CORTEX 63 (2015) 145–154

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research report

Transcranial direct current stimulation over the left prefrontal cortex increases randomness of choice in instrumental learning

Zsolt Turi [a,*,1], Matthias Mittner [b,1], Alexander Opitz [a,1], Miriam Popkes [a], Walter Paulus [a] and Andrea Antal [a]

[a] Department Clinical Neurophysiology, University Medical Center, Georg-August University, Göttingen, Germany
[b] Department of Psychology, University of Tromsø, Norway

ARTICLE INFO

Article history:
Received 6 January 2014
Reviewed 1 April 2014
Revised 13 May 2014
Accepted 26 August 2014
Action editor Jacinta O'Shea
Published online 11 September 2014

Keywords:
Transcranial direct current stimulation
Dorsolateral prefrontal cortex
Probabilistic learning task
Exploration
Exploitation
Working memory

ABSTRACT

Introduction: There is growing evidence from neuro-computational studies that instrumental learning involves the dynamic interaction of a computationally rigid, low-level striatal and a more flexible, high-level prefrontal component.
Methods: To evaluate the role of the prefrontal cortex in instrumental learning, we applied anodal transcranial direct current stimulation (tDCS) optimized for the left dorsolateral prefrontal cortex, by using realistic MR-derived finite element model-based electric field simulations. In a study with a double-blind, sham-controlled, repeated-measures design, sixteen male participants performed a probabilistic learning task while receiving anodal and sham tDCS in a counterbalanced order.
Results: Compared to sham tDCS, anodal tDCS significantly increased the amount of maladaptive shifting behavior after optimal outcomes during learning when reward probabilities were highly dissociable. Derived parameters of the Q-learning computational model further revealed a significantly increased model parameter that was sensitive to random action selection in the anodal compared to the sham tDCS session, whereas the learning rate parameter was not influenced significantly by tDCS.
Conclusion: These results congruently indicate that prefrontal tDCS during instrumental learning increased randomness of choice, possibly reflecting the influence of the cognitive prefrontal component.

1. Introduction

In most everyday situations, we constantly have to adapt and optimize our behavior to cope with various, often conflicting, demands and constraints posed by each specific environment. An important aspect of adaptive behavior is the capability of choosing those actions that lead to a high amount of cumulative reward. One way to achieve this goal is by successively

* *Corresponding author*. Department Clinical Neurophysiology, University Medical Center, Georg-August University, Robert-Koch-Str. 40, Göttingen D-37075, Germany.
E-mail address: zsoltturi@gmail.com (Z. Turi).
[1] These authors contributed equally.
http://dx.doi.org/10.1016/j.cortex.2014.08.026
0010-9452/

generating predictions about the consequences of each action. Generating and using these predictions to guide behavior is known as instrumental learning (Dayan & Balleine, 2002).

Instrumental learning in humans recruits multiple, functionally interacting and parallel brain systems (for reviews see Dolan & Dayan, 2013; Samson, Frank, & Fellous, 2010); these involve a striatal reinforcement learning (RL) component and a cognitive, prefrontal control component (Collins & Frank, 2012; Daw, Niv, & Dayan, 2005; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), also known respectively as the model-free and model-based controls of instrumental learning (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw et al., 2005; Wunderlich, Smittenaar, & Dolan, 2012). The low-level RL (or model-free) component is characterized by computational rigidity and it requires a large number of learning trials to gradually integrate the long-term probability of reinforcement values in response to probabilistic reward associations (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007).

The high-level cognitive (or model-based) component, driven by the prefrontal system, has greater computational flexibility as it dynamically computes the policy to optimize behavior by evaluating the instrumental requirements of the decision situation (Daw et al., 2006). On the one hand, this is achieved by actively maintaining the reinforcement history in working memory (WM) which permits fast goal-directed decisions, albeit with the restriction of a limited capacity (Collins & Frank, 2012; Frank, et al., 2007). On the other hand, functional neuroimaging evidence also suggests that the prefrontal system controls adaptive exploration (Daw et al., 2006). Further evidence also indicates the role of prefrontal involvement specifically, as individual genetic differences in regulating prefrontal dopamine (DA) Catechol-O-methyltransferase (COMT) rs4680 single nucleotide polymorphism has an impact on exploratory behavior but not on the level of striatal DA (Frank, Doll, Oas-Terpstra, & Moreno, 2009).

Nevertheless, genetic studies are correlational in nature and a more direct demonstration of the involvement of the prefrontal component in cognitive control in instrumental learning requires a focal interference with prefrontal regions. Transcranial direct current stimulation (tDCS) has the potential to temporarily shift neuronal membrane potentials of a given neuronal population by passing a low-intensity electrical current through the brain (Nitsche & Paulus, 2000). These physiological effects have been linked to changes in a wide range of cognitive functions, including those that are related to the prefrontal cortex, such as WM (e.g., Zaehle, Sandmann, Thorne, Jäncke, & Herrmann, 2011) or prototype learning (Ambrus et al., 2011).

Modeling studies investigating the tDCS-induced current profile characteristics indicate that the effect of tDCS, at least from electrodes in close spatial proximity, is primarily limited to the neocortex (Datta, Elwassif, Battaglia, & Bikson, 2008; Faria, Hallett, & Miranda, 2011), although tDCS may have the ability to remotely activate deeper brain structures, such as the striatal system (Chib, Yun, Takahashi, & Shimojo, 2013). The common notion that anodal tDCS leads to an increase and cathodal tDCS leads to a decrease in neuronal excitability in the brain area underneath the electrode have been challenged by recent evidence (Reato et al., 2013). First, the electric field induced by tDCS can both de- and hyperpolarize within the same gyrus (Reato et al., 2013) and second, different types of neurons are differentially modulated depending on their morphology and axonal orientation (Radman, Ramos, Brumberg, & Bikson, 2009). Hence, a simple mechanistic relation between polarity and expected behavioral changes may be difficult to establish. Indeed, recent evidence suggests that tDCS has less consistency in polarity effects in cognitive tasks compared to basic motor functions (Jacobson, Koslowsky, & Lavidor, 2012).

