Examining the effects of transcranial direct current stimulation on human episodic memory with machine learning

Aleksandra Petrovskaya[1,2☯], Bogdan Kirillov[3*☯], Anastasiya Asmolova[1,2], Giulia Galli[4], Matteo Feurra[1,2‡], Angela Medvedeva[5‡],

1 Psychology Department, National Research University Higher School of Economics, 109316, Moscow, Russian Federation
2 Institute of Cognitive Neuroscience, National Research University Higher School of Economics, 109316, Moscow, Russian Federation
3 Center of Life Sciences, Skolkovo Institute of Science and Technology, 121205, Skolkovo, Russia
4 Department of Psychology, Kingston University, Penrhyn Road, Kingston Upon Thames KT1 2EE, United Kingdom
5 Vivian L. Smith Department of Neurosurgery, University of Texas Medical School at Houston, Houston, Texas 77030, United States

☯These authors contributed equally to this work.
‡These authors share senior authorship.
* Bogdan.Kirillov@skoltech.ru

Abstract

We aimed to replicate a published effect of transcranial direct-current stimulation (tDCS)-induced recognition enhancement over the human ventrolateral prefrontal cortex [1] and analyse the data with machine learning. We investigated effects over an adjacent region, the dorsolateral PFC. We found weak or absent effects over the VLPFC and DLPFC. We conducted machine learning studies to examine the effects of semantic and phonetic features on memorization, which revealed no effect of VLPFC tDCS on the original dataset or the current data. The highest contributing factor to memory performance was individual differences in memory not explained by word features, tDCS group, or sample size, while semantic, phonetic, and orthographic word characteristics did not contribute significantly. To our knowledge, this is the first tDCS study to investigate cognitive effects with machine learning, and future studies may benefit from studying physiological as well as cognitive effects with data-driven approaches and computational models.

Author summary

Non-Invasive Brain Stimulation techniques (in our case, transcranial direct current stimulation) are widely used among neuroscientists to map cognitive processes, for example, memory, decision making, emotional processing. In many cases, the sample size is limited, or methods applied for the data analysis are questionable. Here we propose a replication study aiming at confirmation of memory enhancement as a result of the application of anodal transcranial direct stimulation over the ventrolateral prefrontal cortex. Additionally, the involvement of the dorsolateral prefrontal cortex into episodic memory was analyzed. We used a larger sample size and applied different data analysis methods, including machine learning techniques. Surprisingly, we did not

replicate the results of the original study, and we found weak effects of memory impairment after stimulating the DLPFC. However, what is most remarkable is that we have observed no significant effect of tDCS involvement on memory performance. Machine learning methods revealed no effect of linguistic factors on tDCS effect for both the original study and the replication at the level of individual trials and participants. Our findings highlight the importance of considering individual differences and data on the level of a single trial; in our data, participants' memory responses resembled guessing behavior when recognition performance was measured by AUROC and highlight the need for modifications in the memory test and the use of other performance measures.

1 Introduction

Transcranial direct-current stimulation (tDCS), a method of safely and non-invasively delivering a weak electric current through the cortex, has been gaining increasing attention [2] as a tool for studying and possibly enhancing episodic memory [3]. Anodal tDCS involves increasing cortical excitability in a target region such as the dorsolateral prefrontal cortex (DLPFC), which may in turn facilitate or enhance memory performance by decreasing reaction time and/or increasing memory accuracy. However, there remains well-founded skepticism about widespread applications of tDCS, partly because of the lack of knowledge about mechanisms of action [4] and difficulty in replicating results [5].

Most tDCS studies in episodic memory have targeted the DLPFC (see [6] for a review) because of its role in selective attention [7], strategic retrieval [8], and other executive functions involved in episodic memory [9]. In line with the hemispheric encoding/retrieval asymmetry model [10], previous studies with tDCS show functional asymmetry in the DLPFC, with stimulation over the left DLPFC affecting encoding and stimulation over the right DLPFC affecting retrieval [11–13]. However, memory enhancements due to atDCS have been inconsistent over the DLPFC, in both individual studies [9, 11, 12, 14, 15] and meta-analyses [16–18]. Most studies over the DLPFC examined effects when atDCS was delivered predominantly during encoding. At least two studies examined purely offline encoding effects: [19] delivered atDCS over F3 for 15 minutes before encoding and found no effects on accuracy or reaction time, although atDCS over the contralateral hemisphere (F4) led to faster reaction times. Lu and colleagues [20] delivered atDCS over a more lateral site (FC5) for 20 minutes before encoding and found an increase in correct memory responses for previously-presented items.

To the authors' knowledge, no study to date has systematically examined effects of timing of administration on the DLPFC, comparing online vs offline effects at encoding or retrieval. However, a previous study by the authors showed significant differences in online and offline effects over an adjacent site, the left VLPFC [1]. The authors administered atDCS before the study phase (offline encoding) or during the study task (online encoding) that involved intentionally memorizing each presented word (Experiment 1). As expected given support from fMRI and TMS studies, the left VLPFC seemed to be strongly modulated by atDCS at encoding, specifically online but not offline encoding.

