregersen, P. K., & Allison, C. (2013). Is synaesthesia more common in autism? *Molecular Autism*, *4*(1), 40.

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251.

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group Affiliation and Altruistic Norm Enforcement. *American Economic Review*, *96*(2), 217–221.

Blakemore, S.-J., Bristow, D., Bird, G., Frith, C., & Ward, J. (2005). Somatosensory activations during the

observation of touch and a case of vision-touch synaesthesia. *Brain : A Journal of Neurology*, *128*(Pt 7), 1571–1583. \

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. http://doi.org/10.1111/j.1469-8986.2008.00654.x

Cameron, L. A. (1999). Raising the stakes in the ultimatum game: experimental evidence from Indonesia. *Economic Inquiry*, *37*(1), 47–59.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy-altruism relationship: when one into one equals oneness. *Journal of Personality and Social Psychology*, *73*(3), 481–494.

Clark, W.C., Yang, J. C. (1983). Pain Measurement and Assessment. *Psychological Medicine*, *14*(3), 717.

Croson, R. T. A. (1996). Information in ultimatum games: An experimental study. *Journal of Economic Behavior & Organization*, *30*(2), 197–212.

Davis,M., M. H. D. (1980). A multidimensional approach to individual difference in empathy.

Doherty, R. W. (1997). The Emotional Contagion Scale: A Measure of Individual Differences. Journal of Nonverbal Behavior, 21(2), 131–154.

Eckel, C. C., & Grossman, P. J. (1996). Altruism in Anonymous Dictator Games. *Games and Economic Behavior*, *16*(2), 181–191.

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, *14*(4), 583–610.

Fehr, E., & Rockenbach, B. (2004). Human altruism: economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology*, *14*(6), 784–790.

Fitzgibbon, B. M., Enticott, P. G., Bradshaw, J. L., Giummarra, M. J., Chou, M., Georgiou-Karistianis, N., & Fitzgerald, P. B. (2012a). Enhanced corticospinal response to observed pain in pain synesthetes. *Cognitive, Affective & Behavioral Neuroscience*, *12*(2), 406–418.

Fitzgibbon, B. M., Enticott, P. G., Rich, A. N., Giummarra, M. J., Georgiou-Karistianis, N., & Bradshaw, J. L. (2012b). Mirror-sensory synaesthesia: exploring "shared" sensory experiences as synaesthesia. *Neuroscience and Biobehavioral Reviews*, *36*(1), 645–657.

Goller, A. I., Richards, K., Novak, S., & Ward, J. (2013). Mirror-touch synaesthesia in the phantom limbs of amputees. *Cortex*, *49*(1), 243–251.

Gregersen, P. K., Kowalsky, E., Lee, A., Baron-Cohen, S., Fisher, S. E., Asher, J. E., Li, W. (2013). Absolute pitch exhibits phenotypic and genetic overlap with synesthesia. *Human Molecular Genetics*, *22*(10), 2097–2104.

Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*(1), 36–41.

Güth, W., Ploner, M., & Regner, T. (2009). Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? *Journal of Economic Psychology*, *30*(5), 814–827.

Hochel, M., & Milán, E. G. (2008). Synaesthesia: the existing state of affairs. *Cognitive Neuropsychology*, *25*(1), 93–117.

Holle, H., Banissy, M. J., & Ward, J. (2013). Functional and structural brain differences associated with mirror-touch synaesthesia. *NeuroImage*, *83*, 1041–1050.

John, O. P., & Srivastava, S. (1999). *The Big Five Trait taxonomy: History, measurement, and theoretical perspectives.* Guilford Press.

Kollock, P. (1998). Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, *24*(1), 183–214.

Kudielka, B. M., Schommer, N. C., Hellhammer, D. H., & Kirschbaum, C. (2004). Acute HPA axis responses, heart rate, and mood changes to psychosocial stress (TSST) in humans at different times of day. *Psychoneuroendocrinology*, *29*(8), 983–992.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 39-58.

Maister, L., Banissy, M. J., & Tsakiris, M. (2013). Mirror-touch synaesthesia changes representations of self-identity. *Neuropsychologia*, *51*(5), 802–808.

Munck, A., Guyre, P. M., & Holbrook, N. J. (1984). Physiological functions of glucocorticoids in stress and their relation to pharmacological actions. *Endocrine Reviews*, *5*(1), 25–44.

Nicolson, N., Storms, C., Ponds, R., & Sulon, J. (1997). Salivary Cortisol Levels and Stress Reactivity in Human Aging. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *52A*(2), M68–M75.

Premack, D., & Woodruff, G. (2010). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515.

Rushton, J. P. (1982). Altruism and Society: A Social Learning Perspective. *Ethics*, *92*(3), 425.

Tomson, S. N., Avidan, N., Lee, K., Sarma, A. K., Tushe, R., Milewicz, D. M., Eagleman, D. M. (2011). The genetics of colored sequence synesthesia: Suggestive evidence of linkage to 16q and genetic heterogeneity for the condition. *Behavioural Brain Research*, *223*(1), 48–52.

van Leeuwen, T. M., Den Ouden, H. E., & Hagoort, P. (2010). Bottom-up versus top-down: Effective connectivity reflects individual differences in grapheme-color synesthesia. In *FENS forum 2010 - 7th FENS Forum of European Neuroscience.*

Waal, F. B. M. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology*, *59*, 279–300.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–70.

Zung, W. W. K. (1965). Self-Rating Depression Scale in an Outpatient Clinic. *Archives of General Psychiatry*, *13*(6), 508.

Quantifying the Subjective Value of Distinct Working Memory Processes

Danai Papadopetraki[1]
Supervisors: Monja I. Froböse[1], Bram Zandbelt[1]

[1]*Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands*

Background: Distracter resistance is the ability to focus in the face of intervening stimuli and flexible updating is the ability to insert new stimuli into our working memory. Both working memory processes are important for adaptive living, yet we often observe performance failure. This failure has been traditionally attributed to a fixed cognitive capacity. Nonetheless, it has been consistently shown that motivation can strongly impact performance decrements. How does motivation affect performance? Value-based decision-making theories propose that task engagement is regulated by our brains via a cost-benefit analysis rendering task valuation key to understanding performance. Cognitive effort discounting studies quantified the subjective value of working memory tasks and showed that the attributed value decreased with increasing task demand. However, the subjective values of distracter resistance and flexible updating have not yet been quantified separately. Aims: First, we aimed to quantify the subjective values of distracter resistance and flexible updating, two key working memory processes. We hypothesised that the value of task engagement will decrease as a function of task difficulty. Our second aim was to assess if one of the two processes is valued more than the other. We hypothesised that distracter resistance is perceived as costlier. Methods: We designed a delayed-match-to-sample task to expose participants to increasing levels of distracter resistance and flexible updating. To quantify the subjective values, we employed a modified cognitive effort discounting task. Results: We provided strong evidence that subjective value decreases as a function of demand and weak evidence that distracter resistance is assessed as costlier compared to flexible updating. Discussion: Our results extend our knowledge about cognitive effort valuation and value-based decision-making. We corroborate other reports that people tend to conserve mental effort and suggest that distinct working memory processes can have differential subjective values.

Keywords: working memory, distracter resistance, flexible updating, value-based decision-making, effort-discounting

Corresponding author: Danai Papadopetraki; E-mail: danaepapadopetraki@gmail.com

Imagine the situation in which a student has to write an essay while her roommate is playing loud music. It is important for the student to remain focused on her essay despite the distraction of the appealing music coming from next door. Distracter resistance is the ability to focus in the face of intervening stimuli and robustly maintain current representations in working memory (Hazy, Frank, & O'Reilly, 2007). Now imagine that the student's roommate, instead of playing music, was calling for help because the house was on fire. In that case, it would be optimal for the student to switch her attention from writing the essay to the information coming from the roommate. Updating is the ability to flexibly insert new stimuli into working memory. Distracter resistance and flexible updating are two distinct working memory functions that are both required for adaptive living (Ernst, Daniele, & Frantz, 2011). Distracter resistance is crucial for maintaining our focus and completing our long-term goals. Flexible updating is important in order to adapt to environmental changes and to explore new opportunities (Hazy et al., 2007). Yet we often observe performance failure. For distracter resistance, performance failure usually occurs in the form of excessive distractibility. As for updating, people sometimes fail to update by exhibiting "stickiness", where information that is no longer relevant tends to persist in working memory.

Why do we often fail when working memory processes are involved? Traditionally, variance in working memory performance has been attributed to variance in cognitive capacity. The higher our working memory capacity, the better our performance in a specific task. However, these fixed models fail to explain situations in which performance can be improved by manipulating motivation or reward (Padmala & Pessoa, 2011). For example, monetary rewards reduce performance decrements that occur as a function of time on task. To account for such observations, newer, more dynamic models have been proposed that can incorporate factors like motivation and reward.

These models advocate that allocation of working memory resources is determined via a cost-benefit analysis where the costs of task engagement are weighted against the rewards (Botvinick & Braver, 2015; Kurzban, Duckworth, Kable, & Myers, 2013). Going back to our original example, the costs of time and/or effort of working on the essay could be weighed against the rewards of getting a good grade and learning more about the specific topic. As the importance (i.e., value) of the benefits increases, so does engagement in writing the essay. These value-based decision-making theories can account for reward effects on performance, but can also incorporate other likely contributing components like emotions, beliefs and past history.

If the above accounts hold, the valuation of working memory functions becomes crucial while trying to interpret and analyse human behaviour. Previously, it has been observed that when faced with a choice, participants preferred less cognitively demanding tasks (Kool, McGuire, Rosen, & Botvinick, 2010). In line with the cost-benefit theory, demand avoidance was reduced when monetary incentives were offered. Thus, all else being equal, people seem to perceive their cognitive effort as costly. This subjective cost of effort can be quantified using discounting paradigms, where the costs of a task are being measured as a function of rewards that participants are willing to forego (discount) in order to avoid performing the given task. Discounting paradigms have been applied extensively in the field of neuroeconomics and have lately been used to quantify physical and mental effort (Westbrook, Kester, & Braver, 2013; Massar, Lim, Sasmita, & Chee, 2016).

Such a cognitive effort discounting task was introduced in a recent study (Westbrook et al., 2013). Cognitive load was manipulated using the well-established N-back working memory task. Participants made choices between a higher level of the N-back task for a higher reward (i.e., more money) or a lower level for a lower reward (i.e., less money). The offer of the easy task at which participants were indifferent between the two options -their indifference point (IP)- was used as an estimate of subjective cognitive effort. The subjective value of the task decreased as a function of demand, suggesting that participants evaluate cognitive effort as costly enough to forego significant rewards in order to avoid it.

This was a landmark study for cognitive effort valuation, but a lot of questions remain unanswered. For instance, are all working memory functions/tasks perceived as costly? And, are some working memory functions perceived more costly than others? For example, previous studies have shown that participants perform better at flexible updating compared to distracter resistance (Fallon & Cools, 2014; Fallon, Van Der Schaff, Ten Huurne, & Cools, 2015). Does that difference partly reflect a difference in valuation? Using existing paradigms does not allow to address these questions. The N-back task requires both distracter resistance and flexible updating intermixed, while studies that have disentangled the two processes (Fallon & Cools,

2014; Fallon et al., 2015) were not designed to be discountable because they lack different levels of difficulty. Here, we aimed to address these remaining issues by designing a novel paradigm that allows to quantify the subjective values of distracter resistance and flexible updating individually.

Based on the above we proposed two research questions. 1) Are distracter resistance and flexible updating perceived as costly? We formulated the null and alternative hypotheses as follows. H_0: The subjective value of distracter resistance and flexible updating is not discounted by participants and the discounting does not increase as a function of demand. H_A: The subjective values of distracter resistance and flexible updating are discounted by participants and the discounting increases as a function of demand. 2) Is flexible updating perceived as less costly than distracter resistance? We generated the null and alternative hypothesis as follows. H_0: The subjective value of distracter resistance is the same as the subjective value of flexible updating. H_A: The subjective value of distracter resistance is smaller than that of flexible updating.

To address our two research questions, we designed a novel working memory paradigm that can evoke varying levels of distracter resistance and updating separately in different trials. During the working memory task participants experienced different demands of both relevant processes. Then using a cognitive discounting task, we quantified and compared the costs of the two working memory functions.

Methods

Participants

28 participants (19 women), aged between 18-29 years old were tested in total. Participants had normal or corrected-to-normal vision. Colour blind participants were excluded. The study was approved by the local ethics committee (CMO region Arnhem/Nijmegen, The Netherlands) and all participants provided written informed consent, according to the declaration of Helsinki. They were financially compensated by €8 per hour for their participation. Four data sets were discarded due to technical problems, so we ended up with 24 data sets (17 women, 18-29 years old, mean = 23.5, insert s.d.

Experimental design

The experiment lasted about 130 minutes and consisted of four tasks performed at a computer and questionnaires that participants filled in the end. The first task (~7 min) was a colour sensitivity test aiming to check if participants were sensitive to the colourful stimuli used later. Then participants proceeded with the colour wheel memory task to acquire experience with varying demand of the two working memory processes of interest (~10 min practice and 30 min task). The third task (~5 min practice and 55 min task) was a cognitive effort discounting paradigm that was used to estimate the subjective values and address our research questions. The last computer-based task was a redo of the color wheel task (~10 min). Finally, participants filled in some experiment-related questionnaires (~5 min).

Paradigms

All three paradigms were entirely programmed in MATLAB (release 2013a) using the Psychophysics Toolbox extensions (version 3.0.12) on a Windows 7 operating system. The screen resolution was 1920 × 1080 pixels. The background colour for all paradigms was grey (R: 200, G: 200, B: 200).

Colour sensitivity task. For our working memory task (described in 3.3.2) we used colourful stimuli and a colour wheel, so it was crucial that our participants' colour vision was not impaired. To test their sensitivity to our stimuli we developed a version of the colour wheel task without a working memory component.

The stimuli used for the colour sensitivity task were a colour wheel, black lines and coloured squares. The colour wheel was created by 512 successive coloured arcs of equal angle (512/360° = 1.42°), each arc carrying a different colour. The radius of the wheel was 486 pixels. To form the wheel into a ring a smaller circle was superimposed, whose radius was ~362 pixels. The centre of both the wheel and the circle coincided with the centre of the screen. The 512 colours of the colour wheel arcs were generated using the hsv MATLAB colourmap. The black lines were 0.4° black arcs.

In every trial of this task, participants viewed the colour wheel and a coloured square in the middle of the screen (Fig. 1). They were instructed to look at the colour of the square and use the mouse to click on the corresponding shade on the colour wheel. To indicate that their response was recorded a black line appeared on the colour wheel and successively another black line appeared designating the location of the correct colour. Feedback consisted of the actual deviance plus a positive message ('Good job! You deviated only __ degrees.') and was provided only when responses deviated less than 10°.

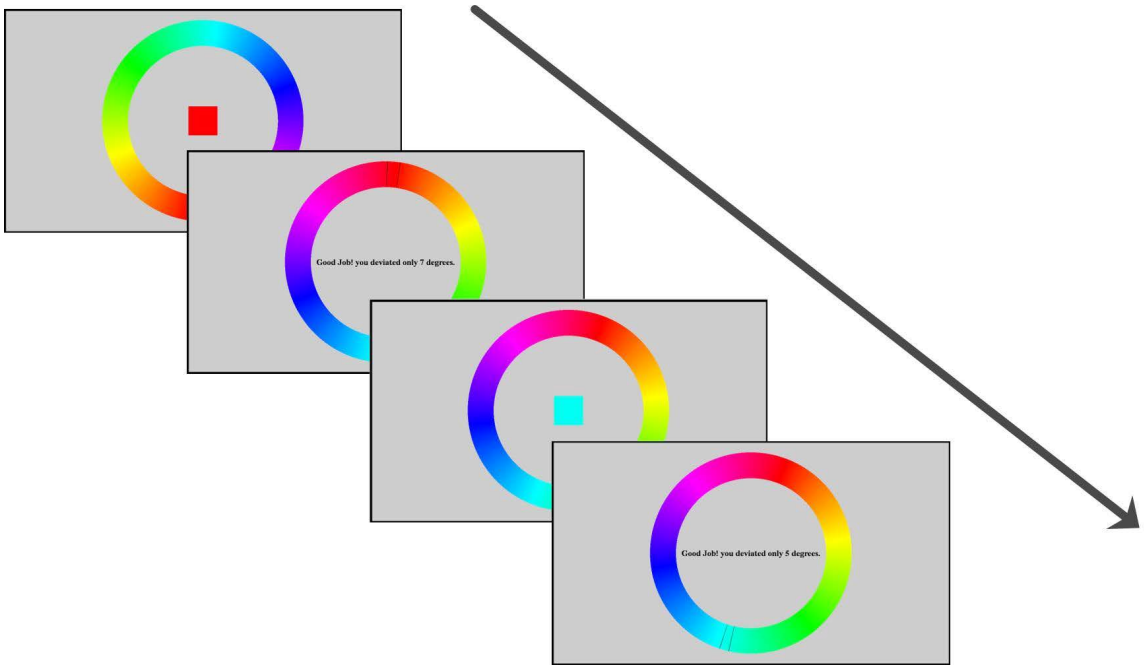


Fig. 1. Two example trials of the colour sensitivity task. Participants viewed a coloured square in the middle of the screen and they had to click on the corresponding colour on the colour wheel. A black line indicated the selected colour and a successive line the correct colour. If the selected colour deviated 10° or less from the correct colour they received feedback that they performed well. The task was self-paced and participants performed 24 trials. They successfully completed the task if their average deviance was less than 15°.

To test a representative sample of the colour wheel we split the wheel in 12 main arcs, each of which consisted of 512/12 colour categories. Participants were tested in two different shades from each colour category. So, they performed in total 24 trials of this task. The presentation of the trials as well as the orientation of the colour wheel was randomised. The responses were self-paced and total task duration was approximately 7 min.

The main dependent variable in this task was deviance in degrees from the correct colour. If their average deviance was less than 15° by the end of the task, the experiment continued. Otherwise, they had one more chance to perform the colour sensitivity task, but if failed again they would be excluded.

Colour wheel working memory task.

After successfully completing the colour sensitivity task, participants proceeded with the colour wheel working memory task. In this part, participants experienced varying demands of distracter resistance and flexible updating. This task was based on a short-recall task (Zhang & Luck, 2008) and delayed-match-to-sample tasks (Fallon & Cools, 2014) that have previously been used to disentangle between the two working memory processes of interest.

The stimuli displayed during this paradigm were a colour wheel, coloured squares, black frames of squares, a fixation cross, black lines and central letter cues. The colour wheel was generated as described above. The number of squares varied from one to four and they could be located in four different positions. The centres of the squares formed a rectangle with dimensions 248 × 384 pixels. Each of the four squares was 100 × 100 pixels in size. To choose the colours of the squares we split the colour wheel in 12 main arcs of 42 colours each and only used the 15 central colours of each arc. The arcs from which the colours would be sampled per trial were defined manually, but the exact shade (Red-Green-Blue values [RGB]) was randomly selected. The letter cues were "I" and "U", coloured black and presented at the centre of the screen.