The aim of the present experimental work has been to study, which component of instrumental learning was influenced by prefrontal tDCS by evaluating the effect of anodal tDCS on behavior as measured by accuracy and computational model parameters. Advances in computational modeling of RL using Q-learning algorithms allow distinct processes to be modeled in detail. This entails the ability to derive information about how performance is affected by specific behavioral influences or strategies by fitting the RL model to behavioral data (e.g., Frank et al., 2009).

In the classical model we employed in this study (Jocham, Klein, & Ullsperger, 2011), the learning rate parameter α reflects the impact of the prediction error (i.e., the difference between the previous outcome estimate and the actual estimate after a certain action). Larger α values reflect trial-to-trial fluctuations (a recency effect), whereas lower values indicate a gradual value integration and more stable value estimation (Frank et al., 2007). If prefrontal anodal tDCS biases participants to rely more on the WM component, we expected to observe a trial-to-trial behavioral adjustment (i.e., change of decision after negative response) during learning and an increased α value. In contrast, if anodal tDCS compels participants to rely less on the WM component, then a lower α value and less trial-to-trial behavioral adjustment will be observed – which would increase outcome-dependent exploitation of the better symbol. In addition, the β parameter, also known as the "temperature" or "noise" parameter, reflects the learners' bias towards either exploitation (i.e., choosing the better option in case of lower β values) or exploration (i.e., choosing the items more randomly; higher β values) (Frank et al., 2007; Jocham, et al., 2011). This model is designed to capture behavior in a probabilistic environment where not only the expected value (determined by integrating past outcomes with learning rate α) determines the decision, but choices are also characterized by intrinsic randomness, reflected in the noise parameter β (Beeler, Daw, Frazier, & Zhuang, 2010). If anodal tDCS affects exploration and induces randomness in choices, participants will demonstrate increased shifting behavior (i.e., a tendency to change, rather than repeat a response to the same stimulus) and a decreased preference for symbols that are associated with the higher reward probability, reflected by higher β values.

2. Material and methods

2.1. Participants

Sixteen right-handed, healthy, native German-speaking participants took part in the study (mean age of 22.9 ± 2.2 years). In order to avoid menstrual cycle-dependent level changes of

the gonadal steroid hormones and their neurofunctional modulation of the reward system, only male participants were included in the study (Dreher et al., 2007; Jocham et al., 2011). All participants gave written informed consent. The study was conducted in accordance with the Declaration of Helsinki and it was approved by the local ethics committee.

2.2. *Stimulation*

A battery-driven CE-certified medical device (DC-Stimulator-Plus, NeuroConn GmbH, Ilmenau, Germany) was used to deliver the direct current to the head. Two rubber electrodes (3 × 3.5 cm) were covered with conductive paste and positioned on the scalp using the standard 64 channel 10/20 EEG caps in different sizes (small, medium and large; ANT-waveguard: https://www.ant-neuro.com/products/waveguard). The vertex was identified as the intercept of the half-way distance between the nasion and inion and the half-way distance between the pre-auricular points. The Cz electrode location of the EEG cap was placed over the vertex and this position was re-measured after the EEG cap was fitted to the participant's head. The electrode montage was based on electric field simulations using a realistic MR-derived finite element model (Opitz, Windhoff, Heidemann, Turner, & Thielscher, 2011) employing SimNibs (Windhoff, Opitz, & Thielscher, 2013). In total, 136 different electrode montages were simulated. Two circular-shaped electrodes with a diameter of 32 mm were used in each of the simulations. Electrodes were placed such that coverage of almost any location in the brain could be achieved in at least one montage. Out of all combinations, the electrode montage was selected that maximized absolute electric field strength in the dorsolateral prefrontal cortex, as determined based on anatomical landmarks (Mylius et al., 2013).

The anodal electrode was adjusted to the F3 location corresponding to the left dorsolateral prefrontal cortex (DLPFC) by moving it in the anterior and superior directions, such that the F3 location was in the lower-right corner of the vertically aligned electrode (see Fig. 1). The cathodal electrode was placed over the temporal cortex, where the middle point of the horizontally aligned electrode was exactly located over the T7 position.

Two stimulation protocols were used; one for the anodal tDCS and one for the sham tDCS condition. In the anodal tDCS condition, the stimulation was administered for 16 min, comprising a 30 sec fade-in/fade-out period and 15 min of stimulation at 1.0 mA intensity. In the sham tDCS condition, the stimulation protocol was identical to the anodal stimulation, except the stimulation duration, which lasted for only 30 sec (Ambrus et al., 2012). Although the stimulation duration in the real session was 4 min shorter than the learning phase, tDCS studies conducted on the motor cortex showed that the excitability changes following anodal or cathodal stimulation outlasts the stimulation duration by an hour, provided the stimulation duration is about 10 min or longer (Nitsche & Paulus, 2001).

2.3. *Experimental design*

The study employed a double-blind, placebo-controlled, repeated-measures experimental design. Subjects attended two separate experimental sessions, in which they completed two versions of the behavioral task (see later), which used two different sets of stimuli. Both the order of the version of the task presented first, as well as the order of the stimulation conditions (tDCS *vs* sham), were randomized for each participant and counterbalanced such that half of the participants

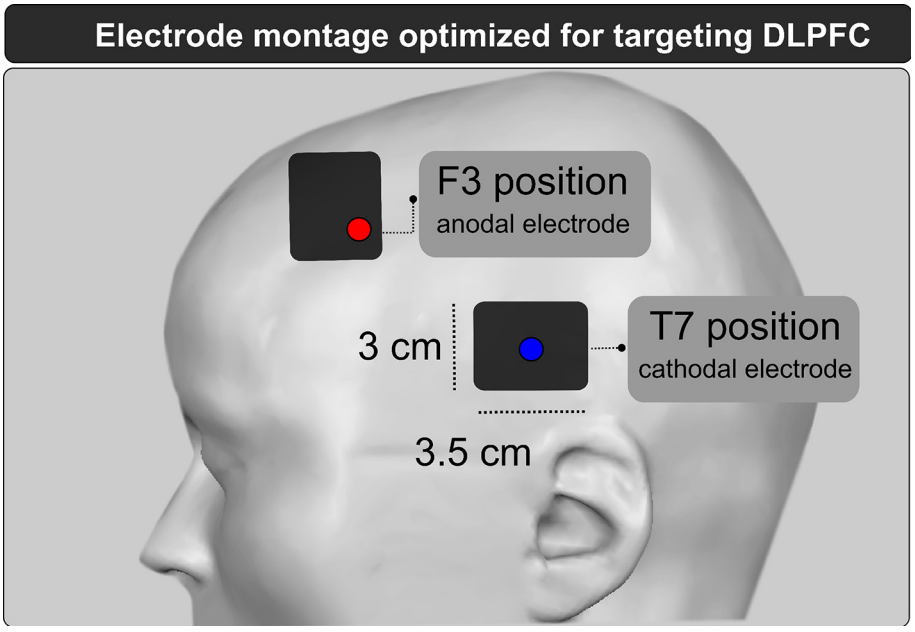

Fig. 1 – The electrode montage was optimized by using a realistic MR-derived finite element computational model in order to maximize the current flow in the DLPFC. The anodal electrode was shifted in the anterior and superior direction and aligned vertically, whereas the cathodal electrode was placed over the temporal cortex and aligned horizontally. DLPFC: dorsolateral prefrontal cortex.