The VLPFC and DLPFC are thought to play functionally distinct roles in long-term verbal memory, and the left hemisphere in the VLPFC appears selectively engaged in verbal but not non-verbal material. Specifically, the VLPFC may be more involved in encoding for individual items, while the DLPFC is more engaged in associative or relational encoding [21–24]. Moreover, activation in the DLPFC may predict long-term memory success through DLPFC involvement in domain-general working-memory or executive processes such as mental manipulation of information (e.g. visualizing rotating

objects; [25]) and applying a specific strategy (e.g. for retrieval of information; [26]). In one tDCS study, [27] found that cathodal stimulation over the left DLPFC (F3) in a cued-word-completion learning task where errors were evoked through guessing the wrong completion lowered memory accuracy in contrast to a non-error condition or anodal stimulation in either condition (no effect). They concluded that the DLPFC was only modulated when its processing demands were higher for conflicting information. In contrast, the VLPFC appears to be engaged in maintaining information in working memory [28] and processing semantic information including accessing lexical identity and connecting words to other words in the semantic network [29].

The current study involved administering atDCS over the left DLPFC offline or online encoding to identify the effective time of administration and confirm the causal role of the DLPFC in encoding processes. Participants were assigned to a Sham group (over the DLPFC) or one of three tDCS groups: DLPFC Online, DLPFC Offline, and VLPFC Online. Previous research suggests that atDCS could potentially enhance memory performance when delivered online or offline encoding over the DLPFC [20, 30–33], so we predicted higher memory performance for either condition (DLPFC Online or DLPFC Offline). In addition, the study provides a novel comparison of atDCS effects during encoding over the VLPFC vs DLPFC. We predicted a successful replication of Experiment 1 from [1] with higher memory accuracy for the VLPFC Online group compared to Sham, but we also predicted a larger effect size for VLPFC Online than DLPFC Online or DLPFC Offline based on previous rTMS studies comparing memory disruption in the VLPFC vs DLPFC. For example, [34] found that stimulation over the VLPFC during encoding led to a greater disruption in memory performance than over the DLPFC, suggesting that the VLPFC may play a more important role in encoding processes. Finally, there remains a lack of knowledge about the mechanisms of atDCS on verbal memory and whether atDCS effects can be influenced by word characteristics, specifically semantic (meaning), orthographic (letters), and phonological (sounds). We also aimed to examine how the tDCS effects interacted with language and linguistic factors, since the replication study was conducted in Russian and the original study was conducted in English. Thus, data were applied from a previously-published experiment ([1]; Experiment 1) and the current experiment from the Online Encoding (VLPFC) group to an AUROC analysis. After comparisons of group means in reaction time and accuracy on the recognition test, an ML regression algorithm was applied to infer the quality of recognition using semantic and phonological features of words. Semantic and phonological features were added to the model and predictive capability was assessed. We predicted that the model would be able to distinguish between words that are hard and easy to remember in Russian and English speakers. Thus, machine learning algorithms were applied in order to reveal the impact of factors including semantics, phonetics and individual variance to episodic memory performance as well as the interaction with tDCS.

2 Methods and Materials

Guidelines can be included for standard research article sections, such as this one. In line with a replication study, we followed the same procedure as in Experiment 1 of [1] with few exceptions (translation of materials to Russian and comparison of DLPFC and VLPFC as stimulation sites). Broadly, participants memorized words presented individually on a screen while undergoing tDCS (before or during this phase), and 24 hours later they performed a recognition test. The study and test stimuli were translated based on the first word meaning in the vocabulary entry, and the translation achieved relatively matched frequency (M English = 24.47; M Russian=40.38; Lyashevskaya Sharov, 2009) and number of letters (M English =6.17; M Russian=6.19).

Specifically, we applied the same tDCS settings (2 mA anodal tDCS and 30 s Sham tDCS) over the DLPFC (all conditions: offline, sham, and online) in addition to the VLPFC (online encoding only). For all groups, the anode was placed over the left VLPFC (F7) or the left DLPFC (F3), and the cathode was placed over the ipsilateral (left) shoulder at both sessions (study and test 24 hours later). However, stimulation was delivered only on the first day and at different sites (DLPFC or VLPFC) and times (offline or online encoding) depending on the group. Stimulation was delivered via a battery-powered, constant-current stimulator (Brainstim, EMS, Bologna, Italy) through 5 x 7 cm2 electrodes. In addition, participants were asked to describe the sensations that they felt during the stimulation and indicate whether they believed they received real or placebo stimulation. Participants were single-blinded, and the questionnaire indicated that blinding was successful: 80% believed that they received real stimulation. Data were analyzed in SPSS (version 24; IBM, Armonk, New York) and machine learning studies were conducted in Python (version 3.7).

2.1 Participants

Participants were randomly assigned to one out of four groups: VLPFC Online, DLPFC Online, DLPFC Offline and Sham. Based on a power analysis to detect a large effect size [1], $d = 1.29; \alpha = 0.05, 1 - \beta = .95$), we aimed to recruit 31 participants per group (124 total), but we analyzed data from 97 participants because of exclusions (see below). Participants (female = 63, male = 49, M age = 20.51, SD=2.89) were native Russian speakers with normal or corrected-to-normal vision and no history of neurological or psychiatric illness. Bonferroni-corrected pairwise comparisons of ae between pairs of groups showed a significant difference in age between VLPFC Online (M = 21.76, SD = 3.54) and DLPFC Online (M = 19.21, SD = 1.10), p = .014. However, there were no significant differences in age between the other groups, ps>.062.

Data from 12 participants were excluded from analysis because 1) the participant was feeling unwell (2 participants : DLPFC Sham and VLPFC Online) 2) there were technical issues (6 participants: two in DLPFC Online, two in VLPFC Online, and two in DLPFC Sham) or 3) the participant was left-handed (one participant: DLPFC Online) and 4) there was an experimenter error (3 participants: one in DLPFC Sham and two in VLPFC Online). The exclusions resulted in 26 in the VLPFC Online group, 31 in the DLPFC Offline group, 28 in the DLPFC Online group, and 27 in the Sham group. Moreover, fifteen outliers were excluded (see data analysis for criteria), leaving a final sample size of 25 for VLPFC Online (one outlier), 25 for DLPFC Offline (six outliers), 23 for DLPFC Online (five outliers), and 23 for DLPFC Sham (four outliers). For all Machine Learning analysis all DLPFC Offline participants except one were used (30 participants).