Every trial of the task consisted of three phases separated by two delay periods (Fig. 2). During the encoding phase, participants viewed the fixation cross and one to four coloured squares for two seconds. The number of squares displayed (set size 1-4) represented the demand of the trial. A delay of two seconds succeeded, during which only the fixation cross was displayed. Then the interference phase followed. In this phase, participants viewed the same number of squares as during encoding, at the same locations, but with different colours. Instead of a fixation cross, one of the two letter

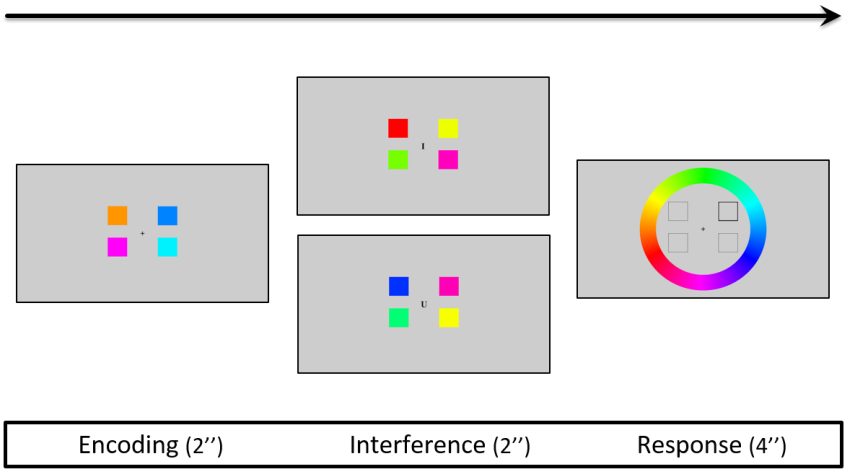


Fig. 2. An illustration of the colour wheel working memory task. Every trial of the task consists of three phases. In encoding phase (2 sec), participants need to memorise coloured squares. After a delay of 2 sec, during interference phase (2 sec), a letter indicates if it is a distracter resistance (I for ignore) or an updating (U for update) trial. In Ignore trials participants need to maintain in their memory the colours from encoding phase and not be distracted by the new intervening stimuli. In Update trials, participants have to let go of their previous representations and update into their memory the stimuli from interference phase. Another delay separates interference with response phase. This delay is 2 sec for ignore and 6 sec for update trials due to time differences in target stimuli. During response phase, participants see a colour wheel and black frames of the same squares; they have 4 sec to click on the target colour for the highlighted square. Demand is manipulated by varying the number of squares from one to four. The example displayed here is of the highest demand.

cues was presented during interference in the middle of the screen. The cue would be determined by the condition of the trial: "I" for distracter resistance trials and "U" for flexible updating. The second delay duration was also dependent on trial condition, being two seconds for distracter resistance and six for updating trials. Finally, during the *response* phase participants saw black frames of the same squares, one of which was highlighted, the colour wheel and the fixation cross. If the participant responded within four seconds, a black line appeared on the colour wheel, otherwise, they were instructed to respond faster. The total duration of response phase was five seconds.

For the encoding phase, participants were instructed to always memorise the colours and locations of all presented squares. The instructions for the interference phase differed based on the condition as suggested by the letter cue. In distracter resistance trials (referred to as ignore trials from now on), they needed to maintain in their memory the colours from encoding phase and not be distracted by the new intervening and distracting stimuli. In flexible updating trials (referred to as update trials from now on), participants had to let go of their previous representations and update into their memory the stimuli from the interference phase. Thus, the colours that needed to be remembered for distracter resistance were the ones from the encoding phase, while for updating trials the ones from interference phase should be remembered. As the encoding phase was four seconds before interference, the second delay was longer for updating trials. Participants were to indicate the colour for only one, highlighted square. They had to identify the target colour on the colour wheel and click using a mouse, within four seconds. Only the first response counted. A black line indicated their response. Only during practice trials, a second line appeared at the correct colour and positive feedback was displayed if they were performing well. During the task, no feedback was provided. Participants were instructed to fixate in the middle of the screen throughout the task. This instruction was given in order to dissuade them from adopting the strategy to close their eyes during ignore trials in order to avoid being distracted.

Participants first underwent a practice session of 16 trials and then performed two identical blocks of the task. A block consisted of 64 trials, resulting from repeating each combination of difficulty (four levels = set size 1-4) and condition (two levels = ignore and update) eight times. Depending on the difficulty level of the trial, a group of two to eight colours was used to create the trial stimuli, each colour coming from one of the 12 arcs.

Colours of the same arc never appeared more than once in the same trial. To make sure that ignore and update trials were as similar and counterbalanced as possible, the colours of the squares used and the target colours were the same for both conditions. However, as the relevant colours appeared during encoding for Ignore and during interference for update, we made sure that the same group of stimuli also appeared in reversed order between these two phases. So, the same groups of coloured squares were presented four times per set size and in total 32 groups of colours were used. To decrease learning effects due to repetition, we split the same stimuli groups between the two blocks. To control for differences between the two hemispheres in representation of colour (Gilbert, Regier, Kay, & Ivry, 2006), the left and right squares were equally highlighted for both conditions. Moreover, the same colours were highlighted for all four set sizes. The total duration of the colour wheel working memory task was approximately 40 minutes.

Discounting choice task. After participants gained adequate experience of all four levels of update and ignore via the colour wheel working memory task, they proceeded with the third part of the experiment: the discounting choice task. The aim of this paradigm was to quantify the subjective value participants assigned to the cognitive engagement they experienced during the colour wheel task. The design of the task was inspired by the temporal and cognitive effort discounting literature (Westbrook et al., 2013; Kable & Glimcher, 2007). There were two versions of choice trials to address our two research questions. In both versions two options accompanied with an amount of money were offered and the options defined what participants would do in the last part of the experiment. Both the rewards and the redo were real and not hypothetical to promote task validity.

The stimuli used for this paradigm were word cues, a fixation cross, a black rectangle and a black square. The rectangle was located in the centre of the screen with dimensions 1600 × 720 pixels. The square dimensions were 250 × 250 pixels. The word cues displayed were "Ignore", "Update", "No Redo" and "for"; the two first were accompanied by a number from one to four (referring to the set size) and the last one by a sum of money varying from €0.1 to €4.

In every trial of the task participants saw a rectangle containing two options and a fixation cross. The options could be "No Redo" or any set size of ignore or update, for example "Ignore 2". Below each option (60 pixels) a monetary reward was displayed, for example "for 2€". Participants could choose the left or right option by pressing 1 or 2 on the keyboard and they had six seconds to respond. When participants made a choice, a black square surrounding the selected offer appeared to indicate their response was recorded.

At this stage, participants were instructed that there were two more parts in the experiment. In the last part, they would have the opportunity to earn a bonus monetary reward by redoing one to three blocks of the colour wheel task. However, the amount of the bonus and the type of trials they would repeat would be based on the choices they made on the choice task. The framework in this task was extremely important because we wanted to minimise influences of research question irrelevant factors. To highlight the importance of every choice, we instructed them that of all the choices they made (of both versions) the computer would select only one randomly and the bonus and redo would be based on that single choice. To minimise effects of error avoidance in valuation, we informed participants that accuracy during the redo part would not influence whether they receive the monetary reward, as long as their performance was comparable to the first time they did the colour wheel task (part 2 of experiment).

Task vs No Redo: Choices between working memory task and no task. These trials addressed the first research question: whether distracter resistance and flexible updating subjective values decrease as a function of task demand. Here, participants had to choose between repeating a level of ignore or update (effort offer) and not redoing the colour wheel task at all (no-effort offer), see Figure 3A. "If they chose the "No Redo" option they were instructed that they would be able to use their time as they pleased (e.g., by using their phones) but they would still have to stay in the testing lab so that time spent on the experiment was the same for both options. Otherwise, if the option to repeat the task was selected, the redo trials would consist of mostly the selected choice condition and level. "Mostly" is important because if they always did the same condition during the redo, they would be able to predict whether they had to update or ignore. We emphasised that they should take their time, not rush their response and consider both the money and their experience while doing the colour wheel task.

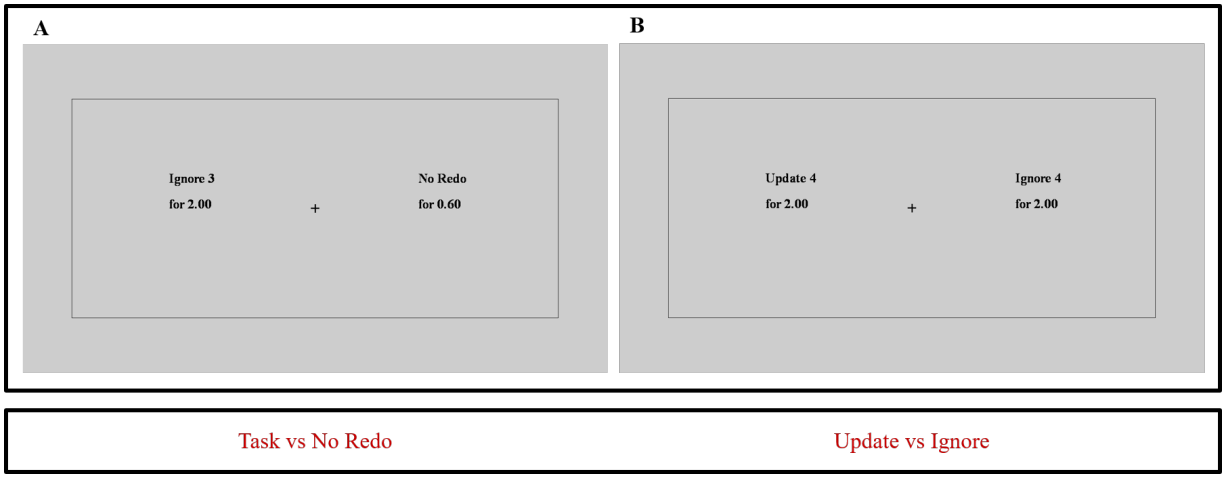


Fig. 3. Two versions of discounting choice task trials. **A.** Task vs No Redo. In this version participants have to choose between repeating a level of ignore or update and not repeating the colour wheel task at all ("No Redo"). The task option offer remains fixed at €2 and the "No Redo" option varies from €0.1 to €2.2. **B.** Update vs Ignore. For these trials participants have to choose between repeating either ignore or update of the same demand. Ignore offers are always fixed at €2 and update varies from €0.1 to €4. Trial duration is 6 sec.

Ignore vs Update: Choices between distracter resistance and flexible updating. This version aimed to investigate whether distracter resistance is perceived as costlier than flexible updating by directly contrasting them. In these trials participants had to choose between doing the same level of either ignore or update (Fig. 3B).

The amount they were offered for the "No Redo" option (no-effort offer) varied from €0.10 to €2.20 in €0.20 steps, while the task option (effort offer) was always fixed at €2.00. The €2.20 option for "No Redo" was included to verify whether there were participants who strongly preferred performing the task, even if that meant foregoing rewards. As we hypothesised that ignore would be costlier, in this case ignore (hard offer) was kept steady at €2 and update (easy offer) was varying from €0.10 to €4 in €0.20 steps. All possible pairs of options were 96 for "task vs no redo" choices (12 amounts × 2 conditions × 4 set sizes) and 84 for "ignore vs update" choices (21 amounts × 4 set sizes). As there is evidence that choice is probabilistic rather than deterministic (Rieskamp, 2008), every pair of options was sampled three times. We decided on three repetitions of the pairs based on a simulation analysis using pilot data (Online supplementary Fig. S1) in order to optimise the trade-off between indifference point estimation and task duration. Each participant performed three blocks that contained in total 288 trials of "task vs no redo" trials and 255 trials of "ignore vs update". There was a short practice session of 12 trials, where the amounts offered were the same for all options (€2) to avoid anchor effects. The trials of the two versions were interleaved (mixed) and randomised within each block. To avoid location effects, we counterbalanced the left-right presentation of the two options. So, for example, "No Redo" was presented left on half of the trials and right on the other half. Total task duration was about 55 minutes.

We decided to use fixed sets of offers and not a titrated staircase procedure to estimate subjective values because staircase procedures do not sample the entire logistic regression curve, rather the curve is estimated. Our version of effort discounting task sampled the logistic regression curves adequately because all participants were faced with the entire range of offer options.

Redo. After participants finished three blocks of the discounting choice task one of their choices was pseudo-randomly selected. Specifically, the computer only sampled from "ignore vs update" choices of level 3 or 4. Participants always did one block of 24 trials of the colour wheel task. Two-thirds of these trials were their preferred condition (ignore/update) and the set size was based on the parity of their subject number. We decided to never choose the "No Redo" option or the lower levels to maintain experimenter credibility for other participants. The redo data were not analysed and participants always received the bonus regardless of their performance.

Debriefing questionnaires. After the end of the experiment we requested participants to complete questionnaires. We explicitly asked them to report their preference by asking "Which trials did you prefer?".

Variables

Colour sensitivity task. The main dependent variable was accuracy as deviance in degrees from the correct colour.

Colour wheel working memory task. The independent variables for this paradigm were set size (four levels: 1-4) and condition (two levels: Ignore, Update) and we measured accuracy as deviance in degrees from the target and response times in seconds from probe onset as dependent variables.

Discounting choice task. For the discounting task, we measured participant choices and the independent variables were set size, condition and easy/no-effort offer.

Data analysis

We analysed our data using both frequentist and Bayesian statistics. All statistical analyses were performed using open source software JASP (version 0.7.5.6; Wagenmakers et al., 2016) on a Windows 7 operating system.

Table 1.

Bayes Factor Interpretation (Lee & Wagenmakers, 2013). At a value of 1, the data are inconclusive and we have no evidence to support either hypothesis. As the BF deviates from 1 evidence for either the alternative or the null hypothesis is enhanced.

B_{10}	Interpretation
> 100	Extreme evidence for H_1
30 – 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
3 – 10	Moderate evidence for H_1
1 – 3	Anecdotal evidence for H_1
1	No evidence
1/3 – 1	Anecdotal evidence for H_0
1/10 – 1/3	Moderate evidence for H_0
1/30 – 1/10	Strong evidence for H_0
1/100 – 1/10	Very strong evidence for H_0
< 1/100	Extreme evidence for H_0

As scepticism against classical statistical tools increases (Ioannidis, 2005), more and more scientists turn to alternative analysis methods such as Bayesian statistics (Wagenmakers et al., 2016). The main strength of Bayesian statistics is that they allow us to quantify evidence for our hypotheses instead of forcing an all-or-none decision and an arbitrary cut-off of significance. Bayesian statistics can also provide evidence for the null hypothesis, thus distinguishing between undiagnostic data ("absence of evidence") and data supporting H_0 ("evidence of absence"). Another important benefit is that we are able to monitor evidence as data accumulate and we can continue sampling without biasing the result. Due to all the above advantages, we decided that our main conclusions will be drawn based on the Bayesian analyses.

However, frequentist statistics are well-established and widely-acknowledged tools, so more scientists are familiar with their rationale and interpretation. To ensure that our results are interpretable by as many researchers as possible and to also compare their outcomes we additionally included classical statistics.

Bayesian statistics allow model comparison, but also provide evidence for individual effects. When possible, we reported Bayesian model comparison (BF_{10}: Bayes factor of model against the null) as well as Bayesian and frequentist effects analyses ($BF_{INC(LUSION)}$: Bayes factor of Bayesian model averaging). Refer to Table 1 for a Bayes Factor interpretation. We used the default JASP Cauchy priors for all Bayesian statistics (Wagenmakers et al., 2016). Regarding frequentist statistics, we considered a *p*-value of .05 or smaller as significant. In the cases where sphericity was violated, we reported the Greenhouse-Geisser corrected *p*-values.

Colour sensitivity task analysis. The data from this task was only used to establish that participants are sensitive enough to our colour wheel. We calculated the overall average deviance in degrees.

Colour wheel task data analysis. We computed the median deviance and median reaction time for all levels of ignore and update. The rationale behind choosing the median was that it is less sensitive to extreme values. For example, 90° and 180° accuracy scores are both wrong responses, but the latter affects the mean much more strongly. Then we used the above scores for the statistical analysis using classical and Bayesian 2 × 4 repeated measures analysis of variance (ANOVA) with condition (Ignore/Update) and set size (levels 1-4) as within-subject factors.

Table 2.

Descriptive statistics for colour wheel task deviance.

Condition	Set size	Mean	SD
Ignore	1	9.20	3.29
	2	10.55	4.96
	3	16.57	18.70
	4	16.11	12.34
Update	1	7.92	3.18
	2	7.99	3.05
	3	9.04	3.85
	4	10.96	6.48

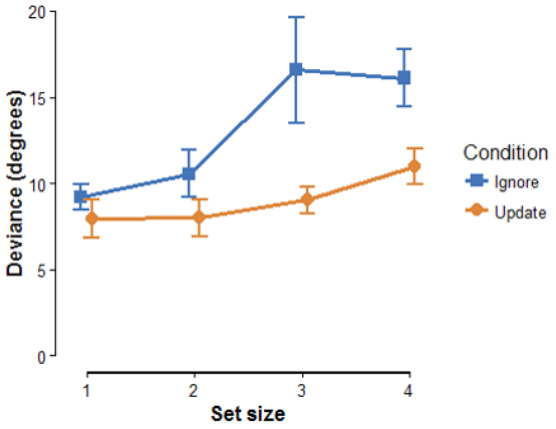


Fig. 4. Accuracy across 24 participants in colour wheel working memory task. Accuracy represented as deviance in degrees from the correct colour is displayed here as a function of set size for distracter resistance (ignore) and flexible updating (update) trials. The data are best explained by the model including condition and set size (BF_{10} = 1179). Error bars indicate the standard error of the mean (SEM).

Outlier criteria. As outliers, we defined participants performing below chance level (90° deviance) or whose accuracy in either condition was deviating more than 3 standard deviations from the mean.

Discounting choice task data analysis. As an estimate of subjective value, we computed participants' Indifference Points. The indifference points can be interpreted as the financial amount offered for the presumably less effortful option (no redo or update) at which participants are equally likely to choose one or the other, thus the probability of accepting either option would be .5. With the main dependent variable being choice, a dichotomous variable, we calculated the probabilities of accepting the presumably easier offer using binomial logistic regression analysis in MATLAB and extracted the indifference points for the different conditions.

Choices between working memory task and no task. Having determined the indifference points for all levels of both working memory tasks per participant, we continued with the statistical analysis using classical and Bayesian 2 × 4 repeated measures ANOVA to assess our first hypothesis that subjective value decreases with demand for distracter resistance and updating. Confirmation of this hypothesis would require that the model including set size is more likely than the null model or the presence of a set size effect with *p*-value smaller than .05. We also performed Bayesian and classical one sample t-tests on the indifference points across levels for both conditions to assess if the subjective value of the working memory functions was overall lower than the no task subjective value. The task offer was always €2 so a subjective value lower than 2 would imply that participants were discounting the task option.