started with anodal tDCS and half with sham stimulation, and half with task version 1 and half with task version 2.

In order to meet the criteria of a double-blind design, the "study mode" of the stimulator was used; that is, the two stimulation conditions, anodal tDCS *vs* sham tDCS, were randomly encoded as A or B modes, respectively. For each session, the investigator selected the stimulation mode according to a predetermined randomized list. The association between modes and stimulation conditions was unknown to the investigator who conducted the experiment. The study mode encoding was secured with a 5-digit code that was only accessible to the principal investigator (A.A.), who was not involved in the data collection and analysis process. The study mode was further advanced by the so called "pseudo-stimulation" mode, which resulted in identical display information (i.e., stimulation duration and impedance information) for the anodal tDCS and sham condition. In order to maintain the participant's unawareness of whether tDCS or sham stimulation was used, the standard "fade-in/short stimulation/fade-out" procedure was used in the sham condition (Ambrus et al., 2012), which is effective at 1.0 mA for up to 20 min (Gandiga, Hummel, & Cohen, 2006). In addition, participants filled out a short questionnaire after each session in order to discover whether adequate blinding was in fact maintained.

2.4. *RL and choice task*

The experimental task was adapted from Jocham et al. (2011), originally developed by Frank, Seeberger, and O'Reilly (2004). The task consisted of a learning and a test phase. In the learning phase (see Fig. 2), participants saw three pairs of symbols (labeled AB, CD and EF for reference), one pair at a time. Each symbol was probabilistically associated with a reward, which followed an inverse relationship within a pair (.8/.2, .7/.3 and .6/.4 for A/B, C/D and E/F, respectively). For example, symbol A was 80% correct and 20% incorrect, whereas symbol B was 20% correct and 80% incorrect. The task of the participants was to select the "better" symbol from the pair (i.e., the one with higher reward probability). The value of the reward probability was unknown to the participants. The learning phase consisted of 6 learning blocks, where each symbol pair was presented 20 times, resulting in 120 presentations of each symbol pair during the entire learning phase (360 presentation trials in total). For each symbol pair, the location of each symbol (left or right) was counterbalanced. The total trial duration was 3.3 sec. The sequence of events within a trial was similar to the study by Jocham et al. (2011): Each trial started with the presentation of a fixation cross for a duration of either 200, 500 and 800 msec (randomly chosen) followed by the symbol pair until a response was given. If no response was made after 1700 msec, the trial was canceled. Finally, the selected symbol was highlighted for 200 msec and feedback was displayed for 200 msec. The feedback was either a 'happy' or a 'sad' emoticon (i.e., a meta-communicative pictorial representation of facial expressions in Western style) for the positive or for the negative feedbacks, respectively. An additional 'confused face' emoticon was used in case of no answer.

In the transfer phase, participants were randomly presented with all possible combinations of the symbols (3 learned combinations plus 12 new combinations; 15 in total) repeated 12 times each. To prevent the participants from additional learning in the transfer phase, no feedback was provided at this time.

Before the start of the experiment, participants were given written instructions about the learning and the transfer phase

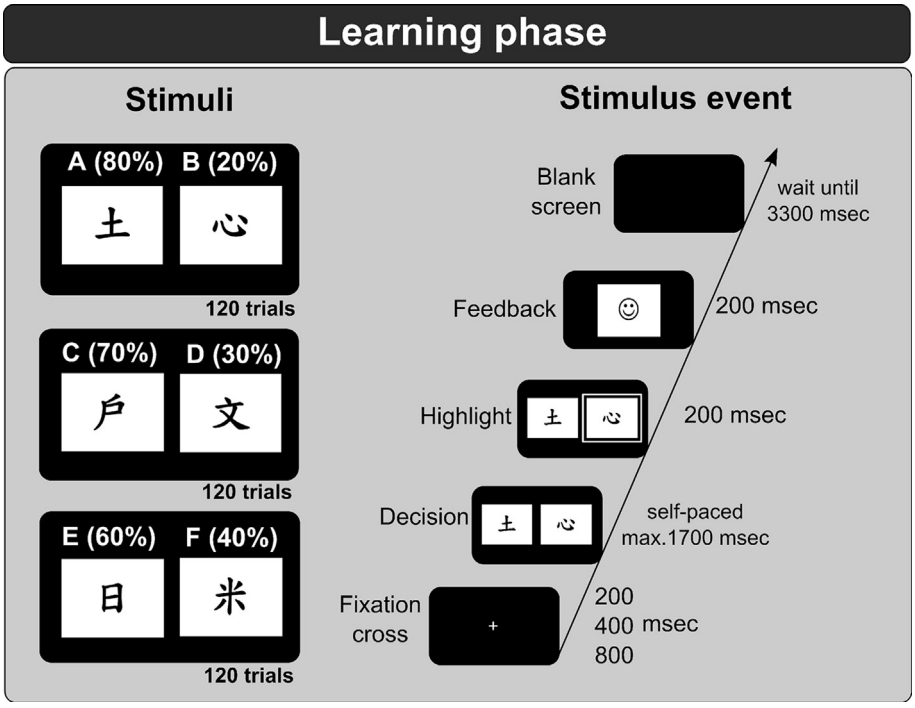

Fig. 2 – The learning phase consisted of 3 symbol pairs (Chinese characters; left), each of which probabilistically associated with the reward (see text for details).