The study was approved by the ethics committee of National Research University Higher School of Economics (Moscow, Russia) and followed the corresponding ethical guidelines. All participants provided written informed consent and were given monetary compensation (500 rubles) for their time.

2.2 Procedure and Experimental Design

Each participant, regardless of group assignment, came to the laboratory twice within an 24 hour-interval: on the first day participants memorized the verbal stimuli one word at a time through pleasantness judgements (pleasant or unpleasant), and on the second day they performed an old/new recognition memory task. Stimulation was only delivered on the first day, but electrodes were placed on the head on the second day without stimulation (to replicate the method of [1] exactly). For all groups, stimulation was delivered on the first day only. For DLPFC Sham, stimulation was delivered for 30

seconds before the stimulator was turned off, and the start of the stimulation corresponded to the start of the reading task for half the Sham participants and the start of the memory task (study phase) for the other half. Stimulation was applied for the entire duration of the reading task for the DLPFC Offline group and for the entire duration of the memory task for the VLPFC Online and DLPFC Online groups.

2.3 Data analysis

One-way ANOVAs were conducted for each measure of recognition accuracy (Snodgrass Corwin, 1988): discrimination index (Pr: combined index involving hits and false alarms), proportion of hits (correctly identifying an "old" item as studied), and proportion of false alarms (incorrectly identifying a new item as studied). In addition, one-way ANOVAs were conducted for average reaction times and reaction times for hits and false alarms separately. Finally, a one-way ANOVA was conducted on response bias, an index involving hits and false alarms (higher values indicate a more conservative pattern of responding to old items). Significant effects were followed up with Bonferroni-corrected planned contrasts between each stimulation group and Sham (one-tailed). Outliers were excluded based on two standard deviations from the mean on any of the following dependent measures: discrimination index Pr, proportion of hits, proportion of false alarms, and reaction times for hits or false alarms.

2.4 Area Under the Receiver Operating Characteristic Curve as a measure of predictive performance for episodic memory changes

For the second analysis including data from the [1], we included all 17 participants from the Sham group and 17 from the Online tDCS group, resulting in the same mean age with no significant differences between the groups. From the current study, we only included data from the DLPFC Sham (N = 23) and VLPFC Online groups (N = 25) with no significant differences in age between groups (Bonferroni-corrected paired comparisons), ps>.852. We implemented a 2 x 2 design with the first factor as language (English or Russian) and the second factor as tDCS group (VLPFC or Sham).

Applying Area Under the Receiver Operating Characteristic Curve (AUROC) was appropriate given that the task involved binary classification (coded as 1 if the word was presented at study and 0 if unpresented), which fits any ML approach to measuring predictive performance of classification models (Bradley, 1997). In ML Classification models, true positives represent the model predicting the positive class correctly and false positives represent incorrect prediction of positive class.

AUROC 1) provides information about true and false positives in a single measure 2) shows the source of model error, with larger values indicating that the model predicts better than chance (greater true positives) and 0.5 indicating the opposite and 3) does not assume a normal distribution and is robust to unequal sample sizes (whereas average accuracy measures would overestimate performance). Thus, it is the most suitable measure for the current data.

We included a trial-based approach, computing AUROC for each word based on the entire sample's responses for that word, and a participant-based approach, computing AUROC for each participant based on responses (old or new) for all words presented to that participant.

AUROC values can show whether tDCS is effective in enhancing memory performance, and the subsequent ML analyses using AUROC can show the effects of various factors (semantic, phonetic, orthographic word characteristics) on recognition of individual words. We attributed any differences in memory performance not explained

by differences in sample size, linguistic characteristics, or group (tDCS vs Sham) to individual differences in memory function.

If stimulation enhances episodic memory performance, we expect a significant increase or decrease in AUROC (both over words and over participants in both VLPFC groups compared to Sham). If tDCS impairs memory performance, we expect values close to 0.5, which reflects that participant response decisions (old or new) approached chance.

If tDCS has no effect, there would be no difference in AUROC values across words or AUROC distributions between VLPFC and Sham. Since tDCS enhanced episodic memory in the original experiment and we expect a successful replication, we expected AUROC values for VLPFC to be significantly different from chance, with significantly different AUROC values and distributions compared to Sham. Moreover, tDCS significantly increased reaction time in [1], so we predicted a significant difference between VLPFC and Sham in reaction time. Since we did not assume a normal distribution, we applied the two-sample Kolmogorov-Smirnov test for accuracy and reaction time. We have also computed median test for differences in median AUROC.

2.5 Extraction of semantic, phonological and orthographic determinants of episodic memory performance

To investigate the effect of semantic and phonetic features on memorization and the interaction with tDCS, we extracted words with significantly more true positive rates (reflecting hits) and false positive rates (reflecting false alarms) in each group (VLPFC and Sham). We examined the top ten words from each category (see S1 Appendix). A preliminary descriptive analysis of the highest and lowest AUROC-words showed no association between individual variance and memory accuracy in VLPFC and Sham groups. Therefore, we do not report these results further.