Choices between distracter-resistance and updating. We then computed each participants' indifference points collapsing across levels of "ignore vs update" choice trials to evaluate our hypothesis that ignore has a lower subjective value than update using Bayesian and classical one sample t-tests. As ignore was set at €2, subjective values lower than 2 confirm that participants were willing to forego rewards to repeat update instead of ignore trials. Additionally, we calculated indifference points for all levels separately and used a 1 × 4 ANOVA with set size as a factor to assess if the preference for update varies with demand.

Exclusion criteria. Participants who consistently chose only one option (presumably easier or hard) would be excluded from the analysis, as we would not be able to estimate an indifference point for them. Similarly, participants who deviated more than 3 standard deviations from the mean were also excluded as outliers.

Results

Colour sensitivity task

All participants passed the colour sensitivity task and continued to the main paradigm. Their average

Table 3.

Model comparison for accuracy in colour wheel working memory task.

Models	p(M)	p(M | data)	BF_M	BF_{10}	% error
Null model (including subjects)	.2	5.685e-4	0.002	1	
Condition	.2	.041	0.173	73	0.82
Set size	.2	.006	0.023	10	0.88
Condition + Set size	.2	.690	8.914	1214	1.57
Condition + Set size + Condition × Set size	.2	.262	1.420	461	11.34

Note. All models include subjects.

deviance was 6.79 (SD = 1.20; median = 4.8, SD = 0.87) degrees. We report the median here as well for easy comparison with the colour wheel working memory task results.

Colour wheel working memory task

Having determined performance outside a working memory context, we analysed performance under conditions requiring distractor resistance and flexible updating. All participants performed overall above chance level (mean deviance less than 90°). Based on our criteria, no outliers were detected.

Deviance. Figure 4 shows colour wheel working memory task accuracy across set sizes for both conditions. See Table 2 for descriptive statistics. The Bayesian model comparison (Table 3) showed strongest support for the model including set size and condition (BF_{10} = 1179); the runner-up model was ~2.5 times less likely and it was the one including both main effects and their interaction (BF_{10} = 460). The effects analysis for deviance confirmed that accuracy decreased with increasing set size (F(3, 23) = 4.676, p = .022, BF_{INC} = 15.234) and that participants performed better at update compared to ignore trials (F(1, 23) = 9.986, p = .004, BF_{INC} = 104.358), see Table 4 for Bayesian effects. The interaction effect here was not significant (F(3, 1.597) = 2.329, p = .122, BF_{INC} = 1.420). Overall, the findings demonstrate that accuracy decreased with demand and was better for update trials.

Reaction times. Figure 5 depicts reaction times for ignore and update trials as a function of set size and Table 5 presents descriptive statistics for RTs. According to the Bayesian model comparison (Table 6), the best model was the one including condition, set size and the interaction between the two (BF_{10} = 1.667e+26). Effects analyses (Table 7) confirmed that participants were faster in ignore compared to update trials (F(1, 23) = 20.111, p < .001, BF_{INC} = 310), a very strong set size effect (F(1.6, 23) = 51.617, p < .001, BF_{INC} = ∞) and a strong interaction effect (F(2.22, 23) = 7.21, p = .001, BF_{INC} = 111). The analyses suggest that RTs varied with demand and that participants were faster for distracter resistance trials, but this difference depended on task demand.

Discounting choice task

Choices between task and no task. After analysing performance on increasing levels of updating and distracter resistance, we proceeded to quantify the subjective value participants assigned to them.

Figure 6A-B depicts the logistic regression curves of an example participant whose indifference points could be adequately sampled for both update (A) and ignore (B) conditions. As the task offer was

Table 4.

Bayesian analysis of effects in accuracy.

Effects	p(inclusion)	p(inclusion | data)	$BF_{Inclusion}$
Condition	.6	.994	104
Set size	.6	.958	15
Condition × Set size	.2	.262	1.4

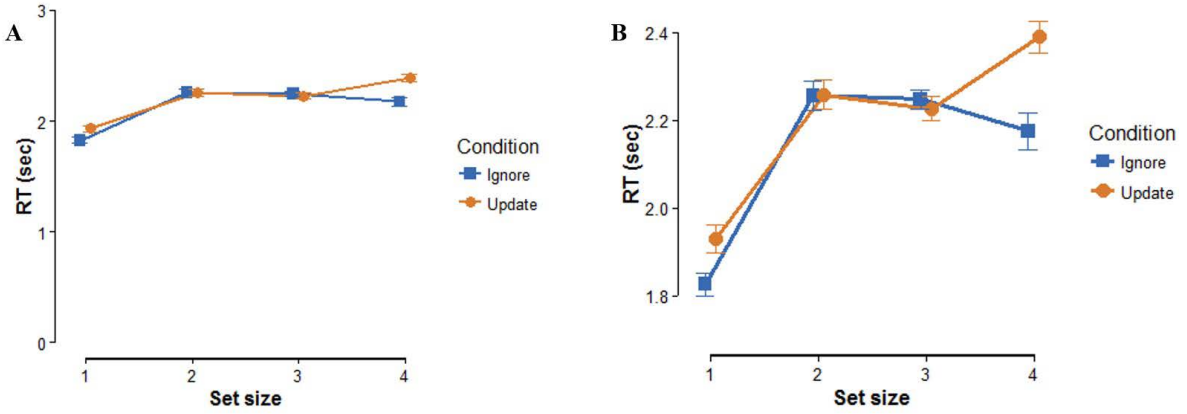


Fig. 5A. Reaction times for 24 participants in colour wheel working memory task. RTs are shown as a function of set size for distracter resistance (ignore) and flexible updating (update). **B.** Same results scaled. These results are best explained by the model including condition, set size and their interaction (BF_{10} = 1.667e+26). Error bars represent the standard error of the mean.

always €2, the smaller the indifference point from €2, the more the task value was discounted. For four participants, we could not estimate indifference points for at least one of the two conditions, so they were excluded. Two of them always chose the "No Redo" option, one of them always chose the task option and one of them always chose "No Redo" for update trials and task for ignore trials.

The indifference points across set size for 20 participants for distracter resistance and updating are displayed in Fig. 7A-B and descriptive statistics in Table 8A-B. The analysis (Table 9) showed that the model including set size and condition (BF_{10} = 85) is the model that is best supported by our data and that this model is ~1.4 times more likely than the one including only condition (BF_{10} = 60). The best model according to Bayesian model comparison is condition and set size (BF_{10} = 85), 1.4 times more likely than the runner-up model which is set size alone (BF_{10} = 59.86). Individual effects analyses (Table 10) provide very strong evidence for a set size effect (F(1.297, 19) = 4.145, p = .043, η^2 = 0.179, BF_{INC} = 48). The one sample t-test showed extreme support for both processes being discounted (for ignore, t-test [IP < 2]: t = -5.552, p < .001, Cohen's d = -1.235, BF_{-0} = 1917; for update, t-test [IP < 2]: t = -4.66, p < .001, Cohen's d = -1.042, BF_{-0} = 346). Regarding the effect of condition on the data frequentist statistics show significance, but the Bayesian analysis signifies that the data are inconclusive (F(1, 19) = 6.901, p = .017, η^2 = 0.266, BF_{10} = 1.081). Finally there is limited evidence against an interaction effect (F(3, 19) = 1.342, p = .270, η^2 = 0.066, BF_{INC} = 0.34). Overall, the results show that participants significantly

Table 5.

Descriptive statistics for RTs in colour wheel working memory task.

Condition	Set size	Mean	SD
Ignore	1	1.83	0.32
	2	2.26	0.28
	3	2.25	0.26
	4	2.17	0.36
Update	1	1.93	0.29
	2	2.26	0.29
	3	2.23	0.28
	4	2.39	0.27

Table 6.

Model comparison for RTs in colour wheel working memory task.

Models	p(M)	p(M | data)	BF_M	BF_{10}	% error
Null model (including subject)	.2	5.791e-27	2.316e-26	1.0	
Condition	.2	7.691e-27	3.076e-26	1.3	1.05
Set size	.2	.002	0.009	3.710e+23	0.75
Condition + Set size	.2	.033	0.135	5.620e+24	2.68
Condition + Set size + Condition × Set size	.2	.965	111	1.667e+26	5.74

Note. All models include subjects.

Table 7.

Bayesian analysis of effects for RTs.

Effects	p(inclusion)	p(inclusion | data)	$BF_{Inclusion}$
Condition	.6	.998	310
Set size	.6	1.0	∞
Condition × Set size	.2	.965	111

Table 8A.

Descriptive statistics for "task vs no redo" indifference points.

	Mean	*SD*
Ignore	1.44	0.46
Update	1.50	0.48

Table 8B.

Descriptive statistics for "task vs no redo" indifference points across set size.

	Mean	SD
1	1.50	0.50
2	1.52	0.41
3	1.42	0.47
4	1.31	0.55
1	1.53	0.46
2	1.54	0.45
3	1.54	0.52
4	1.41	0.58

discounted the working memory task option and that discounting increased with task difficulty, in line with our first hypothesis. We also show preliminary evidence that distracter resistance is discounted more than updating.

Choices between Distracter resistance and Updating. Having established that both our working memory processes were discounted by participants, we aimed to see if they were willing to discount rewards in order to perform updating over distracter resistance. As ignore offer was fixed at €2, a subjective value (indifference point) smaller than 2 indicates a preference for update, while a subjective value larger than 2 a preference for ignore. Figure 8 depicts regression curves of two example participants, one discounting ignore and the other discounting update. We excluded two participants from this analysis. One because we could not estimate any indifference points (always chose ignore) and another because they deviated more than 3 standard deviations from the mean.

For descriptive statistics see Table 11. In Figure 9 we report the average indifference points per set size for 22 participants. In accordance with our second hypothesis, the overall average subjective value of "ignore vs update" choices is less than 2

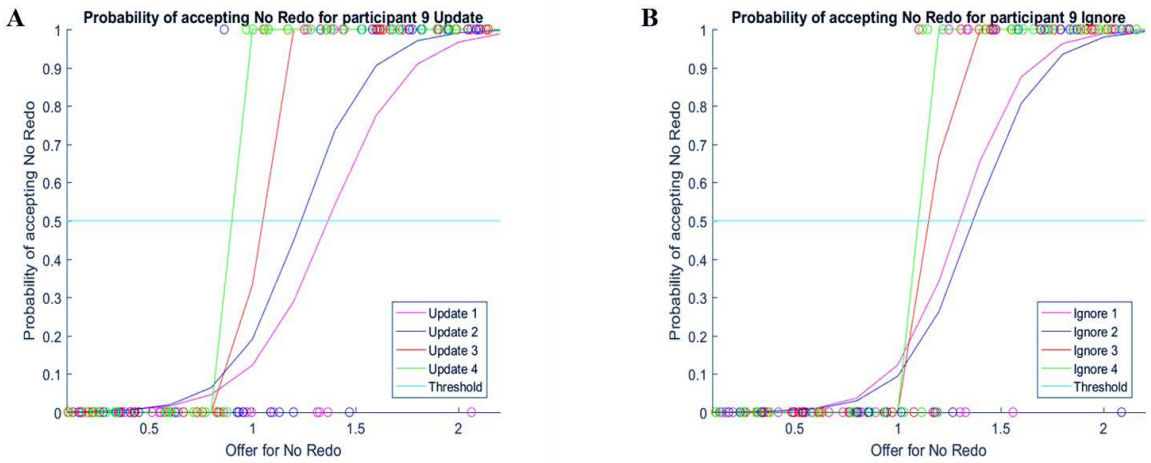


Fig. 6. Logistic regression curves for "task vs no redo" choices of one participant. We present the probability of accepting the "no redo" (no task) offer (y-axis) as a function of the amount of money offered for "No Redo" (x-axis). Task offer is always €2 for both conditions and all set sizes. The estimated indifference point is the offer for "no redo" where the possibility of choosing to do the task or the "no redo" option is equal (i.e., .5). **A. & B.** Example participant for update (A) and ignore (B) condition. Indifference points decrease for the higher demand levels.

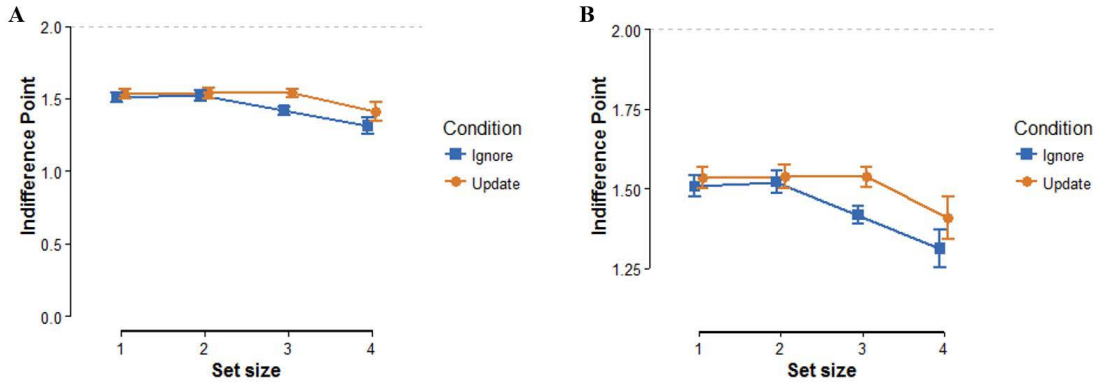


Fig. 7A. "Task vs no redo" Indifference Points as a function of set size across 20 participants. **B.** Same results scaled. As the task offer was fixed at €2, the more the indifference points deviate from 2 the more participants were willing to discount the task option. The results displayed are better explained by a model including set size and condition (BF_{10} = 85). Error bars represent the SEM.

Table 9.

Model comparison for "task vs no redo" indifference points.

Models	*p*(M)	*p*(M | data)	BF_M	BF_{10}	% error
Null model (including subject)	.2	.006	0.025	1.0	
Condition	.2	.007	0.030	1.2	3.92
Set size	.2	.375	2.403	60	0.49
Condition + Set size	.2	.532	4.554	85	1.19
Condition + Set size + Condition × Set size	.2	.079	0.342	13	2.20

Note. All models include subjects.

Table 10.

Analysis of effects for "task vs no redo" indifference points.

Effects	*p*(inclusion)	*p*(inclusion | data)	$BF_{Inclusion}$
Condition	.6	.618	1.08
Set size	.6	.986	48
Condition × Set size	.2	.079	0.34

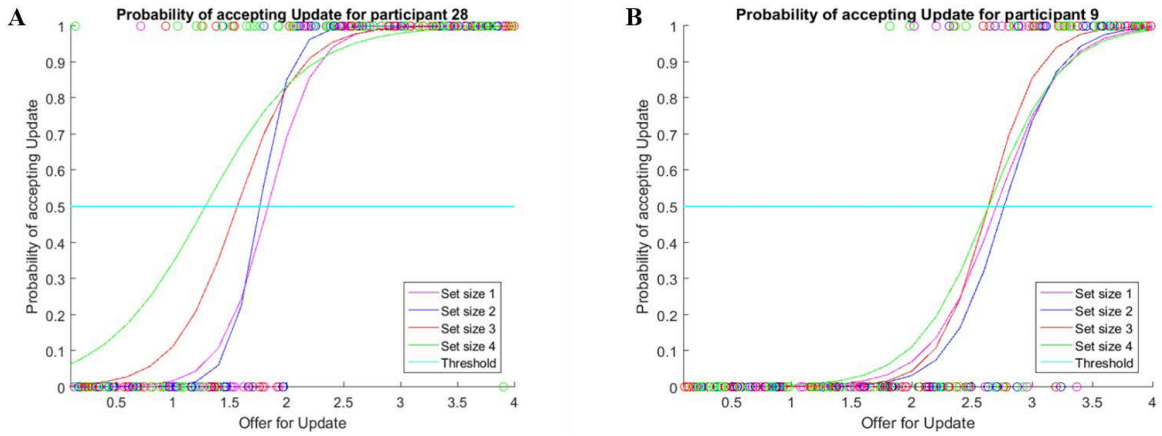


Fig. 8. Example logistic regression curves for "ignore vs update" indifference points. We see the probability of choosing the update offer as a function of the amount of money offered for update. Ignore offer is always €2 for both conditions and all set sizes. The indifference points are the offers for update that the probability of accepting it is .5, meaning that participants are equally likely to choose on offer or the other. **A.** Example participant willing to discount rewards in order to avoid ignore trials (preference for update). **B.** Example participant willing to discount rewards in order to avoid the harder levels of update trials (preference for ignore).

Table 11.

Descriptive statistics for "Ignore vs Update" indifference points (IP 1-4: indifference points for set size 1-4).

	IP	IP1	IP2	IP3	IP4
Mean	1.9	1.92	1.89	1.90	1.89
SD	0.26	0.24	0.28	0.25	0.32

(1.899), indicating a preference for flexible updating trials. This hypothesis is ~1.8 times more likely than the null and the classical t-test indicates a statistical significant effect (t-test [IP < 2]: t = -1.848, p = .039, Cohen's d = -0.394, BF_{-0} = 1.8). The output of the one-way repeated-measures ANOVA shows weak support for the data under the null hypothesis that subjective value is not influenced by set size (F(2.197, 63) = 0.407, p = .687, η^2 = 0.019, BF_{10} = 0.096). Our results provide anecdotal confirmation for our second hypothesis that participants are willing to discount rewards in order to repeat flexible updating trials over distracter resistance.

As the evidence for our second hypothesis was small we performed a sequential analysis to see how evidence accumulated as a function of sample size (Fig. 10). The analysis suggested that evidence for the alternative hypothesis was increasing with increasing sample size, so a larger sample could provide greater confidence in favour of our second hypothesis.

Questionnaires

During debriefing, out of 22 participants included in the analysis, 17 indicated that they preferred update trials and 5 reported a preference for ignore trials.

Two groups of preference. We considered the idea that there are two groups of participants with opposing preferences and by grouping them together we masked underlying effects. Indeed, we saw in this study, as well as in previous pilot studies (Online supplementary Fig. S2), that the majority

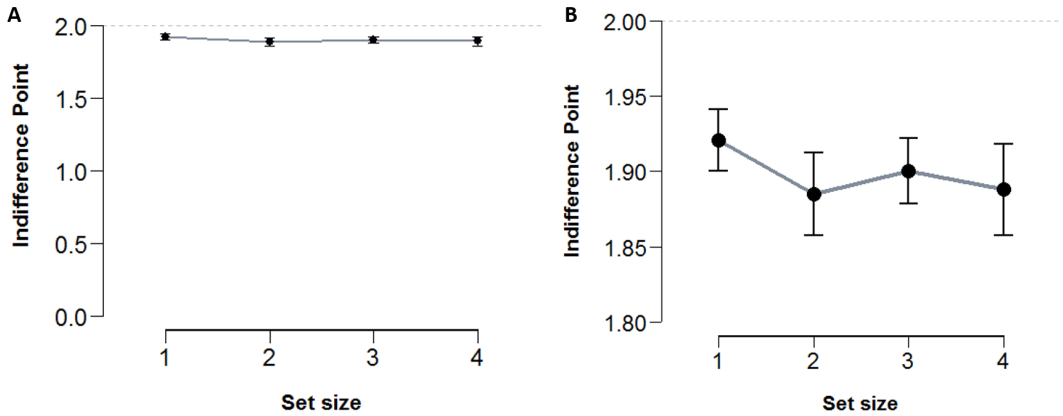


Fig. 9. Indifference points for "ignore vs update" choices across 22 participants. **A.** "ignore vs update" Indifference Points as a function of set size. **B.** Same results scaled As ignore was fixed at €2, indifference points smaller than 2 indicate a willingness to discount rewards to avoid repeating ignore compared to update trials. The statistical analysis showed anecdotal evidence for a preference for update as the hypothesis that overall indifference points are smaller than 2 is 1.8 times more likely than the null hypothesis (BF_{-0} = 1.803).

of participants preferred update, but a smaller percentage reported preference for ignore (4.4). We followed this idea further and divided participants in two groups based on their written preference for either condition and then analysed their choices again.