(translated to German from Frank et al., 2007). Then, participants were asked to perform a training session of 13 trials, to ensure that they were comfortable with the experimental setup. Before the start of the experiment proper, participants were shown the 6 individual symbols twice, presented separately for 5 sec, to familiarize them with the stimuli.

2.5. Analysis of the RL and choice task

In the learning phase, our main interest focused on whether the participants' decisions following a positive or negative feedback were influenced by the stimulation. After receiving positive (win) or negative (lose) feedback to their decisions, participants could choose the previously chosen symbol (stay) or select the alternative symbol (shift) in the subsequent trial containing the same symbol pair (note that the three symbol pairs were randomly intermixed, each learning block containing 20 trials for each symbol pair). Therefore, win-stay behavior was defined when participants chose the same symbol after having received positive feedback on the previous trial in which the same pair had been presented. Win-shift behavior was defined when participants chose the alternative symbol even though they had previously received a positive feedback for this choice. Lose-stay and lose-shift behavior were described respectively as staying on the previous symbol or shifting the decision to the alternative symbol after having received a negative feedback. Each trial was assigned to one of these four categories, however, only win-stay and lose-stay behaviors were included into the analysis, as stay and shift behavior complement each other and add up to the total percentages rewarded.

In the transfer phase, we analyzed the accuracy according to the standard "choose A – avoid B" classification scheme (Frank et al., 2004). The percentages of correct choices were separately calculated for "choose A" trials (AC, AD, AE, and AF) and for "avoid B" trials (BC, BD, BE, and BF).

2.6. RL model

We used the RL model described in (Jocham et al., 2011). In brief, action values Q(A) through Q(F) for each item A through F were estimated based on the individual history of sequence of choices and the corresponding feedback experienced after each decision during the learning phase. The action values for each item were initialized to zero and were gradually updated using a modified version of the Rescorla–Wagner algorithm: $Q_{t+1}(i) = Q_t(i) + \alpha(r_t - Q_t(i))$ for $i \in \{A, B, C, D, E, F\}$ and t the trial number. The prediction error defined as $r_t - Q_t(i)$ is the difference between the actual and the expected feedback, where r_t represents the received reward on trial t (either 0 or 1 for negative and positive feedbacks, respectively). The learning rate parameter α reflects the impact of the prediction error; lower α-values indicate that the Q-values are integrated gradually over multiple-trials, whereas higher α-values reflect the recency effect (Frank et al., 2007). The probability of choosing one item over the other from a given pair was calculated using the soft-max rule. Thus, the probability of choosing A when AB was presented was calculated using the following rule: $P_t(A) = \exp(Q_t(A)/\beta)/[\exp(Q_t(A)/\beta) + \exp(Q_t(B)/\beta)]$. The parameter β reflects the participant's bias towards either exploration or exploitation: lower β-values indicate that the participant exploits the decision (i.e., choosing the better option with higher probability), whereas higher β-values reflect exploration (i.e., choosing the items more randomly) (Frank et al., 2007; Jocham et al., 2011). The maximum-likelihood (ML) parameter estimate (MLE) was selected by choosing parameters α, β that maximized the log-likelihood $l(d|\alpha, \beta) = \sum_{t=1}^{n} \log P_t(d_t)$, where $d_t \in \{A, B, C, D, E, F\}$ is the participant's decision on trial t. We maximized the parameters for each participant separately using the Nelder–Mead simplex algorithm (Nelder & Mead, 1965). The optimization algorithm was run 100 times for each subject from randomly generated starting points in the interval [0,1] for α and [0,3] for β to ensure uniqueness of the solution.

2.7. Statistical analysis

For each symbol and participant, we calculated the percentages (i.e., the percentage of choosing a symbol), accuracy and reaction time (RT). The percentage values for choosing the symbols were calculated relative to the total number of decisions corrected for the missing values. Accuracy was defined as "correct", when the statistically better symbol (i.e., A, C and E) was chosen from a given pair. Therefore, when participants received negative feedback after choosing the better option (e.g., A), the decision is still considered to be an accurate decision. Similarly, when participants chose B (e.g., suboptimal) and received positive feedback, the decision is considered incorrect. The probability to stay after positive, p(stay|win), or negative feedback, p(stay|lose), was calculated as the number of stays after positive or negative feedbacks, divided by the total number of positive or negative feedbacks.

A Shapiro–Wilk test was performed, which indicated that in the case of accuracy in the "choose A – avoid B" classification scheme data, the assumption of normality was violated (all $ps < .004$); therefore, an arcsine square root transformation was applied on these data such that the assumptions for the ensuring parametric tests were fulfilled (all $ps > .05$). Data were analyzed using repeated-measures Analyses of Variance (ANOVA). The assumption of sphericity was tested using the Mauchly test. If there was violation of sphericity, a Huynh-Feldt correction was applied that adjusts the p-values and degrees of freedom, and the latter values were rounded up to the first decimal place. Statistical analyses were conducted using a significance level of $p < .05$. If significant interactions occurred, post-hoc multiple comparisons were performed, where the p-value was always adjusted for multiple comparisons using the Bonferroni–Holm method (Holm, 1979).

In the learning phase, the within-subject factors were stimulation (2 levels: sham and anodal tDCS), block (6 levels: 1–6 blocks), block part (2 levels: first 40 decisions and last 80 decisions) and behavioral shifting (2 levels: win-stay and lose-stay). In the transfer phase, within-subject factors were stimulation (2 levels: sham and anodal tDCS), feedback learning (2 levels: choose A and avoid B classification) and symbols for the final Q-values (6 levels: for A, B, C, D, E and F symbols).

3. Results

3.1. Analysis of the learning phase

3.1.1. General accuracy and RT

During the course of the experiment, participants learned to reliably choose the statistically better symbol from the pairs in both stimulation conditions, evidenced by a significant increase in the arcsine square root transformed accuracy across blocks ($F_{3.4,51.1} = 15.417$, $p < .001$, $\eta_p^2 = .507$). The general learning performance was not influenced by tDCS. There was neither a main effect of stimulation nor a stimulation × block interaction (all $ps > .189$). The RT data revealed the same pattern of results, that is, the significant main effect of block indicates that participants became faster ($F_{5,75} = 49.519$, $p < .001$, $\eta_p^2 = .768$), but the lack of a significant main effect of stimulation and the stimulation × block interaction suggests that in general, RT was not modulated by the stimulation ($F < 1$) (see Table 1).