Moreover, we used word embeddings, which approximate the semantic similarities between words. In line with the trial-based and participant-based approaches, we conducted two kinds of ML studies: 1) Participant-independent (trial-based) analysis tests the success of a model that predicts the AUROC for each word using either word embeddings or letter-based one-hot encoding and 2) Participant-specific (participant-based) analysis tests the success of a model that predicts whether a participant was able to recall the word given either word embeddings or letter-based one-hot encoding for each word for individual participants. Word embeddings were applied from FastText [35], a library developed in Facebook that incorporates semantic information from each word as well as subwords contained within to embed vectors. The pretrained embedding was aligned following [36] so that English and Russian vectors could lie in the same vector space and we could use it to train a model that uses both English and Russian vectors to check whether there are any language-independent semantic determinants. To artificially enlarge the sample size for our models we use the augmentation approach described in [37] (see S2 Appendix for details).

For prediction, we applied TPOT [38], an automated Machine Learning library that enables searching for the best classification and regression model using Evolutionary Algorithms, and AutoPyTorch [39], an automating Machine Learning library for the PyTorch [40] deep learning framework. Pipelines found by TPOT or AutoPyTorch are usually better than those found by grid search or manual construction. TPOT was used for experiments with word vectors, and AutoPyTorch was used for experiments with one-hot encoded word images.

We applied the following parameters: for TPOT, default ones for evolutionary algorithm; 20 for number of generations and 5 for population size; 10 folds for cross-validation; mean absolute error as the scoring function, and “TPOT light” as the

configuration dictionary, for AutoPyTorch, default validation setup, "tiny-cs" config for participant-independent trials and "medium-cs" config for participant-dependent trials.

Prior to training, for each experiment, we have randomly extracted 10% of the dataset and used it as a hold-out set for test. We trained our AutoML on the rest 90% with the respective cross-validation schemes and we examine the generalization ability of a model by assessing the difference between the quality metrics on training and hold-out sets. The model could perform very well on training set, but show lack of predictive ability on the hold-out. Such an outcome would be a sign of poor generalization ability which would show that based on this data we can not infer the existence of learnable connection between features and a label.

3 Results

3.1 ANOVA

Accuracy was significantly above chance for all groups, $ts > 3.88, ps < .005$. Levene's test ($Fs > 2.84, ps < .042$) and frequency distributions suggested that most data did not follow a normal distribution. Although non-parametric statistics may be more appropriate than an ANOVA, to enable comparison with the original effect sizes, corrected parametric statistics (Brown-Forsythe's F statistic) are reported in text for corresponding dependent measures. There were no significant differences between groups in recognition accuracy $F(3, 79.78) = 2.37, p = .076, = .082$, response bias $F(3, 93) = 2.15, p = .100, \eta_p^2 = .065$, or average reaction time $F(3, 74.26) = 0.70, p = .556, \eta_p^2 = .028$. There was also no significant difference in the proportion of hits $F(3, 76.88) = 1.37, p = .259, \eta_p^2 = .051$, or associated reaction times $F(3, 75.56) = 0.49, p = .694, \eta_p^2 = .019$. There was a significant difference in proportion of false alarms $F(3, 93) = 2.85, p = .042$ but not associated reaction times $F(3, 70.82) = 0.91, p = .439, \eta_p^2 = .037$. Planned contrasts for false alarms revealed significant differences between VLPFC tDCS and Sham $t(93) = 2.78$ and between Offline DLPFC tDCS and Sham $t(93) = 2.14, p = .018$, with higher false alarm rates for VLPFC tDCS and Offline DLPFC tDCS. There were no significant differences between Online tDCS and Sham $t(93) = 1.36, p = .178$. See Table 2 for the mean accuracy for each group.

Table 1. Means and standard deviations for memory accuracy across groups.

	Online tDCS M	Online tDCS SD	Offline tDCS M	Offline tDCS SD	Sham M	Sham SD
DLPFC **N**	**24**		**25**		**23**	
Discrimination Pr	0.15	0.16	0.07	0.12	0.16	0.18
Br response bias	0.61	0.14	0.60	0.12	0.56	0.17
Pr hits	0.68	0.10	0.63	0.12	0.63	0.16
Pr false alarms	0.53	0.18	0.57	0.14	0.47	0.18
VLPFC **N**	**25**					
Discrimination Pr	0.09	0.11				
Br response bias	0.66	0.12				
Pr hits	0.69	0.11				
Pr false alarms	0.59	0.12				

Table 2. Means and standard deviations for reaction time accuracy across groups.

	Online tDCS M	Online tDCS SD	Offline tDCS M	Offline tDCS SD	Sham M	Sham SD
DLPFC **N**	24		25		23	
Average RT	501.49	144.91	503.00	103.62	535.72	156.16
RT hits	504.82	151.20	505.85	104.71	532.52	142.72
RT false alarms	498.17	139.84	500.16	103.66	538.91	171.43
VLPFC **N**	25					
Discrimination Pr	484.00	82.58				
Br response bias	490.18	79.85				
Pr hits	477.82	86.10				
Pr false alarms	484.00	82.58				

3.2 AUROC Analysis and ML models

The range of performance as measured by AUROC (0.38-0.62; see Figure 1) indicates that participant decisions were close to random choice, since an AUC coefficient of 0.75 or higher reliably reflects accurate performance. The VLPFC group shows a similar distribution of individual AUROC coefficients to the Sham group, indicating that there was no effect of tDCS on memorization and perhaps other factors (word characteristics, individual differences, error) contributed more highly.