In Figure 11 we graphed the indifference points per set size for the two groups separately. The indifference points of the 17 participants who reported a preference for update are indeed clearly smaller than 2 (mean = 1.835, SD = 0.1578) and the evidence of the hypothesis that the mean is smaller than 2 is extreme (t-test [IP < 2]: t = -4.306, p < .001, Cohen's d = -1.044, BF_{10} = 128). On the other hand, IPs of the 5 subjects who indicated to prefer ignore tend to be higher than 2 (mean = 2.117, SD = 0.4), but the sample size is of course very small to produce reliable statistics (t-test [IP > 2]: t = 0.645, p = .277, Cohen's d = 0.288, BF_{10} = 0.663). With the exception of two participants, written preference and preference expressed by indifference points were aligned (one reported preference for ignore, while discounting ignore and the other a preference for update while discounting update).

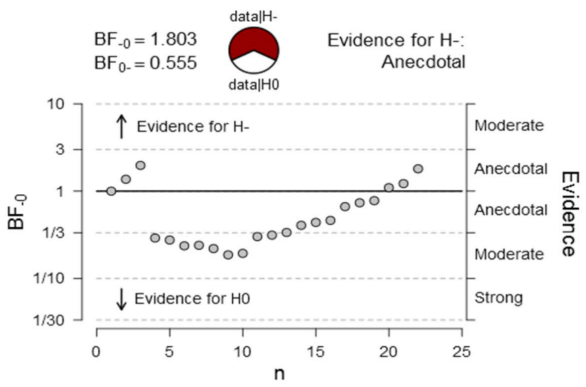


Fig. 10. Sequential analysis for the hypothesis that indifference points are smaller than 2 (preference for update) and the null. Evidence for the alternative hypothesis is growing with increasing sample size.

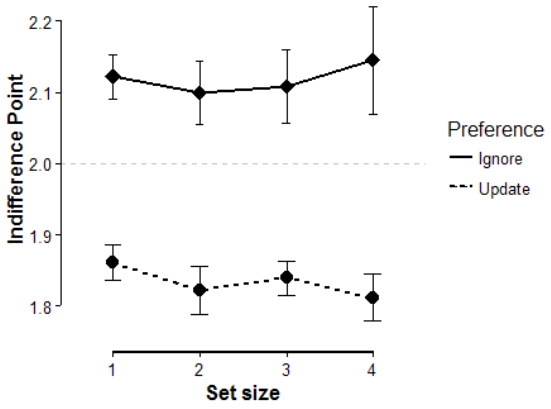


Fig. 11. "Ignore vs Update" indifference points of two groups of participants split based on written preference for either distracter resistance (ignore) or flexible updating (update) trials. The update group consists of 17 participants and ignore group consists of 5. Indifference points of 2 signify no discounting, thus, no preference between the two functions. Error bars represent SEM.

Discussion

In the current project, we set out to quantify the subjective valuation of distinct working memory processes to address our research questions. We asked whether working memory processes are perceived as costlier when demand increases and also whether two distinct key processes of working memory carry a differential cost.

The results show that engaging in updating and distracter resistance is costly and that costs increase as working memory demand grows. The results also provide some support for distracter resistance being perceived as more costly than updating.

Are working memory tasks costly?

Regarding our first research question, we asked whether distracter resistance and flexible updating are costly. Indeed, we have strong evidence to conclude that value discounting increased with demand, being highest for the higher set sizes. Participants discounted the value of distracter resistance overall by 28% and the value of flexible updating by 25%.

These findings extend current knowledge on the value of cognitive engagement. First of all, we show that people are aversive to cognitive demand, even willing to decline rewards in order to avoid demanding tasks. This is in line with earlier work showing that participants prefer to avoid higher levels of the N-back task (Kool et al., 2010). Our results further generalise these conclusions in a new working memory task while at the same time disentangling the subjective value of distracter resistance and flexible updating. Here we show that both functions are perceived as costly. In addition, our design strengthened the validity of reported findings by using a discounting procedure that accounts for the possibility of choices being probabilistic. Earlier studies on cognitive effort used staircase procedures that sample every choice option only once (Westbrook et al., 2013; Massar et al., 2016).

Unlike previous discounting studies we also gave participants the opportunity to choose the effortful option for less money. As expected, most participants declined this offer, but the subjective value of three participants was higher than 2 for at least one of the two working memory processes, indicating a preference for repeating the working memory task. This outcome may seem incongruent with our hypothesis, but it is not necessarily the case. One account for this may be that for those participants, cognitive engagement may be perceived as more valuable than both the monetary rewards we offered and the cost of engagement, in line with the concept of "learned industriousness". For example, socially reinforced rewards or an internal sense of accomplishment might lead to these choices. Another reasoning could be that these participants preferred to repeat the task than to be bored. Indeed, there have been studies suggesting that people would rather receive electrical shocks than do nothing at all (Wilson et al., 2014), suggesting that boredom carries a cost in itself. We aimed to minimise that prospect by offering them the option to surf the internet or use their phones while waiting. Either way, the majority of participants were inclined to discount rewards to avoid the working memory task and not the other way around.

We interpret this significant discounting of our task as evidence that people are averse to high working memory demand. An alternative explanation for the observed effects could be error-avoidance. To diminish such influences, we highlighted that accuracy during the redo would not define whether participants receive the monetary rewards or not. So, mistakes did not bear any external costs in our design, but we cannot exclude intrinsic costs. Moreover, simple error-avoidance seems like an unlikely explanation for the still significant discounting of the easiest set size, at which participants performed very well (less than 10° deviance from target colour) and comparable with performance without a working memory component (~3° more).

But what makes working memory tasks costly? The answer to this question remains an enigma and is the source of a lot of debate in the scientific community. One promising theory inspired by cost-benefit decision-making theories views the cost of cognitive engagement as an opportunity cost (Kurzban et al., 2013). As per this account, our working memory resources cannot be allocated to an infinite number of tasks simultaneously, which means that we perform any task at the expense of all other alternative tasks. So, while the value of these alternative options increases, the cost of focusing on the current task increases as well up to the point where performance fails or we even disengage completely. This model could potentially explain our results. In the "task vs no redo" version of the discounting choice paradigm, the "No Redo" option clearly carries a smaller opportunity cost compared to the "Task" option because participants can use their time as they please. In addition, our data show that willingness to do the task can be manipulated with incentives. Despite the above, with the

current design, we cannot make a case between the opportunity cost and other theories such as resource depletion. To assess that in the future, we could vary the opportunity of pleasurable alternative activities during tasks or free time. However, the lines of research are not necessarily mutually exclusive. For example, as proposed by Harvey (2013), dopamine is a "resource" depleted in Parkinson's patients drastically affecting performance, but it is also a key neurotransmitter in valuation.

Are some working memory functions perceived as costlier?

In accordance with our second hypothesis, we showed anecdotal evidence that distracter resistance has lower subjective value than flexible updating. Overall discounting, in this case, was around 5% and it showed no consistent variance with set size. The results of the frequentist analysis showed significant discounting of ignore, but Bayesian statistics support for this hypothesis is not strong. Further evidence that there may be a difference is that in "task vs no redo" choices the best model involved condition in addition to set size; the effects analysis was significant for frequentist statistics but inconclusive for Bayesian. This discrepancy between classical and Bayesian analyses is not surprising or uncommon. An empirical comparison of the two methods in 855 t-tests showed that that *p*-values between .01 and .05 often correspond to anecdotal evidence in favour of the alternative hypothesis in Bayesian terms (Wetzels et al., 2011).

We are replicating previous accounts that participants perform better at updating (Fallon & Cools, 2014; Fallon et al., 2015). We adapted previous versions, such that the two processes are contrasted without the confounding factor of a shorter time delay between relevant stimulus and response for update trials. However, this made update trials overall longer by 4 seconds, so it is even more interesting that participants preferred update despite a higher cost of time.

How can we interpret this preference? Again, the opportunity cost framework might be able to elucidate this observation, if we consider attending to the incoming stimuli in the ignore condition as a missed opportunity. Furthermore, it has been often stated that processing salient stimuli is an automatic, easy and fast bottom-up procedure while resisting this processing is goal-directed top-down and controlled (Corbetta & Shulman, 2002; Ernst, Daniele, & Frantz, 2011). Consequently, distracter resistance is more computationally costly, although participants only need to encode new stimuli once. When comparing ignore and update, error-avoidance might also contribute to the observed results. Most participants were more accurate in updating trials, but the average difference between the two conditions was only 5 degrees for the highest demand level and even lower for the lower levels. This is not a very striking difference, but participants could still be able to identify it and be affected by it. To account for error-avoidance effects, following studies could attempt to match performance between the two conditions or provide "fake" feedback to influence participants' beliefs about their performance.

Nonetheless, the preference for update was small and the support for this preference limited. The latter may very well be because our sample size was inadequate. Indeed, the sequential analysis indicated that a larger sample size would most likely solidify this conclusion. To draw confident conclusions, Bayesian analysis gives the possibility to continue sampling until either hypothesis reaches a Bayes factor of at least 10. Another factor for a small effect might be the time difference between the trials of the two conditions. Despite minuscule, it may have been picked up by time-sensitive participants and caused a research question-unrelated aversion to updating trials. It seems reasonable that discounting was smaller for "ignore vs update" choices than "task vs no redo" choices. As we have shown that both processes are perceived as costly, the value of no working memory task is understandably higher than the value of one over the other. The opportunity cost of no task is also much lower compared to a better preferred, but still experimentally-defined task. We also examined the idea that there are individual differences in preference for ignore or update processes that were masked when we averaged. Although our study was not a priori designed to sample for groups and we cannot make any such statements, in an exploratory analysis we split participants in two groups based on reported preferences on a questionnaire. This analysis indicated that reported preferences generally aligned with choices on the discounting task and that most participants preferred update. However, a minority preferred the ignore trials. If these individual differences in valuation do in fact exist, it would be interesting to investigate in future studies what the underlying reasons for this differential valuation are. Past work has shown that dopaminergic medication improved overall distracter resistance at the expense of flexible updating (Fallon et al., 2015), however we also know that effects of psychostimulants

vary greatly with individual baseline measures of dopamine (Cools & D'Esposito, 2011). How does a preference for ignore versus update relate to baseline measures of dopamine and psychostimulant effects on cognition? A role for dopamine in effort-based decision-making would be consistent with studies in physical effort, where it has been shown that in Parkinson's patients dopamine medication increases selection of high effort/high reward trials (Chong et al., 2015), while dopamine depletion decreases willingness to exert effort in humans and rodents (Salamone, Correa, Farrar, & Mingote, 2007; Floresco, Tse, & Ghods-Sharifi, 2008). There is recent suggestive evidence from rodent studies that dopaminergic medication also modifies valuation and choices of cognitive effort (Cocker, Hosking, Benoit, & Winstanley, 2012). These findings together raise the questions whether dopamine, a neurotransmitter implicated in motivation (Salamone & Correa, 2012), is involved in valuation of cognitive effort and invites future research. Another potentially relevant neurotransmitter is noradrenaline that seems to be involved in switching modes between task engagement (exploitation) and disengagement (exploration) (Jepma, Te Beek, Wagenmakers, Van Gerven, & Nieuwenhuis, 2010).

Value-based decision-making

Our findings are also consistent with a value-based decision-making process for cognitive resource allocation. We saw that participants overall showed aversion to both working memory processes, preferring "No Redo" option when the rewards were comparable. However, as the offer for "No Redo" was substantially decreasing, most participants were willing to shift their preference and actually chose to do the task. Likewise, most participants shifted their preference for ignore or update. This finding showcases the importance of motivation on task engagement, similarly to results by a recent study in sustained attention (Massar et al., 2016). Finally, the possibility for a role of error-avoidance in our results does not necessarily challenge value-based decision-making because fear of failure can be assessed as a cost in itself.

Limitations

One likely caveat of the study is that there were participants for whom we were not able to sample an indifference point. To avoid that in the future, we could increase the offer range. Another limitation is that we did not perform eye-tracking to exclude that participants closed their eyes in ignore trials. Nonetheless, in order to know that it was an ignore trial participants had to at least initially attend to the stimuli. Additionally, the performance results themselves suggest that participants were indeed at least to some extent distracted during ignore trials, evidenced by lower performance.

Future directions

Although preliminary, our results suggest that different working memory processes may carry different subjective costs. If confirmed, it could highlight the importance of choosing a working memory paradigm when studying cognitive effort. Additionally, identifying differences in valuation can also help us understand performance failure and variance, but also pave the way to. In this direction, future studies could sample for two groups, one valuing updating more and one valuing distracter resistance, and then positron emission tomography (PET) and/or pharmacological studies for dopamine and noradrenaline could help us gain some insight to the underlying mechanisms of this variance in preference.

Our results could also have potential implications for attention deficiency hyperactivity disorder (ADHD) research. For one, reduced performance may reflect differences in valuation. Moreover, we know that ADHD patients show deficiencies in resisting distraction and operate in a more stimulus-driven fashion (Swanson et al., 1998). On the other hand, flexible updating has been linked with creativity and innovation and there are indeed reports that ADHD patients show increased levels of creativity (Abraham, Windmann, Siefen, Daum, & Güntürkün, 2006). Future studies could assess whether ADHD patients discount distracter-resistance more than updating and if this valuation can be manipulated with incentives. If that is the case, novel educational strategies could be developed that aim to increase motivation and take their flexibility into account.

Conclusions

Concluding, this study provides new insights to the novel and growing fields of cognitive effort discounting and value-based decision-making. Specifically, we showed that with increasing demand on working memory processes the subjective valuation decreased, both for the process of distracter resistance and flexible updating. We

also provided evidence that distracter resistance is perceived as relatively costlier. These results remain to be further established and their underlying mechanisms investigated by future research.

References

Abraham, A., Windmann, S., Siefen, R., Daum, I., & Güntürkün, O. (2006). Creative thinking in adolescents with attention deficit hyperactivity disorder (ADHD). *Child Neuropsychology, 12*(2), 111-123.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual Review of Psychology, 66,* 83-113.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433-436.

Chong, T. T. J., Bonnelle, V., Manohar, S., Veromann, K. R., Muhammed, K., Tofaris, G. K., ... & Husain, M. (2015). Dopamine enhances willingness to exert effort for reward in Parkinson's disease. *Cortex, 69*, 40-46.

Cocker, P. J., Hosking, J. G., Benoit, J., & Winstanley, C. A. (2012). Sensitivity to cognitive effort mediates psychostimulant effects on a novel rodent cost/benefit decision-making task. *Neuropsychopharmacology, 37*(8), 1825-1837.

Cools, R., & D'Esposito, M. (2011). Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biological psychiatry, 69*(12), e113-e125.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience, 3*(3), 201-215.

Ernst, M., Daniele, T., & Frantz, K. (2011). New perspectives on adolescent motivated behavior: attention and conditioning. *Developmental cognitive neuroscience, 1*(4), 377-389.

Fallon, S. J., & Cools, R. (2014). Reward acts on the pFC to enhance distractor resistance of working memory representations. *Journal of cognitive neuroscience, 26*, 2812-2826.

Fallon, S. J., Van Der Schaff, M. E., Ten Huurne, N., & Cools, R. (2015). Methylphenidate improves cognitive stability at the expense of cognitive flexibility. Manuscript submitted for publication.

Floresco, S. B., Tse, M. T. L., & Ghods-Sharifi, S. (2008). Dopaminergic and glutamatergic regulation of effort- and delay-based decision making. *Neuropsychopharmacology, 33*(8), 1966-1979.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 489-494.

Harvey, N. (2013). Depletable resources: necessary, in need of fair treatment, and multi-functional. *Behavioral and Brain Sciences 36*(06), 689-690.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1485), 1601-1613.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS med, 2*(8), 0696-0701.

JASP Team. (2016). JASP (version 0.7.5.6) [Computer software].

Jepma, M., Te Beek, E. T., Wagenmakers, E. J., Van Gerven, J. M. A., & Nieuwenhuis, S. (2010). The role of the noradrenergic system in the exploration-exploitation trade-off: a psychopharmacological study. *Frontiers in human neuroscience,* 4, 170.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature neuroscience, 10*(12), 1625-1633.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*(4), 665-682.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36*(06), 661-679.

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

Massar, S. A. A., Lim, J., Sasmita, K., & Chee, M. W. L. (2016). Rewards boost sustained attention through higher effort: A value-based decision making approach. *Biological Psychology, 120*, 21-27.

Padmala, S., & Pessoa, L. (2011). Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of cognitive neuroscience, 23*(11), 34193432.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1446-1465.

Salamone, J. D., & Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. *Neuron, 76*(3), 470-485.

Salamone, J. D., Correa, M., Farrar, A., & Mingote, S. M. (2007). Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology, 191*(3), 461-482.

Swanson, J. M., Sergeant, J. A., Taylor, E., Sonuga-Barke, E. J. S., Jensen, P. S., & Cantwell, D. P. (1998). Attention-deficit hyperactivity disorder and hyperkinetic disorder. *The Lancet, 351*(9100), 429-433.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., & Morey, R. D. (2016). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review.*

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., & Morey, R. (2016). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review.*

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS*

One, 8(7), e68210.

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.

Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., ... & Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science, 345*(6192), 75-77.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233-235.

Listening in the Wrong Language: The Role of Language Dominance and Accent in Cross-language Speech Misperceptions

Mónica A. Wagner[1]
Supervisors: Kristin Lemhöfer[1], James M. McQueen[1]

[1]*Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands*

As has been well-documented in the literature, bilinguals possess a remarkable ability to switch between their languages while interacting with other bilinguals in mixed language contexts. Despite this, sometimes problems can still arise. This study addresses the rare phenomenon of speech misperceptions due to non-target language processing or "listening in the wrong language" (LWL). Our primary goal was to induce LWL states using an auditory sentence verification task with a twist: Participants were led to believe the experiment would be in one of their languages (the base language of the experiment) only to unexpectedly hear speech in their other language (the guest language) one-third of the way through the task. Failures to comprehend were measured by including a response option for when listeners did not understand an utterance. In addition, we investigated how the occurrence of misperceptions was affected by speaker accent (native vs. non-native) and whether the language was the listener's native (L1) or non-native (L2) language. The results revealed more comprehension failures for items in the guest language relative to the base language. Furthermore, the results suggest that language and accent may only initially play a role, with no significant effects observed after the guest language was presented for the first time. Moreover, listener familiarity with an accent may modulate its surprise effect. The results are discussed in the context of theories of bilingual speech comprehension.