3.1.2. Analysis of the stay and shift behavior

When analyzing the amount of stay behavior separately for the symbol pairs, we found a significant main effect of stimulation for the AB pair ($F_{1,15} = 5.09$, $p = .04$, $\eta^2 = .07$) (see Fig. 3). Neither the main effect of stay type ($F_{1,15} = 2.07$, $p = .17$, $\eta^2 = .06$), nor the stimulation by stay type interaction ($F_{1,15} = .04$, $p = .84$, $\eta^2 = .0000$) reached a level of significance (see Table 2).

For the CD and the EF pair, the analysis revealed neither main effect for stimulation, nor a stimulation × behavioral shifting interaction (all $ps > .336$; all $Fs < 1$) (see Table 2 for descriptive statistics).

3.2. Analysis of the transfer phase

3.2.1. General accuracy and RT

General accuracy in the anodal tDCS and sham tDCS session was compared with paired *t*-tests, which revealed that participants performed significantly better when receiving sham tDCS compared to anodal tDCS [$t(15) = 2.887$, $p = .012$]: $M_{sham} = .82$, $SEM_{sham} = .02$; $M_{anodal} = .75$, $SEM_{anodal} = .03$. RT was not affected by the stimulation [$t(15) = 1.232$, $p = .237$].

Table 1 – Mean (untransformed) accuracy (ACC) and reaction time (RT) in the sham and the anodal tDCS sessions in the six learning blocks. SEM: standard error of mean.

Block number	Mean ACC ± SEM		Mean RT ± SEM (msec)	
	Sham	Anodal	Sham	Anodal
1	.69 ± .04	.69 ± .04	948.3 ± 41.0	951.0 ± 49.0
2	.79 ± .04	.75 ± .04	858.0 ± 45.9	862.1 ± 57.1
3	.82 ± .04	.81 ± .04	785.2 ± 48.7	816.8 ± 37.3
4	.85 ± .04	.81 ± .05	792.4 ± 38.5	778.0 ± 44.8
5	.85 ± .04	.80 ± .05	770.8 ± 47.2	751.8 ± 43.4
6	.88 ± .04	.83 ± .04	727.6 ± 42.3	734.0 ± 39.2
Mean ± SEM	.81 ± .04	.78 ± .04	813.7 ± 43.9	815.6 ± 45.1

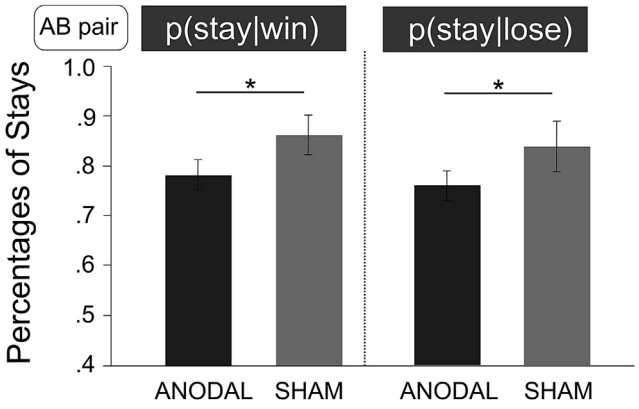

Fig. 3 – In the anodal tDCS session, participants stayed significantly less after receiving reward or punishment in the AB pair. Values represent mean percentages calculated for the six experimental blocks in the learning phase. Error bars represent standard error of mean. Asterisk indicates significant differences.

3.2.2. Analysis of the "choose A – avoid B" trial classification scheme

A repeated-measures ANOVA on the arcsine square root transformed accuracy measure revealed a significant main effect of stimulation ($F_{1,15} = 6.412$, $p = .023$, $\eta_p^2 = .299$) and significant stimulation × feedback learning ($F_{1,15} = 5.115$, $p = .039$, $\eta_p^2 = .254$). Post-hoc comparisons showed that in the anodal tDCS condition, participants performed less well on "choosing A" (calculated from the AC, AD, AE, and AF trials) when compared to the sham tDCS session [$t(15) = -3.017$, $p = .018$] (Fig. 4). No significant differences were found on the "avoid B" [$t(15) = -.691$, $p = .5$] (calculated from BC, BD, BE, and BF trials).

3.3. Analysis of the final Q-values at the end of the learning phase

The final Q-values showed a significant main effect ($F_{1,75} = 33.40$, $p = 2.0 \times 10^{-7}$, $\eta^2 = .338$) of symbol. However, none of these values were modulated by the stimulation ($F_{1,15} = .93$, $p = .35$, $\eta^2 = .017$) and stimulation by symbol interaction ($F_{1,75} = 1.19$, $p = .32$, $\eta^2 = .0008$). These results may indicate that the participants did not differ in the two stimulation conditions with regard to the ability to learn expected reward values for each symbol.

3.3.1. Analysis of the RL parameters at the end of the learning

In order to maximize the likelihood of each participant's trial-by-trial sequence individually for each participant, we fitted the two free parameters α and β to the data from the learning phase. Since the data were not normally distributed for either the α or β parameters even after the arcsine square root transformation procedure (Shapiro–Wilk test; all $ps < 4.7 \times 10^{-6}$), Wilcoxon signed rank test was performed on the untransformed values. The analysis revealed a significant difference in the β parameter [$Z(15) = 2.07$, $p = .04$] between the anodal (.16 ± .02) and sham (.11 ± .02) stimulation

Table 2 – The mean propensity to stay following positive or negative outcome calculated separately for the three different symbol pairs. SEM: standard error of mean.