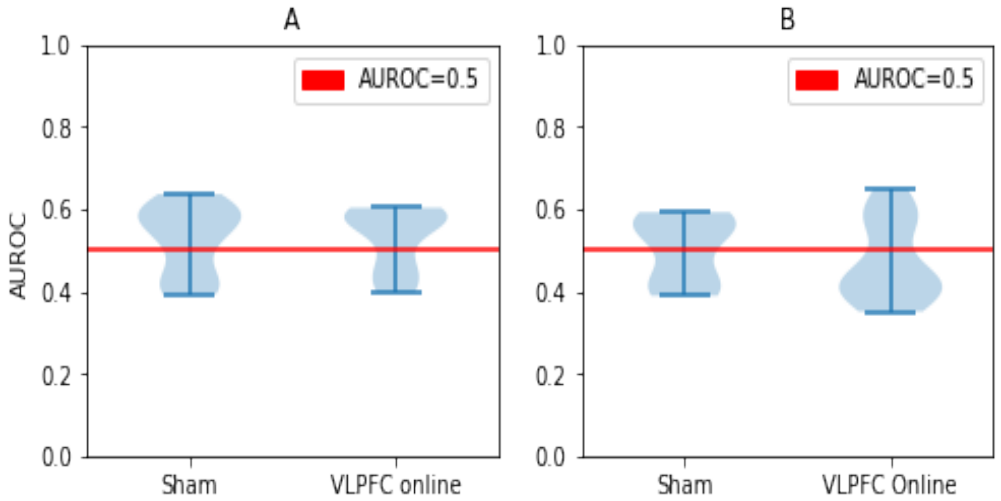

Fig 1. Violin plots of AUC distributions for Russian vs English participants in each group. Panel A shows the Russian sample (p-value of median test is 0.267) and Panel B shows the English sample (p-value of median test is 0.17).

Although the distributions of AUROC coefficients for individual words appeared to vary depending on the sample (English or Russian) and group (VLPFC or Sham; see Figure 2), with the English VLPFC group showing more variance than either of the Sham groups, there was no significant difference in AUROC distributions between English or Russian words or experimental groups, suggesting that the greater variance in the English group reflected smaller sample size.

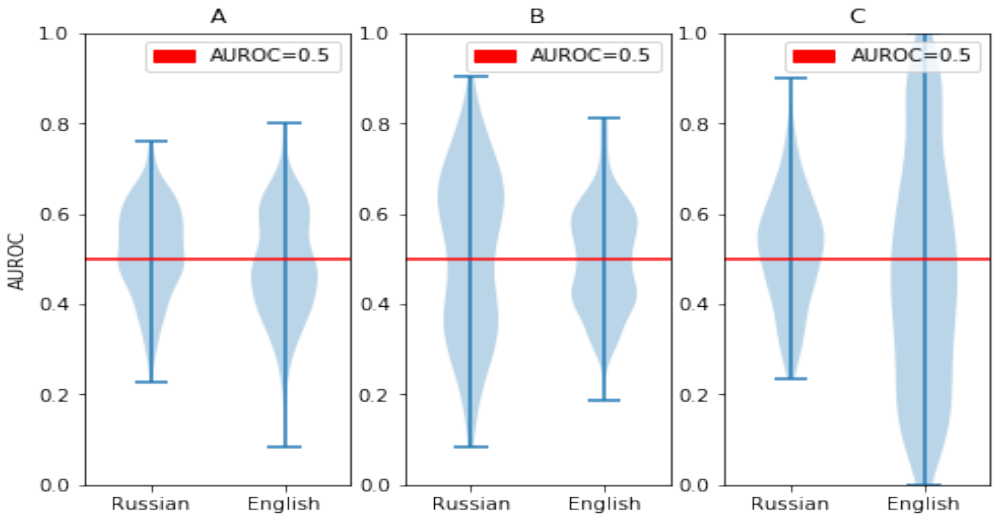

Fig 2. Violin plots showing AUROC distributions for Russian vs English words in each group. A – all groups (p-value of median test is 0.004), B – sham (p-value of median test is 0.274), C – VLPFC online (p-value of median test is 0.001).

We found significant differences in distributions for reaction time between Russian and English words as well as significant differences in distributions between sham and VLPFC online across both samples, $p<.001$, but no significant difference in median AUROC (p-values are either way larger than 0.001 or borderline – about 0.001-0.004), suggesting that as for the accuracy data, differences in distribution reflected differences in sample size.

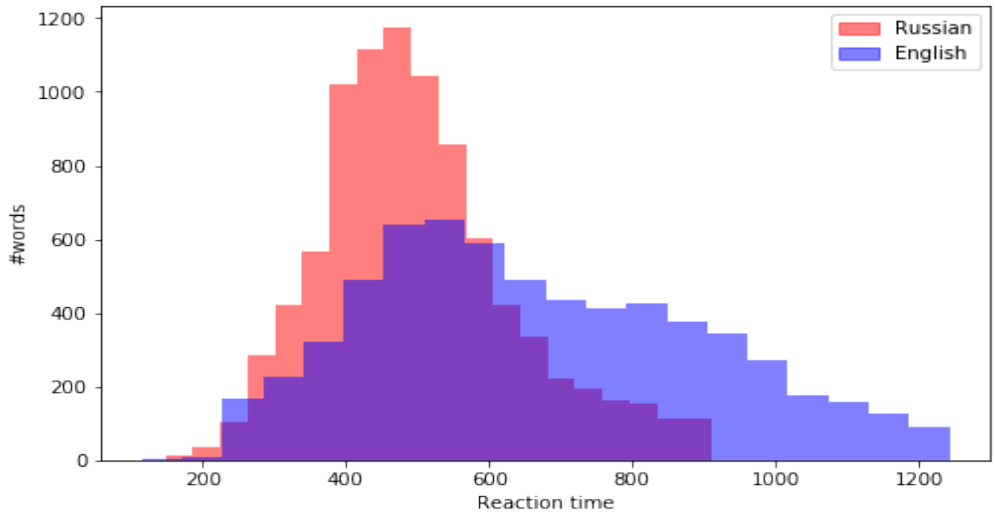

Fig 3. Reaction time distribution for English and Russian words.