Keywords: cross-language speech misperceptions, bilingual speech comprehension, non-native accent, language dominance

Corresponding author: Mónica Wagner; E-mail: M.Wagner@donders.ru.nl

"Selavy."

You have probably heard this phrase a thousand times. Out of context, it may be hard to locate its meaning, but if you were to say it out loud, you might recognise this string of letters as an English homophonic translation of the French phrase "c'est la vie" ("that's life").

Visually, cross-linguistic illusions like this can be induced by manipulating word boundaries and language-specific spelling rules, sending readers on a wild goose chase down the wrong language garden path. Similar phenomena can also occur during speech comprehension: Just imagine that, instead of reading the above word, someone said it to you using English pronunciation. You might not immediately recognize the utterance as French and try to process it as English.

The present study concerns itself with situations like the one described above where bilinguals experience difficulties or failures to comprehend due to non-target language processing or "listening in the wrong language" (LWL). Crucially, in order for a speech misperception to be considered LWL, the listener must be proficient in the target language and/or normally understand the utterance with ease. In a preliminary survey of 402 bilinguals[1] from all over the world, we found that nearly 60% of all respondents and over 80% of Dutch participants (N = 95) reported having experienced LWL at some point in their lives. Furthermore, of those familiar with LWL, most (80%) indicated that they experienced it rarely to occasionally. Therefore, LWL appears to be a rare but real phenomenon.

How exactly LWL states come about is a question researchers are yet to answer. Understanding how speech comprehension normally proceeds in bilinguals may help determine how this process goes awry in LWL states. One model of bilingual word recognition, the Bilingual Model of Lexical Access (BIMOLA; Grosjean, 1988, 1997; Lévy, 2015), was even initially developed to account for LWL. According to this model, phonemes and words are stored separately for each of the bilingual's languages (Lévy, 2015; Shook & Marian, 2013). The language that is accessed at any given moment depends on the global language mode the bilingual is in. Usually, [1]the bilingual will mainly process words in the "base language" or main language of an interaction due to a relatively greater activation of this language. However, when the other language, or "guest language", is also activated, the bilinguals can be said to be in a bilingual mode, and processing ensues in both languages in parallel, although independently (Grosjean, 1988; Lévy, 2015). Thus, following the BIMOLA, LWL could result from selective lexical access in the unintended or "non-target" language.

Many studies have demonstrated that non-target language activation in bilingual word recognition may actually be more common than suggested by the BIMOLA. For example, using a visual world paradigm with eye-tracking, Spivey and Marian (1999) showed that, after hearing instructions to pick up an item in English (e.g., *Pick up the marker*), Russian-English bilinguals fixated more on objects (e.g., a stamp) whose names in the non-target language (*marka*) were similar to the target object than they did on unrelated controls. In light of this and similar findings (Ju & Luce, 2004; Weber & Cutler, 2004), most models of bilingual speech comprehension advocate integrated lexicons with language nonselective lexical access (Li & Farkaš, 2002; Shook & Marian, 2013; Zhao & Li, 2007, 2010). Despite this, bilinguals usually manage to "zoom in" to the target language, as Elston-Güttler, Gunter, and Kotz (2005) call it, and eventually access the meanings of words in the right language. According to integrated lexicon models, this is accomplished in bilinguals, after initially nonselective access, by language-specific patterns of activation. These activation patterns can be explained by words in the same language being more strongly associated to each other as a result of repeated co-activation due to shared language-specific phonology (Li, 1998; Li & Farkaš, 2002; Shook & Marian, 2013). From this perspective, LWL could be explained by non-target language-specific patterns of activation, which would bias the system against the target language.

In some cases, non-target language lexical access may result in meaning, albeit not the one intended by the speaker. This is the case of, for example, near interlingual homophones ("false friends") such as *pet* which means *cap* in Dutch (although phonetically realized differently). Misperceptions of speech resulting in meanings different than the one originally intended are known as "mondegreens" (Hendriks, 2014). The term was coined by writer Sylvia Wright who, as a child, misheard a line from the Scottish song "The Bonny Earl O'Moray" as *They hae slain the Earl O'Moray, and Lady Mondegreen* instead of the original …*and laid him on the green* (Beck, Kardatzki, & Ethofer, 2014; Wright, 1954).

1 Some people prefer to reserve this term for simultaneous and/or balanced bilinguals to distinguish them from unbalanced and/or consecutive bilinguals, often called "second language learners." Here we use the term "bilingual" to refer to all speakers of two or more languages (technically "multilinguals"), clarifying with qualifiers when relevant.

Mondegreens can also occur across languages, due to near interlingual homophones, in which case they are known as Hobson Jobsons (Yule & Burnell, 1903) or Soramimi, when the objects of misperception are song lyrics. Otake (2007) analyzed 194 song lyrics misheard in English as Japanese, the listeners' native language (L1). He found that only 4% were due to purely segmental errors. The rest involved errors extending beyond word boundaries, at the phrasal level. The study of Soramimi can shed light on another way LWL can occur, namely, segmentation.

In contrast to written words, spoken words are not usually separated by pauses (Cole & Jakimik, 1980; McQueen, 1998). Instead, during speech perception, listeners are faced with the task of extracting discrete words from a continuous speech signal (McQueen & Cutler, 2010). In order to accomplish this, listeners make use of different cues that help them detect word boundaries, such as acoustical features (e.g., Gow & Gordon, 1995; Quené, 1992), metrical structure (e.g., Cutler, Dahan, & Donselaar, 1997), and phonotactic information (e.g., McQueen, 1998). The specific cues listeners use, vary depending on the language at hand (Cutler, 2012; Cutler, Mehler, Norris, & Segui, 1986; Tyler & Cutler, 2009). Studies on Soramimi suggest that LWL could be explained by the listener attending to segmentation cues from the wrong language, essentially segmenting the signal at the wrong points, a phenomenon called "juncture misperception" (Kentner, 2015; Otake, 2007).

Most studies on cross-language speech misperceptions have made use of interlingual homophones and Soramimi to demonstrate how bilinguals can end up perceiving words from the non-target language. However, these can be seen as a special case of LWL, where the end product is a (non-target) meaning. More often, though, misperceptions do not result in meaning but rather solely a failure to understand (Bond, 2008). So far, we have discussed two ways in which this might occur: non-target lexical access and segmentation, both intrinsic to the process of speech comprehension. But what extrinsic factors can cause the train to derail, if you will, and proceed in the non-target language?

One likely culprit is a factor briefly touched upon before: The context that the bilinguals find themselves in can lead them to preferentially expect one or the other language. One such contextual factor is the linguistic context, as revealed by the base-language effect (Macnamara & Kushnir, 1971), "a momentary dominance of base-language units (phonemes, syllables, words) at code-switched boundaries…" (Grosjean & Miller, 1994, p. 201). A similar effect was found in an event-related potential (ERP) study on visual comprehension in the L2 (Elston-Güttler et al., 2005). German-English bilinguals performed a lexical decision task on targets (i.e., *poison*) preceded by sentences such as "The woman gave her friend an expensive *gift*" (*poison* in German; control: *item*). Prior to the experiment, participants watched a short film with subtitles in German or English. The results revealed a semantic priming effect behaviourally and modulations in the N200 and N400 ERP components but only following the German version of the film. Moreover, these effects were temporary, disappearing after the first block.

More evidence for a potential role of context comes from work on bilingual speech production. A series of studies suggests that visual cues from the context, such as the speaker's face, whether familiar or unfamiliar, as well as cultural symbols, might bias processing towards the congruent language (Woumans et al., 2015; for a review, see Hartsuiker, 2015). Together these findings suggest that cues from the context might play a role in LWL by augmenting expectations for the non-target language.

Another factor that may contribute to LWL occurrences derives from the speech signal itself: phonetic realization. In particular, activation of the non-target language may increase if the speaker has a non-native accent, especially if the listener speaks the language associated with that accent. Often non-native speakers will even use sounds that only exist in their native language, which could increase expectations for the non-target language. In a series of studies, Grosjean and collaborators (Bürki-Cohen, Grosjean, & Miller, 1989; Grosjean, 1988; Soares & Grosjean, 1984) studied these factors in the recognition of words in the guest language. In a gating study in which participants heard words in increasingly longer fragments, Grosjean (1988) measured how long it took participants to "isolate" (i.e., accurately and consistently identify) guest words in a carrier sentence in the base language. He analyzed the role of three factors: language-specific phonotactics, language-specific phonetics, and the existence of a homophone in the base language. In addition to guest words with guest language-specific phonotactics and guest words that were not homophones, he found an advantage in isolation for guest words pronounced with guest-language phonetics relative to guest words pronounced as in the base language (Li, 1996).

Despite the fact that sentences were used in Grosjean (1988)'s study, it was always the same neutral lead-in phrase: *Il faudrait qu'on… (We should…).* Consistent with the context effects

described above, there is evidence that sentential context can help reduce the amount of non-target language interference to aid selective lexical access during bilingual speech comprehension (for a review, see Fitzpatrick & Indefrey, 2014). This means that phonetic realization may be less important for comprehension in real life than in single-word experiments (Li, 1996). Consistent with this idea, Li (1996), in a replication of Grosjean's (1988) study, found that less of the word was needed for semantically constraining sentences (but see Bürki-Cohen et al., 1989). Lagrou, Hartsuiker, and Duyck (2012) observed similar results using an auditory lexical decision task on the last word in sentences with varying semantic constraint. Critically, the last words were interlingual homophones, which have been shown to cause a delay in processing. They found that the semantic constraint of the sentences, as well as native accents, reduced the effect of interlingual homophones, although not fully eliminating it (see also Chambers & Cooke, 2009; Fitzpatrick & Indefrey, 2010).

Finally, LWL states might also be modulated by a factor pertaining to the listener, that is: whether the target language is the listener's L1 or second language (L2). This is consistent with theories proposing a reduced baseline activation of the L2 (e.g., Pallier, Colomé, & Sebastián-Gallés, 2001). Support for a role of proficiency was provided by a replication of Grosjean (1988)'s study with interlingual homophones in the participant's L1 and L2 (Schulpen, Dijkstra, Schriefers, & Hasper, 2003). There was a disadvantage for the L2, with words in this language being identified less often and, when identified, requiring longer gates. This view is also supported by studies on adverse listening conditions (Bond, 1996), where the difference between the L1 and L2 is found to be exacerbated by adverse conditions (such as noise), causing greater problems for the L2 than the L1.

Furthermore, work on Soramimi have found that the strength of their perception in the L1 correlates positively with verbal fluency in the L1. Moreover, their perception in the L2 was found to not correlate negatively with proficiency in this language. Together, these findings seem to suggest that their occurrence is related to creative solutions to ambiguous acoustic signals, rather than limited linguistic competence (Beck et al., 2014; Beck Lidén et al., 2016).

Studies on switching during language production may also suggest a greater incidence of LWL when the target language is the L1 compared to the L2. In these studies, a common finding is that switching into the L1 is harder than switching into the L2. As the reasoning goes, language-selective access during speech production in bilinguals is accomplished via inhibition. During switching, this inhibition must quickly be lifted and replaced on the non-target language. As the L2's baseline activation is less than that of the L1, speaking in the L2 calls for greater inhibition of the L1 than of the L2 during L1 production. Overcoming this relatively greater inhibition leads to longer reaction times (RTs) for switches into the L1, a phenomenon now known as an "asymmetric switch cost." Following this logic, LWLs should be rarer when the target language is the L1 given the greater amount of inhibition required to keep the L1 at bay (Meuter & Allport, 1999).

In contrast to both of these views on the role of proficiency, yet another possibility is that LWL occurs as often in the bilingual's L1 and L2. An eye-tracking study was conducted in which the effect of semantic constraint of the preceding sentence on L2 auditory sentence processing did not vary with L2 proficiency. This suggests that context effects may play a bigger role than proficiency (Chambers & Cooke, 2009). Similarly, in an ERP study with intra-sentential switching, Fitzpatrick and Indefrey (2014) found no difference between the L1 and L2 in terms of switch costs.

To the best of our knowledge, no studies to date have addressed the occurrence of LWL using a naturalistic experiment. In fact, most studies on bilingual auditory comprehension have focussed on the processing of isolated words, for example in gating or lexical decision tasks (Elmer, Meyer, & Jancke, 2010). Moreover, studies that have looked at sentence-level comprehension have primarily used intra-sentential switching (e.g., Fitzpatrick & Indefrey, 2014). However, these types of tasks are limited in the extent to which they can inform us about how speech comprehension occurs in real life. In addition, many studies have made use of words with form overlap (e.g., cognates or interlingual homophones). These words may be unique, with some studies suggesting that sentential context may better restrict cross-language activation in words without form overlap (Hartsuiker, 2015). Finally, most studies have aimed to evaluate processes resulting in successful comprehension, while here, the interest resides in those cases when comprehension breaks down.

Present study

The present study aimed to induce LWL states in Dutch-English bilinguals. To this end, measures were taken to bias processing towards the

nontarget language: No mention was made of the guest language, and the experiment began with a monolingual block entirely in the base language (see Cheng & Howard, 2008 for a similar set-up in visual comprehension). Comprehension was assessed using an auditory sentence verification task with a twist: In order to gauge misperceptions accurately, in addition to the traditional "true" and "false" choices, participants were provided with a third option to indicate utterances they failed to understand. If our manipulation was successful, we would expect more comprehension failures for guest language items than base language items. Furthermore, even in cases where guest language items managed to be perceived accurately, we expected processing to take longer to be resolved (i.e., slower RTs).

The study also aimed to evaluate whether the incidence of LWL differed if the guest language was the bilingual's L1 or L2. Given the dominance of the L1 in unbalanced bilinguals, the L2 may have a reduced baseline activation and thus an L2 guest language item may be more unexpected than one in the L1. On the other hand, if bilingual speech comprehension in the target language is accomplished via inhibition of the non-target language, as has been suggested for bilingual speech production, guest language processing costs should be greater for the L1 than for the L2.

As explained above, in addition to the actual language being spoken, the phonetics of the utterance can influence the language-selectivity of lexical access during speech comprehension. Another goal of the present study, thus, was to evaluate the effect of speaker accent on the incidence of LWL. This was implemented by having listeners hear utterances produced by native and non-native speakers. Only native Dutch and English speakers were used so the non-native speaker would always be a native speaker of the other language of the experiment. We predicted that guest language items produced by a non-native speaker would be misperceived more often and processed more slowly than those spoken by a native speaker, as the accent would increase the expectation for the non-target language. Similarly, native pronunciation would help disambiguate the language being spoken, in the end facilitating comprehension.

In summary, we predicted greater processing costs (in the form of a higher incidence of misunderstandings and slower RTs) for guest language items than base language items, both in the monolingual and bilingual blocks. Moreover, we suspected that this guest language effect might be different for the L1 and L2, although we were not really sure about the direction of the difference. Finally, we expected non-native accent to exacerbate guest language misperceptions and processing costs by increasing expectation for the non-target language.

Methods

Participants

Forty-nine native Dutch speakers (age: M = 23.5, SD = 3.4, range = 18-33; 15 male) participated in the auditory sentence verification study. Participants provided written informed consent before the start of the experiment and afterwards received a €10 voucher for their collaboration.

Design

The present study was different from most studies on speech comprehension in that it aimed to study failures to comprehend and, what's more, a type of failure that occurs only on rare occasions outside the laboratory. Such an infrequent phenomenon called for a unique approach that would maximise the probability of observing these misperceptions during the experiment. To this end, a design was conceived to induce non-target language expectations, essentially tricking the participant. Two important manipulations were introduced to the experimental design. First, the experimental task began with a monolingual block in the base language to establish the expectation for that language and set participants in that language mode. Furthermore, once the guest language was introduced, in an attempt to maintain expectation biased towards the base language, the frequency of guest language items and, thus, code-switches was kept low, specifically 20% of the items.

In terms of analysis, the effect of the guest language could be observed by comparing performance on items spoken in the guest language with those spoken in the base language. This could be accomplished via comparison of base language items in the initial, monolingual block. However, base language items in the monolingual block and guest language items varied in several aspects that could complicate interpretation of the results. First of all, as called for by the design, these base language items always preceded guest language items as they occurred in the first block of the experiment. This meant that any potential effects due to the order of presentation could not be controlled for. In addition to the common concern for effects of fatigue or

learning, this was particularly problematic for the present study where participants had to adapt to speakers' voices and accents.

An additional point of contrast was that guest language items, by definition, occurred in a bilingual context where both target languages were activated and the participant was required to switch from one language to the other. Therefore, any processing costs observed for guest language items could also be explained by interference from the increased activation of both of the bilinguals' languages or the fact that language mixing was more effortful. In task-switching studies, this is usually resolved by the introduction of nonswitch trials in the switch blocks, allowing for two comparisons: (1) switch trials (in switch blocks) - nonswitch trials in switch blocks and (2) nonswitch trials in switch blocks - nonswitch trials in nonswitch blocks (Hughes, Linck, Bowles, Koeth, & Bunting, 2014; Koch, Prinz, & Allport, 2005; Weissberger, Wierenga, Bondi, & Gollan, 2012). While the former measures the well known switching cost, the latter provides a measure of the cost of task mixing in general. This method was adopted in the present study, allowing for three critical conditions: (1) base language items when the participant was still in a monolingual context, (2) base language items in a bilingual context, and (3) guest language items (necessarily in a bilingual context). Most studies on task- or language-switching focus on the difference between (1) and (2): the effect of context — monolingual vs. bilingual — and/or (2) and (3): the effect of language status — base vs. guest language. However, those studies also usually conduct by-participant analyses. Given the nature of the present design, comparing conditions within participants would not have been very informative as condition differences were confounded with a change in language. Therefore, by-item analyses were preferred. Since we were interested in misperceptions, no sentence was presented twice to avoid priming effects. Thus, within-item analyses were between-participant and, to increase the power of these analyses, condition was kept as a three-level factor, instead of conducting separate analyses for the effect of context and language status.

Materials

For a comprehensive overview of the materials and pilot studies on the basis of which they were selected, see the complete version of the Materials in the online Supplementary Material.

Auditory sentence verification task (aSVT). The critical stimuli consisted of 64 Dutch and 64 English sentences, selected from an original set of 351 Dutch and 333 English sentences on the basis of two pilot studies. For each language, half of the sentences were true statements and half were false. Sentences were kept short — consisting of only three words — to ensure that participants would not be able to guess the meaning of the statement from the end of the sentence but rather had to understand all of the words. Critical sentences belonged to one of four syntactic structures: (1) noun + verbto be + noun (be + N), (2) noun + verbto be + adjective (be + Adj), (3) noun +verbto have + noun (have), or (4) noun + verbcan + verb (can; as in Collins & Quillian, 1969; for examples, see Table 1). Words whose translations were phonetically very similar (i.e., cognates) and interlingual homophones (i.e., false friends) were avoided by calculating normalized Levenshtein distances between the phonetic transcriptions (DISC; Baayen, Piepenbrock, & van Rijn, 1993) of the first and last words (hereafter "content words") words and their translations (Gooskens & Heeringa, 2004; Schepens, Dijkstra, & Grootjen, 2012). None of the critical content words exceeded .5 phonetic similarity (English sentences: M = .14, SD = .14, range = 0 - .43; Dutch sentences: M = .11, SD = .12, range = 0 - .40, $t(254) = 1.776$, $p = .077$).