	AB		CD		EF	
	Sham	Anodal	Sham	Anodal	Sham	Anodal
p(stay|win) ± SEM	.86 ± .03	.78 ± .04	.81 ± .03	.79 ± .04	.75 ± .04	.72 ± .04
p(stay|lose) ± SEM	.84 ± .03	.76 ± .05	.81 ± .04	.76 ± .05	.74 ± .04	.72 ± .04

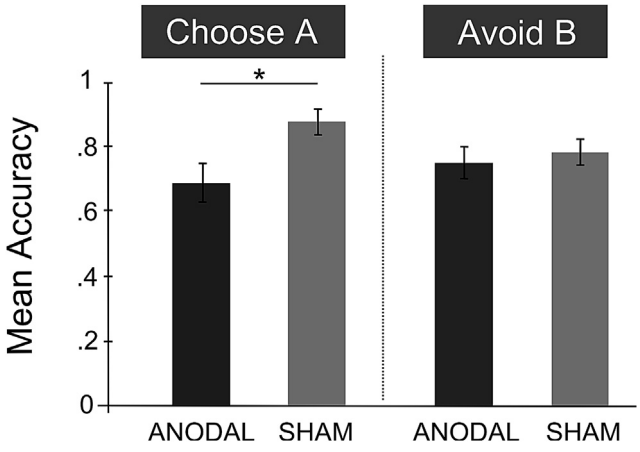

Fig. 4 – Participants performed significantly worse choosing A (calculated from AC, AD, AE, and AF pairs in the transfer phase) in the anodal tDCS compared to the sham tDCS condition, whereas avoid B (calculated from BC, BD, BE, and BF pairs in the transfer phase) performance was not influenced by the stimulation. Error bars represent standard error of mean. Asterisk indicates a significant difference.

conditions, but not in the α values [$Z(15) = -1.09$, $p = .3$] (anodal: .04 ± .01 and the sham .09 ± .06) (Fig. 5).

Because ML-based estimations sometimes have stability issues (parameter identifiability problems; Rutledge et al., 2009), we also ran a hierarchical Bayesian analysis as well as a model incorporating a perseverance parameter (Rutledge et al., 2009), for details on these analyses, see Supplemental Methods and Results.

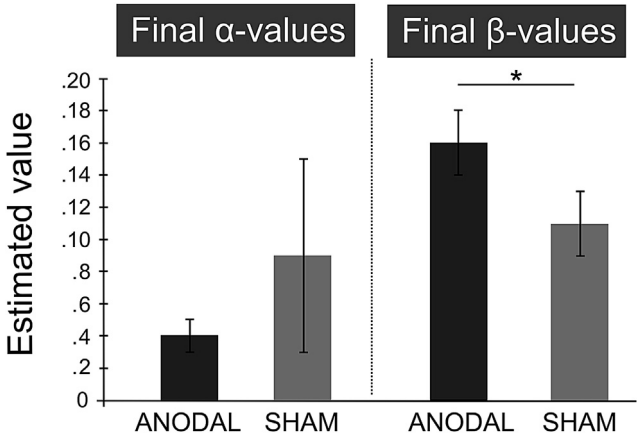

Fig. 5 – The mean final value for the β parameter was significantly higher in the anodal compared to the sham stimulation condition. The α parameter in the anodal and sham stimulation conditions was not significantly different. The vertical axis uses estimated values.

4. Discussion

The aim of the present study has been to investigate the role of the prefrontal high-level cognitive component of instrumental learning. Sixteen male participants were administered sham and anodal tDCS using a double-blind, sham-controlled, repeated-measures study design. Based on computer simulations of the electric current flow in the brain (Windhoff et al., 2013), we applied an electrode montage maximizing the current distribution over left DLPFC, a brain region playing a key role in high-level control of instrumental learning (Collins & Frank, 2012; Daw et al., 2006). During the learning phase, we observed a greater amount of behavioral shifting in the anodal tDCS as compared to the sham tDCS condition in the AB pair. In addition, fitting computational model parameters to the behavioral data also showed that participants were significantly impaired in exploiting the symbols associated with the higher reward probability as evidenced by increased β values (indicating increased randomness of choice) during learning and decreased accuracy for choosing the better option in the transfer phase in the anodal tDCS condition. Our findings complement previous computational, neuroimaging and genetic studies that investigated the role of the prefrontal component in instrumental learning (Collins & Frank, 2012; Daw et al., 2006; Frank et al., 2009; Frank et al., 2007) by means of a tDCS method.

In the anodal tDCS session, we observed more behavioral shifts (i.e., choosing the alternative symbol in next trial) relative to the sham tDCS session for the AB symbol pair during the learning phase. Anodal tDCS decreased the probability of win-stay and lose-stay behavior, in other words, participants shifted more often, both after positive and negative feedbacks. The pattern of these findings indicates that our participants showed increased shifting behavior, since they changed their decision both after positive and negative feedbacks. This was further supported by the increased β parameter in the anodal relative to the sham tDCS, which reflects increased randomness of choice.

A number of possible explanations can be provided for the observed pattern of results. TDCS might have affected instrumental learning through the WM. Previous experimental evidence suggests that the WM component provides the possibility for a flexible behavioral control of instrumental learning by actively maintaining the recently reinforced reward values (Collins & Frank, 2012). Genetic studies and computational modeling data indicate that in COMT Met individuals, the elevated PFC DA level may stabilize WM

representations and participants effectively use this ability to systematically adjust behavior on trial-to-trial basis following negative outcomes (Frank et al., 2007). Evidence also indicates that when participants performed in high and low WM-load conditions, COMT Met homozygotes performed better compared to Val carriers in the high (i.e., when the number of stimuli was higher), but not in the low WM-load conditions (i.e., when the number of stimuli was lower) (Collins & Frank, 2012). Further, a recent study also indicates left DLPFC involvement in the model-based control of decision-making via WM, as participants with low WM capacity (and possibly with low DA level) were impaired more after inhibitory continuous theta burst stimulation (cTBS) than individuals with high WM capacity (Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013). These findings congruently suggest that instrumental learning engages the prefrontal component via the WM.

An important aspect of the present data is that the behavioral effect was observed on the AB pair only even though it did manifest globally in the temperature parameter of the computational model. One possible explanation for this result would be that the low- and high-level components may be differently involved in instrumental learning based on reward probability. When the reward probability can be reliably separated within a pair (e.g., 80/20% as in the AB pair), the instrumental learning benefits more from WM system involvement by actively holding reinforcement outcomes in the WM. On the other hand, the active maintenance of the reinforcement history of the less reliable pairs might be beyond the capacity of the prefrontal system and therefore predominantly recruits the low-level components. In other words, prefrontal tDCS only interfered with the AB pair and not with the other pairs, since the reinforcement history of the AB pair may rely on the WM system, which was affected by tDCS. However, this account fails to explain the increased behavioral shift after both positive and negative feedback.