Finally, we found no significant contributions of semantic, orthographic, or phonological characteristics of words in terms of predicting the success of participant in recalling the word for participant-independent (see S3 Appendix) or participant-dependent (see S4 Appendix) models. The Spearman correlations for predictions and real labels for hold-out set in prediction of word AUROC were close to zero, the AUROCs of participant-independent models for hold-out set were close to 0.5

with small standard deviation, suggesting that the model did not learn anything useful for prediction of experiment outcome.

4 Discussion

The aim of the current study was to replicate and extend the findings of [1](Experiment 1) with the VLPFC and DLPFC. However, the effect of atDCS over the left VLPFC (decrease in false recognition) was not replicated, and there were weak, negative effects over the DLPFC in the Offline group. In fact, there was a trend toward increased false alarms after online atDCS over the VLPFC and offline atDCS over the DLPFC compared to Sham. The trend suggests that an increase in false alarms obscured group differences, and the increase in false alarms could be driven by an increase in semantic elaboration that leads to better memory for features common to multiple items [41, 42]. While the results do not support effects over the DLPFC and VLPFC, this speaks to the lack of tDCS effect rather than the lack of involvement of these regions in episodic memory. Although it is clear from TMS studies that the VLPFC is necessary for episodic memory, TMS disruption of the DLPFC does not consistently impair episodic memory and the specific roles of both regions remain to be clarified. While meta-analyses and studies aimed at replication of tDCS cognitive effects do not support the effectiveness of tDCS generally [43, 44], the majority of tDCS studies lack needed deeper explorations of non-linear effects and individual differences through modelling and ML [45]. More insight is needed into the relationship between biological and cognitive effects. For example, future studies could attempt to classify groups by using concurrent EEG activity during tDCS administration. However, even tDCS-EEG studies alone may not be beneficial because EEG reflects the synchronized firing of large populations of neurons across the brain (not just the stimulated region), and the neural signature is not always modulated by changes in cognitive performance. Thus, more sensitive measures of tDCS cognitive and neural effects can be implemented, such as changes in resting-state connectivity (Stagg et al., 2014), GABA and glutamate concentrations [46], and cerebral blood flow [47]. Nonetheless, the results of the current study support the conclusion that tDCS does not modulate memory in the VLPFC and DLPFC directly but rather a subprocess leading to successful memory formation and retrieval. Specifically, tDCS may influence the earlier (e.g. semantic elaboration) rather than later (i.e. item-context binding) sub-stages of successful encoding, which can lead to a more indirect effect on retrieval success. For example, tDCS has been shown to modulate the speed of vocabulary learning but not final vocabulary recall over multiple sessions [48]. Moreover, atDCS over the VLPFC led to improved semantic processing in language tasks [31] such as naming common objects [49] and working memory tasks [50]. Thus, tDCS could modulate a semantic process or working memory maintenance that contributes to long-term memory rather than the formation of the memory trace, which is thought to rely more on the hippocampus [51, 52]. It is notable that using AUROC we could not discriminate hard and easy to remember words in Experiment 1 of [1]: the AUC curve revealed no significant effect of VLPFC tDCS on English or Russian speakers. Although average memory accuracy was above chance, AUC suggested that when considering individual trials, participants were guessing. Future studies can test the reliability of the recognition test as an assessment for tDCS-induced cognitive effects and try to increase participant motivation and concentration to attain optimal trial-by-trial performance for examining tDCS-induced changes. The ML analysis showed that phonological, semantic, and orthographic features did not influence the episodic memory. Although it is likely that these features contributed to memory formation, their influence was minimal and overshadowed by the large interaction between atDCS and individual differences. This individual variability has been

examined in other studies that found differences between high and low performers (e.g. [9]). Individual differences in baseline memory performance and encoding ability appear to be an influential factor, more than language or word characteristics. The results indicate that individual differences may be one of the most important parameters not only in the current study and tDCS field but also in replication studies. A limitation was that we did not conduct a within-subjects study in which we examined differences in individual performance. A replication of a within-subjects study could reveal that the effect of a between-subjects study was due to differences between participants and when participants are compared to their own performance, there is little effect of tDCS. Indeed, studies including [1] have found varying effects when comparing between-subjects to within-subjects tDCS effects over the same location, with potentially smaller or absent effects for within-subjects studies (e.g. compare between-subjects results in [30, 31, 53, 54] and Experiment 1 in [1] to within-subjects results in Experiment 3 in [1]; Experiment 2 in [55] [13]. Furthermore, it is important to identify the reliability of tDCS effects with direct replications within the same lab as well as other labs [56] with sufficiently large sample sizes. However, continued future examinations of tDCS cognitive effects may have less value if not supplemented with a measure of biological effects. Future data-driven studies should aim to predict when physiological effects such as increases in BOLD activation lead to cognitive outcomes such as higher performance and whether duration and frequency of tDCS are involved. Although the sample size of the current study should have had sufficient statistical power to detect the original effect size ([1], Experiment 1), the true effect size could still be eluded because of a statistical phenomenon known as the "winner's curse". The winner's curse posits that the first studies to find a significant and novel effect will be published, and the reported effect will be exaggerated because these studies tend to be exploratory and include smaller sample sizes. If several small-sample studies are investigating the same effect, random error and sampling variation may lead to one but not all of the studies finding an effect that crosses significance threshold because under-powered studies can only find large effects. The true effect is likely to be smaller, so it would not otherwise emerge in under-powered studies. This phenomenon is illustrated well in the meta-analysis by [3], in which most included articles reported at least one significant effect, but the average effect size was close to zero when all studies were included in the analysis. For example, [57] found an effect of atDCS in their first experiment but not in a subsequent replication in the same paper . The conclusion of [58] is in line with the "winner's curse" effect, while the authors suggest a smaller sample size (between d = .40 and d = .50 according to Cohen, 1988) as more appropriate for tDCS studies. We did not successfully replicate the results of the previous work, although we used a larger sample size and nearly identical method. It is worth mentioning that the results of the replicated study are significant based on the ANOVA (and this statistical model is appropriate for the original study) but the main effect is not detected by the AUROC analysis applied to both datasets. It does not seem that there was a cultural or linguistic component involved, in line with previous tDCS experiments that found similar effects on verbal memory performance between countries (Italy: [31]; England: [1]; USA: [54]). However, culture-dependent tDCS should be tested directly by comparing individuals of different cultures in the same language. We would expect culture-dependent effects in social cognition (e.g. [48]) but not in processes such as memory encoding that are thought to rely on the same neural architecture across people. We conclude that tDCS may exert a subtle modulation that also interacts strongly with individual differences, particularly in baseline activation and neuroanatomy. ML regression approaches could be successful for future studies that model interactions between biological and cognitive effects. In spite of the significant tDCS effects found in previous memory studies, we suggest caution in interpreting these