Since English was not the participants' native language, words thought to be familiar to participants were chosen for the English sentences, resulting in a higher frequency for these, as can be seen in lemma frequency (per million; CELEX: $t(254) = 2.421$, $p = .022$; SUBTLEX: $t(254) = 1.824$, $p = .069$; see Table 2 for averages; Baayen et al., 1993; Brysbaert & New, 2009; Keuleers, Brysbaert, & New, 2010).[2] Nevertheless, given that "the bilingual is not two monolinguals in one person" (Grosjean, 1989), corpus-based frequencies probably do not very accurately reflect subjective frequencies for bilinguals (e.g., Connine, 2004; Duyck, Vanderelst, Desmet, & Hartsuiker, 2008).

The critical items were pretested in a series of off-line sentence verification tasks with native Dutch speakers. Participants read the sentences and were asked to judge, for each, whether they thought the statement was true or false. An additional option was included in the English version for participants to indicate any words they did not know. Each sentence was evaluated by at least 10 raters and the

2 One English word (peels) did not appear in noun form in CELEX, so its frequency value was computed as 0.

Table 1.

Examples of critical sentences per structure, language, and veracity.

	English		**Dutch**	
Structure	True	False	True	False
be + N	Cars are vehicles	Uncles are women	Tafels zijn meubels *(Tables are furniture)*	Schedels zijn spieren *(Skulls are muscles)*
be + Adj	Sugar is sweet	Deserts are wet	Bergen zijn hoog *(Mountains are tall)*	Schuurpapier is glad *(Sandpaper is smooth)*
have	Rabbits have fur	Shrimp have pearls	Uilen hebben ogen *(Owls have eyes)*	Kwallen hebben botten *(Jellyfish have bones)*
can	Airplanes can move	Fish can walk	Nagels kunnen groeien *(Nails can grow)*	Hanen kunnen brullen *(Roosters can roar)*

veracity of each of the final 128 critical sentences was confirmed by at least 9 raters. A complete list of critical stimuli can be found in Appendix A of the online Supplementary Material. In addition to these critical sentences, 144 sentences (72 true; length: 3 words) and their translations were used as fillers and ten (5 true; 3-5 words) as practice items (Appendix B of the online Supplementary Material).

The number of total critical items and fillers was determined by the number of guest language items in the bilingual blocks: 20%, With a cell size of 32 (16 true, 4 per speaker), this meant a total number of 160 items, of which 64 were critical items (32 guest language and 32 base language items) and 96 fillers. Furthermore, the initial monolingual block consisted of 80 base language items (32 critical and 48 fillers).

The sentences were spoken by two native Dutch and two native English speakers selected on the basis of a series of pilot studies (for full details, see the online Supplementary Material).

Several candidate speakers were recorded reading sentences like those used in the aSVT. An online rating study was conducted with native Dutch speakers to obtain measures of accentedness (the perceived strength of non-native accent of the utterance, on a scale from 1 (*no foreign accent*) to 9 (*very strong foreign accent*) (Munro, 2008), comprehensibility (the perceived ease or difficulty in understanding the utterance, on a scale from 1 (*very easy to understand*) to 9 (*very difficult to understand*), and intelligibility (the degree to which an utterance is actually understood

Table 2.

Frequency, sentence length, utterance length, and speech rate of critical sentences.

	Variables							
	Language				**Accent**			
	English		**Dutch**		**Native**		**Non-native**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Word frequency								
CELEX	56.58	136.53	24.60	60.91	37.58	94.41	43.60	118.04
SUBTLEX	95.48	148.33	63.60	130.72	77.02	143.15	82.06	138.19
Sentence length (syllables)	3.89	0.86	5.72	1.05	4.75	1.31	4.86	1.34
Utterance duration (ms)	1137.75	148.17	1317.81	199.27	1207.06	136.38	1248.50	242.19
Speech rate (syllables/s)	3.44	0.73	4.37	0.73	3.93	0.96	3.88	0.76

Note. SUBTLEX-US used for English

by the listener, based on transcription accuracy) in both English and Dutch (Munro, 2008; Saito, Trofimovich, Isaacs, & Webb, 2015). If the raters indicated that the speaker had a native accent (foreign accent = 1), they were then asked to guess the region where the speaker was from/regional accent they had (e.g., a province in the Netherlands or Belgium for Dutch and British/American dialect for English). If, on the other hand, they responded that the speaker had a non-native accent (foreign accent > 1), they were asked to guess what the speaker's native language was. Four speakers (one male and one female per language) were selected based on the following criteria: perceived as (1) native in their mother tongue, (2) free of a strong regional accent, and (3) moderate to strongly accented in their non-native language (English or Dutch), but (4) still understandable. The results of the rating study are shown for the final speakers in Tables 3 and 4.

The final four speakers were recorded reading the final sentences for the aSVT in a sound-attenuated booth using a Sennheiser microphone. Audio files were recorded and saved in Audacity at 44 kHz. Speech was monitored online by the first author and, after reading the entire list, speakers were asked to repeat sentences pronounced with disfluencies or gross pronunciation errors that could hinder understanding. In addition to the four speakers, two different female speakers (one native Dutch speaker and one native English speaker [the experimenter]) recorded the practice items.

Tokens were manually extracted from the audio files by auditory and visual inspection of the waveform and spectrogram in Audacity, removing silence before and after the utterances. The best (i.e., most comprehensible) of the exemplars was chosen. Sentences were quasi-randomly assigned to speakers in such a way as to evenly distribute true and false sentences across speakers.

Audio stimuli were equated in amplitude using the normalize function in Audacity. Overall, utterances averaged 1228 ms in length (*SD* = 196.87). An analysis of variance (ANOVA) with language and speaker L1 confirmed a main effect of language, $F(1, 124) = 40.8889$, $p < .001$, and speaker L1, $F(1, 124) = 26.928$, $p < .001$ on utterance duration. No significant interaction between language and speaker L1 was revealed, $F(1, 124) = 2.165$, $p = .144$.

An additional measure of speech rate was calculated by dividing the number of syllables in a sentence by the duration of its utterance in seconds. An ANOVA with language and speaker L1 showed a main effect of language on speech rate, $F(1, 124) = 54.651$, $p = .000$, and of speaker L1, $F(1, 124) = 9.722$, $p = .002$, but no significant interaction between language and accent, $F(1, 124) = .151$, $p = .699$.

These analyses revealed two things: (1) that Dutch was spoken faster, and (2) that native Dutch speakers spoke faster than native English speakers. However, these differences are not too problematic for the present design because, as explained above (Design) analyses were within-item and, thus, within-speaker, so any potential difference in guest language effect for native and non-native speakers could not be explained by a difference in audio duration or speech rate.

Sentences spoken by native and non-native speakers did not vary in frequency (CELEX: $t(254) = -.451$, $p = .653$; SUBTLEX: $t(254) = -.287$, $p = .774$) nor did sentences spoken by native English and native Dutch speakers (CELEX: $t(254) = .004$, $p = .997$ SUBTLEX: $t(254) = -1.005$, $p = .316$).

Table 3.

Results of speaker ratings for accentedness, comprehensibility, and intelligibility.

	Selected speakers															
	English male				**English female**				**Dutch male**				**Dutch female**			
	English		**Dutch**		**English**		**Dutch**		**English**		**Dutch**		**English**		**Dutch**	
Rating	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
ACC	1.37	0.56	6.33	2.04	1.60	1.10	5.40	2.43	4.63	2.54	1.57	1.25	6.37	2.43	1.03	0.18
COMP	1.80	1.45	5.47	2.58	2.23	2.03	4.03	2.57	2.63	1.75	1.70	1.74	3.43	2.42	1.07	0.25
INT	.97	.10	.84	.28	.94	.15	.97	.10	.94	.20	.99	.06	.99	.06	1	0

ACC = accentednesss (1 = native, 7 = very strong foreign accent)

COMP = comprehensibility (1 = very easy to understand, 7 = very difficult to understand)

INT = intelligibility (proportion of 3 words correct)

Table 4.

Characteristics of final four speakers.

	English male	English female	Dutch male	Dutch female
Age	32	28	22	36
Originally from	California, U.S.A.	California, U.S.A.	Gelderland, The Netherlands	Noord Brabant, The Netherlands
Years lived in Gelderland	2	2	22	17
Speaking English				
Perceived as native (%)	67	67	20	3
Modal perceived regional dialect	U.S.A. (60%)	U.S.A. (50%)	U.S.A. (20%)	U.K. (3%)
Perceived as Dutch speaker (%)	0	7	43	53
Speaking Dutch				
Perceived as native (%)	0	3	60	87
Modal perceived regional dialect	-	Gelderland (3%)	Gelderland (20%)	Noord-Holland (37%)
Perceived as English speaker (%)	13	40	0	3

Note. Percentages are out of a total of 30 tokens rated.

Language background questionnaire (LBQ). Participants completed a LBQ (in Dutch) with questions about their native language(s) and experience with non-native languages. For each non-native language named, listed in order of proficiency, the following measures were obtained: age of acquisition, frequency of use (1-Never, 2-Rarely, 3-Occasionally, 4-Sometimes, 5-Frequently, 6-Very frequently, 7-Always), and self-rated proficiency for speaking, listening, writing, and reading (1-Very poor, 2-Poor, 3-Fair, 4-Functional, 5-Good, 6-Very good, 7-Fluent).

In addition to the information about languages spoken, participants answered a few questions about their previous exposure to the accents presented during the experiment: (1) were they familiar with the accents of the native Dutch and native English speakers from the experiment, (2) what dialect of English they were most familiar with (options: American, Canadian, British, Scottish, Irish, Welsh, Australian, New Zealand, and South African), (3) how often they heard English-accented Dutch and Dutch-accented English (never, less than once a week, once a week, several times a week), and (4), for each, from how many speakers (0-1, 2-5, 6-10, more than 10; following Witteman et al., 2013). In addition, participants were asked where they were from and how long they had lived in the province of Gelderland.

Finally, a few questions were added to inquire about the incidence of LWL, that is, situations where they did not understand what someone said to them, despite speaking the language, because they were expecting the person to speak another language: (1) had they ever experienced LWL, (2) how often (1-*Never*, 2-*Rarely* [less than once a month], 3-*Occasionally* [once a month], 4-*Sometimes* [more than once a month, less than once a week], 5-*Frequently* [once a week, less than once a day], 6-*Very frequently* [once a day], 7-*Always* [several times a day]), and (3) did they experience LWL during the experiment.

LexTALE. In addition to the self-ratings of English ability provided in the LBQ, the English version of LexTALE (Lemhöfer & Broersma, 2012) was administered as an objective measure of proficiency. This test is a brief lexical decision task which measures English vocabulary knowledge. The test consists of 60 items and scores are calculated by weighing both hit and false alarm rates.

Procedure

Experimental list construction. Considering the length of the aSVT (240 items total), it was considered necessary to split the items into blocks. A first block of 80 items (32 critical) coincided with the monolingual context. To increase comparability, the

bilingual context items were divided into two blocks of 80 items, as well, with a sub-set of base and guest language critical items evenly distributed into each block. To this end, critical stimuli were separated into two sets per language and rotated through the three conditions, yielding four experimental lists. The sub-sets were also rotated through experimental block, so each item appeared in each block. Furthermore, care was taken while dividing the stimuli into subsets to ensure each speaker was equally represented in each block and in each language, as well as with an equal number of true and false statements and a similar number of stimuli per verb structure. As mentioned before, controlling the frequency of each manipulation was given such importance since the effect of interest hinged on expectations and, thus, probabilities. Filler sentences were fixed to their blocks and also equally divided in terms of speaker, veracity, and sentence structure.

Speaker's identity changed on every trial, as did the sex of the speaker to avoid having to control for congruence of speaker identity and sex between trials. Similarly, critical items always followed true statements to prevent differential effects from previous statements, as false statements tend to take longer to verify than true ones (Cox, 2005; Gough, 1966). Furthermore, no more than four trials of each condition (veracity, accentedness, speaker L1) appeared consecutively. An attempt was made to make sure critical sentences did not follow sentences with the same verb (be, can, or have), when not possible, care was taken that the sentences did not contain the same conjugation form of the verb (e.g., "is"). Sentences that could be semantically associated were also kept apart. In bilingual context blocks, critical items never immediately followed a guest language item, with three to five intervening trials between guest language items. The order of the variables speaker and veracity were kept constant across the four lists, except for two items so that the first guest language could occur in both native and non-native accent conditions. Base language items in the bilingual blocks always occurred after the first appearance of the guest language item to ensure participants were in bilingual mode.

Testing. Given that the critical manipulation of the study involved an unexpected guest language, special attention was paid to the information participants received about the experiment and several measures were taken to induce a monolingual mode in an effort to maximize the expectation for the base language. Participants were recruited via the Radboud Research Participation System and were prescreened with the following information, provided in the system's general questionnaire: (1) not to suffer from any hearing problems, (2) to be between the ages of 18 and 35, (3) to have Dutch as their native language, (4) to speak Dutch, (5) not to have been raised multilingually, and (6) to speak English. However, only the first requirement was made visible to participants and no mention was made of English or the fact that the study was about language. Recruitment for both base language groups was conducted in Dutch with the premise that separate recruitment in English for the English base language group could result in a differential preselection of the participants (e.g., based on their attitudes towards and confidence in English). Thus, in order to keep English experience constant, no mention of English was made.

On the day of the experiment, participants were assigned to an experimental list and received all information about the study in the corresponding base language, including the informed consent and prescreening forms. The only exception were the instructions received orally from the experimenter, which were always given in English. However, participants of the Dutch baseline group were told that this was a limitation of the experimenter and if they asked (although very few did), participants were led to believe that the experiment would be conducted in the base language. Furthermore, the aSVT instructions were presented visually in the base language and participants were instructed not to talk to experimenter once the aSVT began. After that, a monolingual context was created by presenting the items of the practice session (10 sentences) and entire first block of the experiment (80 sentences) in the base language.

The experimental sessions were conducted in a quiet room where participants were seated in front of a computer, where they read task instructions and filled out the written surveys. After filling out the consent and prescreening forms, participants completed the aSVT task which was administered via PsychoPy (Peirce, 2007). Participants listened to the utterances with headphones and responded by pressing keys on a keyboard. Before the task began, the audio was tested and set at a comfortable listening volume individually for each participant. In order to have a measure sensitive to misperceptions, in addition to the two "true" and "false" options, participants were provided with a separate key (the space bar) to indicate when they failed to understand an utterance (don't understand or DU responses). Furthermore, with the motivation of keeping DU responses as pure as possible, participants were

instructed to guess between "true" or "false" if they managed to understand the sentence but were not sure of the correct answer. Assignment of the true and false responses, informed at the beginning of the task, was counterbalanced to the "z" and "/" keys so that half of the participants provided true responses with their dominant hand and half with their nondominant hand. Response keys were signaled with red illumination. To increase RT sensitivity, participants were told to keep their index fingers resting on these two keys during the task and move them to the DU button as needed. Two self-administered breaks were included after blocks one and two. During this time, which never lasted more than a couple of minutes, participants did not speak to the experimenter.

Recordings were set to play 500 ms after each response or, in case no response was given, 5000 ms after utterance offset. Responses were possible at utterance onset, in line with cascaded theories of speech comprehension (Marslen-Wilson, 1987).

After the aSVT, a manipulation check similar to the speaker rating pilot study was conducted. Participants were presented with a sample of each speaker in English and Dutch and asked to rate their accent (on a scale from 1-native to 9-very strong foreign accent). Participants were asked to guess the regional accent of speakers thought to be native and the native language of speakers rated as non-native.

In addition, participants received a list of all the critical English sentences they had heard during the task in order to indicate words they did not know. Items with unknown words were later removed from analyses on an individual basis. This precaution was taken in order to ensure DU responses reflected failures to understand due to speech processing errors and not to unknown words. Then, participants completed the LBQ in the format of an online survey and the LexTALE, administered in PsychoPy (Peirce, 2007). At the end of the session, participants were debriefed and paid for their collaboration. In all, the session lasted between 30 and 60 minutes.

Results

Of the original 49 participants tested, four were immediately discarded based on the following criteria: technical difficulties during the aSVT which prevented their responses from being recorded (1), having participated in one of the pilot studies (1), having been raised bilingually as indicated on the LBQ (Dutch-German: 1; Dutch-English: 1). The remaining participants were from all over the Netherlands, although nearly half were from the province of Gelderland and all had been living in Gelderland for at least a year (M = 12.0 years, SD = 9.1, range = 1-28). Thus, they all had had exposure to the regional accent of Dutch. When specifically inquired about this, 87.5% indicated being familiar with the regional accent of Dutch spoken by the native Dutch speakers in the aSVT. In terms of native English accents, 78% reported an American accent as one of the accents they are most familiar with (it should be noted that British English was also selected by 75% of participants). Furthermore, 70% responded that they were familiar with the accent of English spoken by the native English speakers during the aSVT. Familiarity with the non-native accents was not as comparable, as can be seen in Table 5, with participants hearing Dutch-accented English more often and from more speakers than English-accented Dutch. This is also apparent from the greater difficulty they had in identifying the native language of the native English speakers in Dutch than of the native Dutch speakers when speaking English.

Since the fact that the study involved English was not mentioned during recruitment and the only indication that participants were proficient in English before the experiment was the question on the SONA prescreening questionnaire, "Do you speak English?," some proficiency selection criteria were considered necessary to ensure that any effects found would not be due to differences in English proficiency. To this end, a general English proficiency score was calculated for each

Table 5.

Familiarity with non-native accents by percent of sample.

	Dutch-accented English	**English-accented Dutch**
Frequency heard		
never	2.5	47.5
less than once a week	45	45
once a week	25	7.5
several times a week	27.5	0
Number of speakers heard from		
0-1	10	80
2-5	50	15
6-10	20	5
>10	20	0

participant by averaging self-ratings across the four skills (speaking, listening, writing, and reading). Two participants were removed from further analyses for not providing information about their English language skills and two more for having a general proficiency score of ≤ 3 (out of 7). The previous preprocessing steps resulted in an unequal number of participants per experimental list: 11 for list 4 and 10 for each of the other lists. Therefore, the last participant tested on list 4 was removed, yielding a final sample size of 40 participants (11 males; Age: M = 23.0, SD = 3.1, range = 18-31). No significant difference ($p > .1$) was found between the lists for any of the measures of English experience, except self-rated reading proficiency, $F(3, 36) = 3.314$, $p = .031$. A post-hoc test revealed that this difference was driven by a difference between lists 1 and 2, both Dutch base language groups, with participants of list 2 presenting greater proficiency than list 1. All participants noted English as their most fluent non-native language except for one who indicated being more proficient in German. The English experience of the final participants is summarized in Table 6.