Although the present experiment employed an electrode montage that maximized the current distribution over the left DLPFC, we cannot claim that anodal tDCS impacted the WM component exclusively. The pattern of the present findings indicates that our participants showed increased shifting behavior, as they changed their decision after both positive and negative outcomes. This is contrary to a previous genetic study, where COMT Met carriers actively maintained recent negative reinforcement experiences and corrected their behavior on a trial-to-trial basis after negative outcomes (Frank et al., 2007). Further, previous studies investigating the left DLPFC found improved WM performance following anodal tDCS (Zaehle et al., 2011), which would lead to more adaptive trial-to-trial adjustment after negative outcomes and to an increased learning rate (α value), similar to COMT Met carriers. Interestingly, we observed only numerical differences in the learning rate parameter between the stimulation sessions by using the ML estimation technique (for the results of the hierarchical Bayesian modeling see Supplemental Fig. 2).

Alternatively, anodal tDCS may have affected the exploratory behavioral component of instrumental learning. Since we observed increased shifting behavior, i.e., an increased β parameter without a significant difference in the learning rate parameter, we conclude that a plausible explanation of the current findings is that anodal tDCS increases the randomness of choice. Current theory and experimental research on exploration suggest that exploration is accomplished by overriding an exploitative tendency of the striatal system by the prefrontal component (Daw et al., 2006). Intriguingly, the competition by mutual inhibition theory holds that decision-making is influenced by the relative degree of inhibition and excitation in the prefrontal cortex and consequently (Hunt et al., 2012), would partially depend on the balance between glutamatergic excitation and gamma-aminobutyric acid (GABA)ergic inhibition (Jocham, Hunt, Near, & Behrens, 2012). As anodal tDCS was shown to locally decrease the cortical GABA level (Stagg et al., 2009), we might speculate that our findings are the results of the decreased inhibitory GABA level in the frontal cortex, which may in turn increase choice randomness. Nevertheless, future neuroimaging experiments are needed to investigate this speculation directly.

Further, our findings are in line with a previous tDCS study, which applied anodal tDCS over the left DLPFC and observed suboptimal decision-making performance following anodal stimulation (Xue, Juan, Chang, Lu, & Dong, 2012). Although the experimental paradigm was somewhat different from that of the present experiment, the behavioral consequence of anodal tDCS was fundamentally equivalent in the two studies. Similar to our results, participants stayed less often after positive feedback, however, they also stayed more often after negative outcomes. Observing a brief and reversible decline in performance during anodal stimulation is not unprecedented in the literature (e.g., Ambrus et al., 2011), although it is commonly thought that anodal tDCS leads to an increase in neuronal excitability in the brain area underneath the electrode that should result in performance augmentation in a given task. However, it is hard to establish such a simple, linear and mechanistic relation between stimulation parameters, direction of the cortical excitability change and expected behavioral influence. In fact, this implicit assumption about the polarity effect of tDCS and its physiological consequences were recently questioned by a modeling study which showed that tDCS electric fields can de- and hyperpolarize within the same gyrus (Reato et al., 2013). Further, even in a homogeneous electric field, different types of neurons are differentially modulated based on their morphology and orientation (Radman et al., 2009). In line with the modeling studies, a meta-analysis on the effect of tDCS also supports the view that, compared to the motor domain, the polarity effect is less consistent on the cognitive domain (Jacobson et al., 2012).

An intriguing remaining question is whether tDCS increased randomness of choice by affecting the prefrontal or the striatal system. Positron emission tomography (PET) studies conducted on transcranial magnetic stimulation (TMS; another non-invasive brain stimulation technique) found that prefrontal TMS can have an impact on striatal DA release (Cho & Strafella, 2009; Ko et al., 2008; Strafella, Paus, Barrett, & Dagher, 2001). Unfortunately, the neurochemical effect of tDCS is still unexplored, and our experimental design did not allow us to directly answer this intriguing question as we lack PET/fMRI data. On purely speculative grounds however, we favor the view that the observed differences are mainly due to prefrontal rather than striatal changes. First, based on the present electrode montage, the computational model of electric current flow predicts that the electric field strengths in the

striatum are several orders of magnitude smaller than in the prefrontal cortex and are thus, very likely, not effective. In addition, the electric field estimation results are in line with previous fMRI findings showing local neurotransmitter change in the neocortex (Stagg et al., 2009). In addition, a previous functional neuroimaging study associated left DLPFC activity with maladaptive decision strategy, which was further influenced by anodal stimulation over the left DLPFC (Xue et al., 2012). Further, recent work indicates that the excitation-inhibition balance in the prefrontal cortex related to glutamatergic and GABAergic neurotransmitter balance (both of these are affected by tDCS) can itself explain value-based choice behavior variability (Jocham et al., 2012). Finally, although an animal study showed that tonic extracellular DA increase can influence exploration (i.e., the temperature or noise parameter) in rats (Beeler et al., 2010), the exact mechanism of how prefrontal tDCS could lead to altered striatal DA release in humans is unknown.

In summary, the present study tested the possible involvement of the prefrontal system in human instrumental learning by means of tDCS. DLPFC was targeted using an optimized montage based on computational electric field simulations (Windhoff et al., 2013). Stimulation with anodal tDCS increased behavioral shifting and decreased adaptive behavior compared to sham tDCS, possibly reflecting interference with the prefrontal system. The complexity of our results indicates that further studies are needed to investigate the interaction between the low-level and high-level components of instrumental learning.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This study was supported by the, DFG (PA 419/15-1) awarded to WP. MM was supported by a NWO Vidi grant awarded to Birte Forstmann. We thank Thomas Petrik for his help in adopting the task instructions to German and Titas Sengupta for her help in collecting participants and for the intriguing discussions on this topic. The authors wish to thank Christine Crozier who assisted in the proof-reading of the manuscript. We also thank the anonymous reviewers for their careful reading and their insightful suggestions.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cortex.2014.08.026.

REFERENCES

Ambrus, G. G., Al-Moyed, H., Chaieb, L., Sarp, L., Antal, A., & Paulus, W. (2012). The fade-in – short stimulation – fade out approach to sham tDCS – reliable at 1 mA for naive and experienced subjects, but not investigators. *Brain Stimulation, 5*(4), 499–504.