effects and applying tDCS as a neuromodulator until replications are conducted with biomarkers.

Supporting information

S1 Appendix. Significantly different words. Russian and English words that differ significantly in the number of hits, false alarms, AUROC and reaction time.

S2 Appendix. Data Augmentations. The procedures used to augment the training set according to the method proused in [37].

S3 Appendix. Participant-Independent ML Trials. The performance of participant-independent TPOT and AutoPytorch models.

S4 Appendix. Participant-Dependent ML Trials. The performance of participant-independent TPOT and AutoPytorch models.

Acknowledgments

The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'. This study used the HSE Synchronous Eye-tracking, Brain Signal Recording and Non-Invasive Brain Stimulation System. We are grateful for the assistance of Anastasiya Bogdanova and Daniil Krivenok in data collection.

Competing interests: We declare no competing interests.

References

1. Authors. Anonymized for the review; 2019.
2. Polania R, Nitsche MA, Ruff CC. Studying and modifying brain function with non-invasive brain stimulation. Nature neuroscience. 2018;21(2):174–187.
3. Galli G, Vadillo MA, Sirota M, Feurra M, Medvedeva A. A systematic review and meta-analysis of the effects of transcranial direct current stimulation (tDCS) on episodic memory. Brain stimulation. 2019;12(2):231–241.
4. Giordano J, Bikson M, Kappenman ES, Clark VP, Coslett HB, Hamblin MR, et al. Mechanisms and effects of transcranial direct current stimulation. Dose-Response. 2017;15(1):1559325816685467.
5. Heroux ME, Loo CK, Taylor JL, Gandevia SC. Questionable science and reproducibility in electrical brain stimulation research. PloS one. 2017;12(4).
6. Manenti R, Cotelli M, Robertson IH, Miniussi C. Transcranial brain stimulation studies of episodic memory in young adults, elderly adults and individuals with memory dysfunction: a review. Brain Stimulation. 2012;5(2):103–109.
7. Szczepanski SM, Knight RT. Insights into human behavior from lesions to the prefrontal cortex. Neuron. 2014;83(5):1002–1018.
8. Fuster JM. Executive frontal functions. Experimental brain research. 2000;133(1):66–70.

9. Habich A, Klöppel S, Abdulkadir A, Scheller E, Nissen C, Peter J. Anodal tDCS enhances verbal episodic memory in initially low performers. Frontiers in human neuroscience. 2017;11:542.

10. Nyberg L, Cabeza R, Tulving E. PET studies of encoding and retrieval: The HERA model. Psychonomic Bulletin & Review. 1996;3(2):135–148.

11. Javadi AH, Cheng P. Transcranial direct current stimulation (tDCS) enhances reconsolidation of long-term memory. Brain stimulation. 2013;6(4):668–674.

12. Javadi AH, Walsh V. Transcranial direct current stimulation (tDCS) of the left dorsolateral prefrontal cortex modulates declarative memory. Brain stimulation. 2012;5(3):231–241.

13. Manenti R, Brambilla M, Petesi M, Ferrari C, Cotelli M. Enhancing verbal episodic memory in older and young subjects after non-invasive brain stimulation. Frontiers in aging neuroscience. 2013;5:49.

14. Lara GAd, Knechtges PN, Paulus W, Antal A. Anodal tDCS over the left DLPFC did not affect the encoding and retrieval of verbal declarative information. Frontiers in neuroscience. 2017;11:452.

15. Zwissler B, Sperber C, Aigeldinger S, Schindler S, Kissler J, Plewnia C. Shaping memory accuracy by left prefrontal transcranial direct current stimulation. Journal of Neuroscience. 2014;34(11):4022–4026.

16. Brunoni AR, Vanderhasselt MA. Working memory improvement with non-invasive brain stimulation of the dorsolateral prefrontal cortex: a systematic review and meta-analysis. Brain and cognition. 2014;86:1–9.

17. Dedoncker J, Brunoni AR, Baeken C, Vanderhasselt MA. A systematic review and meta-analysis of the effects of transcranial direct current stimulation (tDCS) over the dorsolateral prefrontal cortex in healthy and neuropsychiatric samples: influence of stimulation parameters. Brain stimulation. 2016;9(4):501–517.