The data were preprocessed in the following way. First, as explained before, English items containing words that participants indicated not knowing were removed for each participant. Remaining incorrect (true for false statements or false for true statements) responses were counted as errors. Items with ≤70% accuracy in the monolingual block were discarded, as this was the baseline of the experiment. Fourteen items (English: Four items spoken by a native, three by a non-native, Dutch: three by a native, four by a non-native) were removed based on this criterion. The final data set was composed of: 138 errors (4.08% of the observations), 122 DUs (3.61%), and two missing responses (0.06%).

RTs (measured from sentence onset) were processed by first subtracting the duration of the corresponding audio stimuli to adjust them to sentence offset in order to control for differences in sentence duration. RTs for errors and DUs were removed from subsequent RT analyses. Only a participant-based criterion was used for outlier detection. To compensate, a stricter threshold of 3 *SD*s was employed. RTs were considered outliers if they deviated more than 3 *SD*s from factor mean, which factor being determined by language status (base vs. guest), actual language of the item, and context (monolingual vs. bilingual). This resulted in the exclusion of 39 RTs (1.14% of the RT data).

For reasons explained before, here by-item analyses were preferred over by-participant analyses with the main variable of interest being between-

Table 6.

English experience for final participant sample by list.

	List									
	1 base language Dutch		**2 base language Dutch**		**3 base language English**		**4 base language English**		**Overall**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Age of acquisition	10.00	2.58	10.20	1.81	10.90	3.63	11.00	2.21	10.53	2.58
Frequency of use	4.10	1.10	5.00	0.67	4.60	0.70	4.60	0.84	4.58	0.87
LexTALE score	78.88	11.26	82.38	10.73	82.00	9.99	77.13	13.88	80.09	11.32
Word knowledge	98.28	3.72	96.07	7.03	97.89	3.49	96.49	5.73	97.18	5.09
Self-rated English proficiency										
Speaking	4.60	1.35	5.60	0.70	5.30	1.06	5.30	1.16	5.20	1.11
Listening	5.40	1.17	6.00	0.82	5.60	0.84	5.70	0.48	5.68	0.86
Writing	4.70	0.82	5.60	0.97	5.40	1.07	5.00	0.82	5.18	0.96
Reading	5.70	0.67	6.70	0.48	6.10	0.99	6.00	0.67	6.13	0.79
General *(average across skill)*	5.10	0.94	5.98	0.67	5.60	0.83	5.50	0.68	5.54	0.82

Note. Percent of English items with all words known was calculated after removal of items with > 30% errors.

participant. This means that performance on the same item (sentence) produced by the same speaker (and, thus, with the same accent) was compared when it occurred in each of the different critical conditions: in the base or guest language of the experiment and in a monolingual or bilingual context. Considering the impossibility of having guest language items in a monolingual block, the distribution of the factors context and language status were uneven: Base language items could occur in the monolingual and bilingual context, but guest language items only occurred in the bilingual context. Because of this, these two factors were combined into one of three levels which we called "condition." For simplicity's sake, the three levels will from here on be referred to as base monolingual (base language items in the monolingual context), base bilingual (base language items in the bilingual context), and guest language (guest language items necessarily in the bilingual context). Differences between base monolingual and base bilingual items give the effect of context and those between base bilingual and guest language items give the effect of language status. Therefore, repeated measure ANOVAs were run with condition (base monolingual vs. base bilingual vs. guest) as a within-item independent variable and language (Dutch and English) and accent (native vs. non-native) as between-item independent variables and error rate, DU rate, and RT as dependent variables.

On average, participants made 4.10% errors (*SD* = 3.36) and responded DU 4.96% (*SD* = 5.17) of the time. RTs averaged 788.19 (*SD* = 223.18) across all variables. Similarly, items averaged 4.09% errors (*SD* = 5.59), 3.63% DUs (*SD* = 7.18), and RTs of 805.35 ms (*SD* = 223.60). Averages of error rates, DU rates, and RTs by item per factor can be found in Table 7.

An ANOVA on DU rates with condition, language, and accent as independent variables yielded a main effect of condition, $F(2, 220) = 9.050$, $p = .001$, $\eta^2 = .07$. Planned comparisons revealed a difference between the base monolingual ($p = .003$) and bilingual ($p = .001$) conditions, on the one hand, and the guest language condition, on the other. Furthermore, a significant interaction between language and accent on DU rates was also observed, $F(1, 110) = 4.718$, $p = .032$, $\eta^2 = .04$. With planned comparison it was possible to see that this was due to a difference between native and non-native Dutch, $p = .012$. No other significant interactions were observed.

An ANOVA on RTs revealed a main effect of language, $F(1, 110) = 13.476$, $p < .001$, $\eta 2 = .99$, with Dutch sentences being processed faster than English ones. No significant main effect of condition was observed for RT. However, there was also an interaction between condition and language, $F(2, 220)$[3] $= 3.298$, $p = .029$, $\eta^2 = .03$. Planned comparisons indicated that this was due to a difference between the guest condition and the base monolingual and bilingual conditions, but only for Dutch (monolingual-bilingual: $p = .035$. guest-bilingual: $p = .004$; see Table 7 for all values). However, across languages the tendency ($p = .076$) was for the guest language condition to be slower than both monolingual and bilingual conditions. However, as will be seen below, a look at each language individually revealed a different pattern of results.

An ANOVA on error rates revealed no significant main effect nor interactions (all p values > .1). Error rates were included in all of the subsequent analyses but consistently yielded no significant effects. Therefore, they will not be discussed further.

Analysis of the First Guest Language Item

Although much effort was made to make the guest language unexpected, its surprise value probably largely wore off once it began to appear regularly. It follows that the first guest language item was inherently different from the rest and the trial where we thought LWL was most likely to occur. Because of this, the first guest language item was inspected separately. During the construction of the experimental list we made sure that the first guest language item occurred in all of the critical accent-language combinations: native accent in Dutch, non-native accent in Dutch, native accent in English, and non-native accent in English. In Table 8 you will find a summary for these first items in all three conditions. An inspection of DU frequencies revealed that the native English, non-native English, and non-native Dutch first guest language items had the highest DU rates of all sentences in any condition. This was in stark contrast to the native Dutch condition with 0% DU rate for the first guest language item. A chi squared test revealed that there was an association between these groups and DU responses, $\chi^2(3, N = 39) = 17.598$, $p = .001$ (Likelihood ratio). Follow-up analyses (with α Bonferroni-adjusted to .008 to account for the number of comparisons) indicated that there was a significant difference between native

3 Uncorrected degrees of freedom are reported here for aesthetic reasons. However, in actual analyses, degrees of freedom were corrected for violated assumptions.

Table 7.

Average DU rates, error rates, and RTs per variable before and after removal of the first guest language item.

	Variable													
	Language				**Accent**				**Condition**					
	English		**Dutch**		**Native**		**Non-native**		**Monolingual**		**Bilingual**		**Guest**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Before first guest language item removal														
Error rate (%)	4.6	6.1	3.6	5.1	4.6	5.0	4.1	6.2	3.8	6.0	4.1	8.7	4.4	9.6
DU rate (%)	3.6	6.0	3.6	8.2	3.6	6.0	4.6	8.1	2.5	6.6	2.2	7.5	6.2	13.3
RT	884	236	726	180	884	234	799	214	790	258	798	276	839	304
After first guest language item removal														
Error rate (%)	4.5	6.1	3.7	5.1	4.5	4.9	4.1	6.3	3.7	6.0	4.1	8.9	4.4	9.7
DU rate (%)	3.0	5.0	3.1	7.1	3.0	5.3	3.8	6.8	2.4	6.5	2.0	7.4	4.7	9.5
RT	879	238	728	182	879	238	797	212	792	258	796	276	831	299

Dutch and non-native Dutch, $\chi^2(1, N = 20) = 10.208$, $p = .001$ (Continuity correction). The difference between native Dutch and native English was also notable, but not significant at the adjusted α, $\chi^2(1, N = 19) = 4.947$, $p = .026$ (Continuity correction). Analysis of RTs was not possible for all conditions since, due to the large number of DU responses, insufficient data points remained for RTs (only 2-5 for all conditions except native Dutch). An ANOVA for the native Dutch conditions showed that there was a main effect of condition, $F(2, 29) = 11.933$, $p < .001$, $\eta^2 = .88$, with Tukey post-hoc tests revealing that RTs were slower for the first guest language item than the same item in the monolingual ($p < .001$) and bilingual conditions ($p = .014$), while there was no significant difference between these last two ($p = .195$).

Analysis Without First Guest Language Item

Given the relatively high DU rates overall for the first guest language items, it is possible that these items were mainly responsible for the effects observed. What is more, these rates indicate that first guest language items are very different from the rest of the items and thus may involve different processes. Therefore, it was considered prudent to re-conduct the analyses without these sentences to ensure the condition effects reported above remained intact and did not only occur for the first unexpected item (see averages in Table 7) ANOVAs without the first guest language items and their counterparts in the other two conditions yielded, once again, a main effect of language on RTs, $F(1, 106) = 13.105$, $p < .001$, $\eta^2 = .11$, with Dutch sentences being processed faster ($M = 742.76$, $SE = 31.16$) than English ones ($M = 902.26$, $SE = 31.16$). The main effect of condition on DU rates observed before was also found here, $F(2, 212) = 6.030$, $p = .004$, $\eta^2 = .05$. Planned comparisons showed a significant difference between base monolingual ($p = .014$) and base bilingual conditions ($p = .003$), on the one hand, and guest language, on the other, while the difference between base monolingual and bilingual conditions was not significant ($p = .575$). Therefore, the main effects observed with the first guest language item included remained significant. However, the interactions between condition and language for RTs and language and accent for DU rates were only marginally significant here ($p = .078$

Table 8.

Comparison DU rates, error rates, and RTs for first guest language sentences and same sentences in other conditions.

	Condition											
	Base monolingual				**Base bilingual**				**Guest**			
	DU (%)	**Err (%)**	**M RT *(SD)***	***N***	**DU (%)**	**Err (%)**	**M RT *(SD)***	***N***	**DU (%)**	**Err (%)**	**M RT *(SD)***	***N***
Dutch												
Native	0	0	381 (166)	10	0	0	686 (394)	10	0	0	1206 (505)	10
Nonnative	20	0	699 (390)	10	10	0	483 (314)	10	80	0	2157 (1827)	10
English												
Native	0	11	709 (157)	9	20	10	1217 (507)	10	56	11	863 (112)	9
Nonnative	0	10	1158 (826)	10	0	0	1037 (423)	9	50	10	1742 (417)	10

Note. Err = error, *N* = total cases.

and $p = .066$, respectively). No other significant interactions were observed.

Given their volatility, in order to further test the reliability of these interactions, an additional analysis was run after removing items with a small cell size (< 5 out of a maximum of 10) to ensure these were not biasing the results. This led to the exclusion of three items, with a total of 53 and 54 items remaining for English and Dutch, respectively. Following analyses, the main effects of condition and language remained. However, the interactions between language and condition, on the one hand, and language and accent, on the other, were not significant ($p = .087$ and $p = .092$, respectively).

Discussion

The aim of the present study was two-fold: On the one hand, given the scarcity of the literature on the topic, we wanted to see if it was possible to induce LWL states in a laboratory setting. To this end, we came up with a novel design that would bias bilinguals towards expecting one or the other of their languages, called the base language of the experiment. Our inclusion of a "don't understand" response option allowed us to more precisely measure comprehension and, our main interest, failures to comprehend.

Our second aim was to evaluate the role different factors play in the occurrence of LWL. In particular, we thought speech misperceptions would occur differentially in the bilingual's two languages and that they would be augmented when listening to the speech of a non-native speaker.

Analysis of the first guest language item revealed that the manipulation was indeed successful. After a monolingual block in the base language, participants were surprised with an item in their other language, the guest language of the experiment. Failures to comprehend, as indexed by DU rates, those first guest language items were the greatest in the experiment.

First of all, this study brings to light the importance of the sensitivity of your measurement tool. Specifically, many studies in cognitive science, especially those using button presses where RT is vital, do not provide participants with an option to indicate insecurity. This may lead to inaccurate responses and RTs, which actually reflect misunderstandings. This problem is even more serious in studies on non-native speech comprehension, where comprehension is more taxed. Here we hope to have demonstrated the value of data commonly piled together with other error rates.

One concern when deciding to include the DU button was that participants would use it as a "don't know" option when they did actually understand the utterance, but did not know the answer. Several precautions were taken in order to ensure this did not happen, such as telling participants to guess in case they understood the answer but did not know whether the statement was true or false. In addition, items with words participant were not familiar with (in English) were removed. The strongest evidence against this explanation, however, stems from the fact that DU rates, for the same item, were higher when that item occurred in the guest language condition. This, together with the observation that

the final data do not reveal a difference in DU rates for English and Dutch, is strong support for the claim that misperceptions cannot be (entirely) attributed to a lack of knowledge or low L2 proficiency.

On the other hand, DU rates were not null in the other conditions. Furthermore, as a result of our manipulation check, we know that only 70% of participants claimed to have experienced a LWL situation during the experiment. While this could be viewed as a high success rate for a new experimental manipulation, it begs the question: can we be sure that the failures to comprehend that we observed were really due to LWL or could they just be explained as a language switching cost? Indeed, it may be difficult to disentangle these two concepts, and that is because LWL is a form of language switching. It could be defined as a failure or delay in speech comprehension due to non-target language processing. Effectively, resolution of a LWL state requires a language switch in speech perception mechanisms. The effect of condition found here was, by design, caused by the participant having to change from the base language of the experiment to the less frequent guest language. This does not mean that all DU responses were necessarily caused by LWL, but the fact that these rates were higher for the guest language condition suggests that making participants respond to an item in a different language than the previous one increases the likelihood of a comprehension failure. Of course, not unlike many other psycholinguistic processes, the occurrence of LWL does not presuppose consciousness. In fact, people misperceive speech in the same language all the time, without necessarily knowing why. Still, stronger evidence of LWL could be found with other designs, for example using target competitors, to demonstrate activation of the non-target language.

Returning to our predictions, we expected that guest language items would result in more DUs and slower RTs. As concerns the first, guest language items did produce more comprehension failures than base language items and this effect did not differ for the L1 or L2, nor for native vs. non-native speech[4]. While prior studies have demonstrated increased difficulties in processing gated guest words (e.g., Grosjean, 1988; Schulpen et al., 2003) or intra-sentential code-switches (Fitzpatrick & Indefrey, 2014), we are not aware of any previous evidence of complete breakdowns in comprehension during bilingual sentence comprehension. Methodologically, the fact that DU rates were observed at all is promising for future studies on LWL or similar bilingual speech misperceptions. Theoretically, this suggests that bilingual speech comprehension can at least partially proceed in a language-selective manner, for example, when strongly biased by the context, both local (previous item) and global (entire experiment). While the present study cannot be said to support any particular model of bilingual speech comprehension, since they all concern word recognition, the present findings could hint at the fact that comprehension proceeds in a more language-selective fashion during sentence processing, as has been suggested by some researchers (Fitzpatrick & Indefrey, 2014; Lagrou et al., 2012; Li, 1996).

Interestingly, no significant difference was observed between the monolingual and bilingual base language conditions. Studies on non-selective lexical access have been finding a cost for the mere activation of the non-target language since seminal studies (e.g., Kolers, 1966; Macnamara & Kushnir, 1971) showing a processing cost for reading mixed language passages compared to monolingual passages. However, RTs to sentences, measured here, may not have been precise enough to reflect these differences. In addition, as mentioned before, the comparison made here is not ideal as the monolingual block always preceded the bilingual blocks. Further study would be needed to confirm these observations.

Regarding the second part of our prediction on guest language processing costs, we found that DU rates were not accompanied by RT differences. One possible explanation is that DUs and RTs reflect the use of different strategies, with participants using the DU option when deciphering utterances proves too daunting. However, analyses of RTs to DU responses seem to suggest that participants do try to understand these utterances before "giving up," taking, numerically, on average longer than for correct responses (DU: $M = 1480.89$, $SD = 776.16$; correct responses: $M = 790.55$, $SD = 516.13$).

Another possible explanation for the lack of relationship between DU rates and RTs has to do with the nature of the task, which was meant to be challenging. With short words and sentences, and utterances presented shortly after responses were given, the task was designed to make sure participants had to tune in to the utterances quickly. In cases where LWL has to do with a speech segmentation problem, taking longer to process the utterance might not present a real benefit: once a part of the utterance was misperceived there was little chance of recovering it (except, perhaps, via a phonological loop mechanism).

4 It should be noted that these were still rare, as suggested by our initial survey on LWL incidence.

In addition to inducing speech misperceptions, we were interested in examining the role of two factors on its occurrence: language and accent. As regards language, we expected that whether the guest language was the L1 or L2 would influence the incidence of speech misperceptions. In terms of the direction, we equally envisioned the two possible directions. Findings of a reduced baseline activation for the L2 would seem to predict a greater incidence of DU responses for items in this language, in line with Schulpen et al. (2003)'s study. On the other hand, studies demonstrating an asymmetric switch in bilingual speech production would have one believe that during L2 processing, the L1 is heavily inhibited, which could manifest itself as a greater processing cost for L1 guest language items.

Regarding our second variable of interest, accent, we thought that guest language utterances produced with non-target language pronunciation would hinder comprehension by tilting the system towards the non-target language. Support for this idea comes from gating studies showing that guest words pronounced with guest language phonetics are identified faster than those pronounced in base language phonetics (Grosjean, 1988; Li, 1996). However, there is also evidence that sentential context can help reduce the amount of non-target language interference during auditory comprehension in bilinguals (see Fitzpatrick & Indefrey, 2014 for a review).

In line with previous findings, participants were faster to respond in their L1 than in their L2 (Proverbio, Leoni, & Zani, 2004; Schulpen et al., 2003). Furthermore, analysis of the first guest language item revealed that not all conditions were equally surprising. Listeners never failed to understand utterances produced by native speakers of their L1, although RTs revealed a delay in comprehension relative to the same item in the base language. This was in stark contrast to non-native speech in Dutch, despite it being participants' L1. Speaker accent did not modulate comprehension of speech in the L2. These differences are likely due to familiarity with the accents, with Dutch speakers being less exposed to non-native Dutch than native Dutch and both native and non-native English. Some studies have found that previous experience with a particular accent can increase comprehension of that accent (Witteman et al., 2013).

When the first guest language item was excluded from the analyses, language and accent no longer played a role. This suggests that language and accent may only initially play a role, when other information is not available. These findings are in line with studies showing that bilinguals are able to quickly adapt to unfamiliar accents (Witteman et al., 2013). A similar effect was observed in the study on the role of facial cues on bilingual production (Woumans et al., 2015). In that study, prior to the task, bilinguals interacted with speakers in one of their languages. Later, during a noun-verb association task, speakers' faces were presented while producing noun stimuli to elicit participant responses. Crucially, speech could be produced in the same language spoken by that speaker before (congruent trials) or a different language (incongruent trials). The results revealed a difference between congruent and incongruent trials for the first six trials only, with slower RTs observed for incongruent trials. This difference was not evident, however, in later trials. The authors interpreted these findings by proposing that faces are used as cues for language production as long as they are considered reliable. If expectations are violated, however, the association between speaker face and language can be weakened. These findings are also consistent with Elston-Güttler et al. (2005)'s finding that priming effects from prior exposure to a film in one language decreased throughout the experiment.