Ambrus, G. G., Zimmer, M., Kincses, Z. T., Harza, I., Kovács, G., Paulus, W., et al. (2011). The enhancement of cortical excitability over the DLPFC before and during training impairs categorization in the prototype distortion task. *Neuropsychologia, 49*, 1974–1980.

Beeler, J. A., Daw, N. D., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience, 4*, 170.

Chib, V. S., Yun, K., Takahashi, H., & Shimojo, S. (2013). Noninvasive remote activation of the ventral midbrain by transcranial direct current stimulation of prefrontal cortex. *Translational Psychiatry, 3*(6), e268. http://dx.doi.org/10.1038/tp.2013.44.

Cho, S. S., & Strafella, A. P. (2009). rTMS of the left dorsolateral prefrontal cortex modulates dopamine release in the ipsilateral anterior cingulate cortex and orbitofrontal cortex. *PloS One, 4*(8), e6725.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience, 35*(7), 1024–1035.

Datta, A., Elwassif, M., Battaglia, F., & Bikson, M. (2008). Transcranial current stimulation focality using disc and ring electrode configurations: FEM analysis. *Journal of Neural Engineering, 5*(2), 163–174.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876–879.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron, 36*(2), 285–298. http://dx.doi.org/10.1002/0471214426.pas0303.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*(2), 312–325.

Dreher, J.-C., Schmidt, P. J., Kohn, P., Furman, D., Rubinow, D., & Berman, K. F. (2007). Menstrual cycle phase modulates reward-related neural function in women. *Proceedings of the National Academy of Sciences, 104*(7), 2465–2470.

Faria, P., Hallett, M., & Miranda, P. C. (2011). A finite element analysis of the effect of electrode area and inter-electrode distance on the spatial distribution of the current density in tDCS. *Journal of Neural Engineering, 8*(6), 066017. http://dx.doi.org/10.1088/1741-2560/8/6/066017.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12*(8), 1062–1068.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences, 104*(41), 16311–16316.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science, 306*(5703), 1940–1943.

Gandiga, P. C., Hummel, F. C., & Cohen, L. G. (2006). Transcranial DC stimulation (tDCS): a tool for double-blind sham-controlled clinical studies in brain stimulation. *Clinical Neurophysiology, 117*(4), 845–850.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience, 15*(3), 470–476.

Jacobson, L., Koslowsky, M., & Lavidor, M. (2012). tDCS polarity effects in motor and cognitive domains: a meta-analytical review. *Experimental Brain Research, 216*(1), 1–10.

Jocham, G., Hunt, L. T., Near, J., & Behrens, T. E. J. (2012). A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nature Neuroscience, 15*(7), 960–961.

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *The Journal of Neuroscience, 31*(5), 1606–1613.

Ko, J. H., Monchi, O., Ptito, A., Bloomfield, P., Houle, S., & Strafella, A. P. (2008). Theta burst stimulation-induced inhibition of dorsolateral prefrontal cortex reveals hemispheric asymmetry in striatal dopamine release during a set-shifting task–a TMS–[11C] raclopride PET study. *European Journal of Neuroscience, 28*(10), 2147–2155.

Mylius, V., Ayache, S. S., Ahdab, R., Farhat, W. H., Zouari, H. G., Belke, M., et al. (2013). Definition of DLPFC and M1 according to anatomical landmarks for navigated brain stimulation: inter-rater reliability, accuracy, and influence of gender and age. *NeuroImage, 78*, 224–232. http://dx.doi.org/10.1016/j.neuroimage.2013.03.061.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308–313. http://dx.doi.org/10.1093/comjnl/7.4.308.

Nitsche, M. A., & Paulus, W. (2000). Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. *The Journal of Physiology, 527*(3), 633.

Nitsche, M. A., & Paulus, W. (2001). Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. *Neurology, 57*(10), 1899–1901.

Opitz, A., Windhoff, M., Heidemann, R. M., Turner, R., & Thielscher, A. (2011). How the brain tissue shapes the electric field induced by transcranial magnetic stimulation. *NeuroImage, 58*(3), 849–859. http://dx.doi.org/10.1016/j.neuroimage.2011.06.069.

Radman, T., Ramos, R. L., Brumberg, J. C., & Bikson, M. (2009). Role of cortical cell type and morphology in subthreshold and suprathreshold uniform electric field stimulation in vitro. *Brain Stimulation, 2*(4), 215–228. http://dx.doi.org/10.1016/j.brs.2009.03.007.

Reato, D., Gasca, F., Datta, A., Bikson, M., Marshall, L., & Parra, L. C. (2013). Transcranial electrical stimulation accelerates human sleep homeostasis. *PLoS Computational Biology, 9*(2), e1002898. http://dx.doi.org/10.1371/journal.pcbi.1002898.

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *The Journal of Neuroscience, 29*(48), 15104–15114.

Samson, R. D., Frank, M. J., & Fellous, J.-M. (2010). Computational models of reinforcement learning: the role of dopamine as a reward signal. *Cognitive Neurodynamics, 4*(2), 91–105.

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80*(4), 914–919.

Stagg, C. J., Best, J. G., Stephenson, M. C., O'Shea, J., Wylezinska, M., Kincses, Z. T., et al. (2009). Polarity-sensitive modulation of cortical neurotransmitters by transcranial stimulation. *The Journal of Neuroscience, 29*(16), 5202–5206.

Strafella, A. P., Paus, T., Barrett, J., & Dagher, A. (2001). Repetitive transcranial magnetic stimulation of the human prefrontal cortex induces dopamine release in the caudate nucleus. *The Journal of Neuroscience, 21*(15), RC157.

Windhoff, M., Opitz, A., & Thielscher, A. (2013). Electric field calculations in brain stimulation based on finite elements: an optimized processing pipeline for the generation and usage of accurate individual head models. *Human Brain Mapping, 34*(4), 923–935.

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75*(3), 418–424.

Xue, G., Juan, C. H., Chang, C. F., Lu, Z. L., & Dong, Q. (2012). Lateral prefrontal cortex contributes to maladaptive decisions. *Proceedings of the National Academy of Sciences, 109*(12), 4401–4406.

Zaehle, T., Sandmann, P., Thorne, J. D., Jäncke, L., & Herrmann, C. S. (2011). Transcranial direct current stimulation of the prefrontal cortex modulates working memory performance: combined behavioural and electrophysiological evidence. *BMC Neuroscience, 12*(1), 2.