Extended to the present study, a claim could be made for the surprise effect of guest language productions decreasing after this language made its first appearance in the experiment. Rather than completely dissipating, however, a guest language processing cost remained, although this no longer differed for the L1 and L2 nor for utterances produced by native and non-native speakers. Nonetheless, given the small number of items available for the first guest language analysis, additional studies are necessary with a great deal more items and participants, in order to properly test for an effect of an entirely unexpected guest language, as well as a difference between early and later trials in that guest language.

Analyses without the first guest language item also showed that, after the first guest language item, accent did not exacerbate the effect of the guest language. This lends support to the claim that phonetic information may be more relevant when processing isolated words than sentences, where other information can help decipher meaning. These results are similar to those of Fitzpatrick and Indefrey (2014) who reported symmetrical switch costs in ERPs. However, more conclusive evidence could be provided by future studies, particularly where the monolingual block could be counterbalanced and allowing for within-participant analyses and an equal amount of switching into the L1 and into the L2.

It should be noted that, despite not finding evidence for the role of these factors here, it is

likely that LWL is so rare that the roles of these factors are hard to assess. In fact, this is why much psycholinguistic research relies on types of stimuli that are not very common in everyday life, like interlingual homophones. Although we did not find evidence for a role of language or accent in guest-language induced speech misperceptions, we do not deny the possibility that these aspects, and others, can modulate the occurrence of LWL outside the laboratory. Here we have developed a paradigm that has proven effective in inducing speech misperceptions. Further studies can explore ways to increase the likelihood of LWL states in an experimental setting, such as increasing cognitive load or adverse listening conditions.

Here we were interested in a rare failure in comprehension that occurs in bilinguals. Misperceptions are valuable to research in that, by highlighting what can go wrong, they can provide insight into the processes underlying normal speech comprehension. However, it goes without saying that, outside of the laboratory, in the real world, they are probably a lot less detrimental than research would lead one to believe. Indeed, in everyday life, bilinguals manage to communicate and code-switch without major problems.

References

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical data base on (CD-ROM).* Philadelphia, PA: Linguistic Data Consortium.

Beck, C., Kardatzki, B., & Ethofer, T. (2014). Mondegreens and Soramimi as a method to induce misperceptions of speech content - Influence of familiarity, wittiness, and language competence. *PLoS ONE, 9*(1).

Beck Lidén, C., Krüger, O., Schwarz, L., Erb, M., Kardatzki, B., Scheffler, K., … & Ethofer, T. (2016). Neurobiology of knowledge and misperception of lyrics. *NeuroImage, 134,* 12–21.

Bond, Z. S. (2008). Slips of the Ear. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 290–310).

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning, 22*(2), 203–217.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Bürki-Cohen, J., Grosjean, F., & Miller, J. L. (1989). Base-language effects on word identification in bilingual speech: Evidence from categorical perception experiments. *Language and Speech, 32*(4), 355–371.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 1029–1040.

Cheng, Y.-L., & Howard, D. (2008). The time cost of mixed-language processing: An investigation. *International Journal of Bilingualism, 12*(98568), 209–222.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 136–163). Hillsdale, New Jersey: Erlbaum.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*(2), 240–247.

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin & Review, 11*(6), 1084–1089.

Cox, E. A. (2005). *Second language perception of accented speech.* (Doctoral dissertation, University of Arizona, 2005). Retrieved from Digital Dissertations, (UMI No. 3177525).

Cutler, A. (2012). *Native listening: language experience and the recognition of spoken words.* Cambridge, Massachusetts: The MIT Press.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141–201.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*(4), 385–400.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 19*(1), 1–16.

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review, 15*(4), 850–855.

Elston-Güttler, K. E., Gunter, T. C., & Kotz, S. A. (2005). Zooming into L2: Global language context and adjustment affect processing of interlingual homographs in sentences. *Cognitive Brain Research, 25*(1), 57–70.

Fitzpatrick, I., & Indefrey, P. (2010). Lexical competition in nonnative speech comprehension. *Journal of Cognitive Neuroscience, 22*(6), 1165–1178.

Fitzpatrick, I., & Indefrey, P. (2014). Head start for target language in bilingual listening. *Brain Research, 1542*, 111–130.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America, 9*(1), 370–389.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1276–1293.

Gooskens, C., & Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using

Norwegian dialect data. *Language Variation and Change, 16*(3), 189–207.

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language, 122*(2), 92–102.

Gough, P. B. (1966). The verification of sentences: The effects of delay of evidence and sentence length. *Journal of Verbal Learning and Verbal Behavior, 5*(5), 492–496.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 344–359.

Grosjean, F. (1988). Exploring the recognition of guest words in bilingual speech. *Language and Cognitive Processes, 3*(3), 233–274.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*(1), 3–15.

Grosjean, F. (1997). Processing mixed language: Issues, findings, and models. In A. M. B. De Groot & J. F. Kroll (Eds.), *Tutorials in Bilingualism: Psycholinguistic Perspecitives* (pp. 225–254). Mahwah, New Jersey: Lawrence Erlbaum Publishers.

Grosjean, F., & Miller, J. L. (1994). Going in and out of languages: An example of bilingual flexibility. *Psychological Science, 5*(4), 201–202.

Hartsuiker, R. J. (2015). Visual cues for language selection in bilinguals. In R. Kumar Mishra, N. Srinivasan, & F. Huettig (Eds.), *Attention and Vision in Language Processing* (pp. 129–145). New Delhi: Springer India.

Hendriks, P. (2014). *Asymmetries between Language Production and Comprehension.* Dordrecht: Springer.

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods, 46,* 702–721.

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science, 15*(5), 314–318.

Kentner, G. (2015). Rhythmic segmentation in auditory illusions- evidence from cross-linguistic Mondegreens. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 0–4.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650.

Koch, I., Prinz, W., & Allport, A. (2005). Involuntary retrieval in alphabet-arithmetic tasks: Task-mixing and task-switching costs. *Psychological Research, 69*(4), 252-261.

Kolers, P. A. (1966). Reading and talking bilingually. *American Journal of Psychology, 79*, 357–376.

Lagrou, E., Hartsuiker, R. J., & Duyck, W. (2012). The influence of sentence context and accented speech on lexical access in second-language auditory word recognition. *Bilingualism: Language and Cognition, 16*(2013), 1–10.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*(2), 325–343.

Lévy, N. (2015). *Computational psycholinguistics and spoken word recognition in the bilingual and the monolingual.* Université de Neuchâtel.

Li, P. (1996). Spoken word recognition of code-switched words by Chinese–English bilinguals. *Journal of Memory and Language, 35*(6), 757–774.

Li, P. (1998). Mental control, language tags, and language nodes in bilingual lexical processing. *Bilingualism: Language and Cognition, 1*(2), 92–93.

Li, P., & Farkaš, I. (2002). A self-organizing connectionalist model of bilingual processing. In R. Heredia & J. Altarriba (Eds.), *Bilingual Sentence Processing* (pp. 59–85). North Holland: Elsevier Science Publisher.

Macnamara, J., & Kushnir, S. L. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior, 10*(5), 480–487.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1), 71–102.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*(1), 21–46.

McQueen, J. M., & Cutler, A. (2010). Cognitive processes in speech perception. In W. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 489–520). Wiley-Blackwell.

Meuter, R. F. I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40*, 25–40.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 193–218).

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*(3), 289–306.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The Role of Speaking Rate. *Studies in Second Language Acquisition, 23*(4), 451–468.

Nelson, K. E., & Kosslyn, S. M. (1975). Semantic retrieval in children and adults. *Developmental Psychology, 11*(OCTOBER), 807–813.

Otake, T. (2007, August). Interlingual near homophonic words and phrases in L2 listening: Evidence from misheard song lyrics. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, 777–780.

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science, 12*(6), 445–449.

Pecher, D., De Rooij, J., & Zeelenberg, R. (2009). Does a pear growl? Interference from semantic properties of orthographic neighbors. *Memory & Cognition, 37*(5),

541–546.
Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13.
Piske, T., Mackay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*, 191–215.
Proverbio, A. M., Leoni, G., & Zani, A. (2004). Language switching mechanisms in simultaneous interpreters: An ERP study. *Neuropsychologia, 42*(12), 1636–1656.
Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics, 20*, 331–350.
Romero-Rivas, C., Martin, C. D., & Costa, A. (2016). Foreign-accented speech modulates linguistic anticipatory processes. *Neuropsychologia, 85*, 245–255.
Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2015). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives.*
Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15*(01), 157–166.
Schepens, J., Dijkstra, T., Grootjen, F., & Van Heuven, W. J. B. (2013). Cross-language distributions of high frequency and phonetically similar cognates. *PLoS ONE, 8*(5).
Schulpen, B., Dijkstra, T., Schriefers, H. J., & Hasper, M. (2003). Recognition of interlingual homophones in bilingual auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 29*(6), 1155–1178.
Shook, A., & Marian, V. (2013). The Bilingual Language Interaction Network for Comprehension of Speech. *Bilingualism: Language and Cognition, 1*6(2), 304–324.
Soares, C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory & Cognition, 12*(4), 380–386.
Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281–284.
Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America, 126*(1), 367.
Van Der Haagen, M. J. (1998). *Caught between norms: The English pronunciation of Dutch learners.* The Hague: Holland Academic Graphics.
Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50*(1), 1–25.
Weissberger, G. H., Wierenga, C. E., Bondi, M. W., & Gollan, T. H. (2012). Partially overlapping mechanisms of language and task control in young and older bilinguals. *Psychology and Aging, 27*(4), 959–974.
Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics, 75*(3), 537–556.
Woumans, E., Martin, C. D., Vanden Bulcke, C., Van Assche, E., Costa, A., Hartsuiker, R. J., & Duyck, W. (2015). Can faces prime a language? *Psychological Science, 26*(9), 1343–1352.
Wright, S. (1954). The death of Lady Mondegreen. *Harper's Magazine, 209*(1254), 48–51.
Yule, H., & Burnell, A. (1903). *Hobson-Jobson: a glossary of colloquial Anglo-Indian words and phrases, and of kindred terms, etymological, historical, geographical and discursive.* London: John Murray.
Zhao, X., & Li, P. (2007). Bilingual lexical representation in a self-organizing neural network model. *Proceedings of the 29th Meeting of the Cognitive Science Society, 1*, 755–760.
Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism, 13*(5), 505–524.

Abstracts

Proceedings of the Master's Programme Cognitive Neuroscience is a platform for CNS students to publish their Master thesis. Given the number of submissions, we select the articles that received the best reviews, under recommendation of our editors, for the printed edition of the journal. The abstracts of the other articles are provided below, and for interested readers a full version is available on our website: www.ru.nl/master/cns/journal.

Reactivation of Complex Events and Preactivation in Humans

Claudia van Dun, Silvy Collin, Christian F. Doeller

The hippocampus' role in memory integration is starting to be understood. However, the exact mechanisms by which existing memories affect newly acquired information remain a matter of debate. We present two approaches to investigate this topic. It has been demonstrated that reactivation during offline periods is important for memory integration. In previous studies, reactivation of relatively simple stimuli has been investigated but evidence for reactivation of complex events has been lacking. For our first experiment, we used multivariate pattern analyses to assess reactivation of complex events. Analyses of the resting-state blocks of a functional magnetic resonance imaging (fMRI) memory integration paradigm with complex life-like stimuli showed no evidence for reactivation. We discuss methodological limitations that could explain these results.

In rodents, it has been found that not only reactivation of previously encoded information, but also preactivation of to be encountered information can be beneficial for memory integration. Contrary to reactivation, which refers to experiences in the past, preactivation refers to activation corresponding to events in the future. To date, preactivation has not been assessed in humans. In our second experiment, we present a novel behavioural reaction time paradigm to assess preactivation. The results strongly suggest that humans show preactivation of to be encountered events. For future research, we suggest an fMRI version of the preactivation paradigm used here to shed light on the hippocampus' role in preactivation.

Frequency Tagging at High Frequencies in Downstream Areas Under Influence of Attention

Jerome Herpers, Ole Jensen, Jim Herring

Frequency tagging can be used to study the downstream flow of information from visual cortex to higher order areas. This flow of information is modulated by attention, which might be mediated by cross-frequency coupling of low and high frequencies. To investigate the influence of low frequencies on stimulus processing at high frequencies mediated by attention, the feasibility of using frequency tagging at high frequencies in downstream areas was investigated. Furthermore, the effects of spatial and object-based attention on the magnitude of the elicited response by the frequency tags were examined. Stimuli of faces and houses were presented at high frequencies of 63 Hz and 78 Hz during a catch trial flip detection task. Results show that spatial attentional modulation increased activity in occipital cortex contralateral to the attended stimulus. When examining the effect of object-based attention, no interpretable activity patterns were observed. As a result of interaction between spatial and object-based attention, fusiform gyrus and parahippocampal gyrus in the right hemisphere, but not left, showed enhanced activity in tagged frequencies. With frequency tagging at high frequencies, it will be possible to investigate communication between regions, as well as information processing by phase-to-power cross-frequency coupling under influence of spatial and object-based attention.

Early Life Stress Induces Persistent Alteration in Endocannabinoid System and Leads to Dysfunctional Modulation of Emotional Memory Retrieval

Sara Jamil, Piray Atsak, Benno Roozendaal

Early life stress (ELS) is one of the best characterised risk factors for later development of stress-related disorders, such as posttraumatic stress disorder (PSTD). A hallmark feature in PTSD is persistent, uninhibited retrieval of emotional memory. Recent evidence from our lab indicated that glucocorticoids interact with the endocannabinoid system, particularly 2-arachidonoyglycerol (2- AG), to impair the retrieval of emotional memory under stress. Given that adult rats with ELS history show an inability to upregulate 2-AG signalling in hippocampus after acute stress, we hypothesised that glucocorticoids will not impair emotional memory retrieval in ELS animals, whereas a direct augmentation of hippocampal 2-AG signalling will. We first showed that the well-established limited nesting paradigm resulted in fragmented maternal care and elevated plasma corticosterone levels in pups. At adulthood, we trained male offspring on a contextual fear memory paradigm. One hour before retention testing, 24 hours after training, rats were injected with corticosterone (CORT, 3 mg/kg) systemically or administered the 2-AG hydrolysis (MAGL) inhibitor KML-29 (0.2 μg/0.5 μL) directly into the hippocampus. Unlike control rats, we found that systemic CORT injection did not impair retrieval of contextual fear memory in ELS animals. By contrast, direct hippocampal administration of KML-29 impaired memory retrieval in both ELS and control rats in a CB1 receptor-dependent fashion. Thus, these findings support our hypothesis that the inability of ELS rats to modulate memory retrieval under stress might originate from their inability to mount a 2-AG response. Our findings are highly relevant for informing future studies on the link between ELS and maladaptive stress coping, and the increased risk for stress-related psychopathologies.

One Self, Too Many Tasks: Bimanual Interference from a Predictive Coding Framework

Sarit Pink-Hashkes, Luc P.J. Selen, Johan H.P. Kwisthout

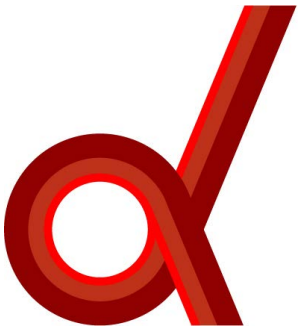
In this work, we used the predictive coding framework to examine bimanual interference in human hand movements. Based on previous experiments, we hypothesise that bimanual interference can be understood from a similar framework as binocular rivalry, as a bi-stable system created as a Bayesian optimal solution dealing with "unecological" conditions under strong hyperpriors learnt by the brain in "ecological" conditions. Specifically, we postulate that the layer of the brain in which a single minimal-self is created to predict the correlation of information from the different modalities includes a hyperprior that only one task goal is possible at any given time. While most tasks require many effectors to work together as one coordinated unit and not interfere with each other, like riding a bike, eating with fork and knife or driving a stick shift car, under usual "ecological" conditions, these actions emerge as a solution to a single task goal and individual motion paths are undefined under the task goal. We tested this hypothesis by manipulating top down task goal and bottom up visual feedback of subjects' own hands in an immersive virtual reality environment. We instructed subjects to either follow an avatar's motion or create a self-motion and manipulated the visual feedback to influence the predictions created by the minimal-self. Our main findings are that providing false visual feedback that is in total opposition to the minimal-self predictions lowered interference in the follow task and increased interference in the self-task. We further discovered that providing a first person view, by showing the subject performing bimanual independent movements, increased the interference of the hands despite subjects' belief that the task is easier. We explain these results using the predictive coding framework and discuss the implications regarding possible rehabilitation programs and notions regarding the relative weakness of the proprioceptive system in comparison to the visual system.

Characterization of Age-Related Myelination Deficits in a Rat Model for Schizophrenia

Marigoula Vlassopoulou, Astrid Vallès Sanchez, Gerard J.M. Martens

Schizophrenia (SZ) is a debilitating neuropsychiatric disorder that affects millions of people around the world. A growing body of evidence points towards the involvement of hypomyelination of the prefrontal cortex (PFC) in the development of the cognitive symptoms of the disorder. Our hypothesis is that SZ patients exhibit elevated oxidative stress (i.e., redox imbalance) throughout the brain early in development and that this disrupts the ongoing process of myelination in the late-maturing PFC, resulting in clinical manifestations of cognitive dysfunction. To tackle the neurophysiological underpinnings and postnatal developmental timing of such a mechanism, we combined molecular and cellular analysis in the apomorphine-susceptible (APO-SUS) rat model for schizophrenia, using apomorphine-unsusceptible (APO-UNSUS) rats as their control counterparts. First, we assessed the mRNA expression levels of genes related to myelin, oxidative stress and oligodendrocytes (OLs, the myelinating glia of the central nervous system) in the medial PFC (mPFC), striatum and corpus callosum of post-natal day (PND) 21 and PND28 male animals. Secondly, we used immunohistochemistry to assess differences in the percentage of OLs and their precursors in the mPFC, striatum, corpus callosum and hippocampus of PND21 and PND90 male rats, as well as the extent of myelination in axons of the mPFC and barrel cortex of these animals. The results showed that, while differential expression in redox-related genes is already present at PND21 in multiple brain regions, myelin- and OL-related molecular and cellular abnormalities are more prominent in animals aged PND28 and older, and occur specifically in the APO-SUS mPFC. The results of this study provide more insights into the neuropathology of SZ and clues for developing therapeutic approaches for the disorder.

Institutes associated with the Master's Programme Cognitive Neuroscience



Donders Institute for Brain, Cognition and Behaviour:
Centre for Cognitive Neuroimaging
Kapittelweg 29
6525 EN Nijmegen

P.O. Box 9101
6500 HB Nijmegen
www.ru.nl/donders/

Max Planck Institute for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen

P.O. Box 310
6500 AH Nijmegen
http://www.mpi.nl

Radboudumc

Radboudumc
Geert Grooteplein-Zuid 10
6525 GA Nijmegen

P.O. Box 9101
6500 HB Nijmegen
http://www.umcn.nl/

Baby Research Center
Montessorilaan 3
6525 HR Nijmegen

P.O. Box 9101
6500 HB Nijmegen
http://babyresearchcenter.nl