18. Dedoncker J, Brunoni AR, Baeken C, Vanderhasselt MA. The effect of the interval-between-sessions on prefrontal transcranial direct current stimulation (tDCS) on cognitive outcomes: a systematic review and meta-analysis. Journal of Neural Transmission. 2016;123(10):1159–1172.

19. Lafontaine MP, Theoret H, Gosselin F, Lippe S. Transcranial direct current stimulation of the dorsolateral prefrontal cortex modulates repetition suppression to unfamiliar faces: an ERP study. PloS one. 2013;8(12).

20. Lu Y, Wang C, Chen C, Xue G. Spatiotemporal neural pattern similarity supports episodic memory. Current Biology. 2015;25(6):780–785.

21. Murray LJ, Ranganath C. The dorsolateral prefrontal cortex contributes to successful relational memory encoding. Journal of Neuroscience. 2007;27(20):5515–5522.

22. Blumenfeld RS, Ranganath C. Prefrontal cortex and long-term memory encoding: an integrative review of findings from neuropsychology and neuroimaging. The Neuroscientist. 2007;13(3):280–291.

23. Blumenfeld RS, Parks CM, Yonelinas AP, Ranganath C. Putting the pieces together: the role of dorsolateral prefrontal cortex in relational memory encoding. Journal of cognitive neuroscience. 2011;23(1):257–265.

24. Blumenfeld RS, Nomura EM, Gratton C, D'Esposito M. Lateral prefrontal cortex is organized into parallel dorsal and ventral streams along the rostro-caudal axis. Cerebral Cortex. 2013;23(10):2457–2466.

25. Blumenfeld RS, Ranganath C. Dorsolateral prefrontal cortex promotes long-term memory formation through its role in working memory organization. Journal of Neuroscience. 2006;26(3):916–925.

26. Hawco C, Berlim MT, Lepage M. The dorsolateral prefrontal cortex plays a role in self-initiated elaborative cognitive processing during episodic memory encoding: rTMS evidence. PLoS One. 2013;8(9).

27. Hammer A, Mohammadi B, Schmicker M, Saliger S, Münte TF. Errorless and errorful learning modulated by transcranial direct current stimulation. BMC neuroscience. 2011;12(1):72.

28. Takahashi E, Ohki K, Kim DS. Diffusion tensor studies dissociated two fronto-temporal pathways in the human memory system. Neuroimage. 2007;34(2):827–838.

29. Raposo A, Han S, Dobbins IG. Ventrolateral prefrontal cortex and self-initiated semantic elaboration during memory retrieval. Neuropsychologia. 2009;47(11):2261–2271.

30. Sandrini M, Brambilla M, Manenti R, Rosini S, Cohen LG, Cotelli M. Noninvasive stimulation of prefrontal cortex strengthens existing episodic memories and reduces forgetting in the elderly. Frontiers in aging neuroscience. 2014;6:289.

31. Pisoni A, Vernice M, Iasevoli L, Cattaneo Z, Papagno C. Guess who? Investigating the proper name processing network by means of tDCS. Neuropsychologia. 2015;66:267–278.

32. Balzarotti S, Colombo B. Effects of unilateral transcranial direct current stimulation of left prefrontal cortex on processing and memory of emotional visual stimuli. PloS one. 2016;11(7).

33. Pergolizzi D, Chua EF. Transcranial direct current stimulation over the parietal cortex alters bias in item and source memory tasks. Brain and cognition. 2016;108:56–65.

34. Blumenfeld RS, Lee TG, D'Esposito M. The effects of lateral prefrontal transcranial magnetic stimulation on item memory encoding. Neuropsychologia. 2014;53:197–202.

35. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;5:135–146.

36. Joulin A, Bojanowski P, Mikolov T, Jégou H, Grave E. Loss in translation: Learning bilingual word mapping with a retrieval criterion. arXiv preprint arXiv:180407745. 2018;.

37. Zhang D, Yang Z. Word embedding perturbation for sentence classification. arXiv preprint arXiv:180408166. 2018;.

38. Olson RS, Moore JH. TPOT: A tree-based pipeline optimization tool for automating machine learning. In: Automated Machine Learning. Springer; 2019. p. 151–160.

39. Mendoza H, Klein A, Feurer M, Springenberg JT, Urban M, Burkart M, et al. Towards automatically-tuned deep neural networks. In: Automated Machine Learning. Springer; 2019. p. 135–149.

40. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in pytorch. In: Proceedings of Neural Information Processing Systems; 2017.

41. Einstein GO, Hunt RR. Levels of processing and organization: Additive effects of individual-item and relational processing. Journal of experimental Psychology: Human learning and Memory. 1980;6(5):588.

42. Staresina BP, Gray JC, Davachi L. Event congruency enhances episodic memory encoding through semantic elaboration and relational binding. Cerebral Cortex. 2009;19(5):1198–1207.

43. Horvath JC, Forte JD, Carter O. Quantitative review finds no evidence of cognitive effects in healthy populations from single-session transcranial direct current stimulation (tDCS). Brain stimulation. 2015;8(3):535–550.

44. Vannorsdall TD, Van Steenburgh JJ, Schretlen DJ, Jayatillake R, Skolasky RL, Gordon B. Reproducibility of tDCS results in a randomized trial: failure to replicate findings of tDCS-induced enhancement of verbal fluency. Cognitive and Behavioral Neurology. 2016;29(1):11–17.

45. Peterchev AV. Neuromodulation: Transcranial electric stimulation seen from within the brain. Elife. 2017;6:e25812.

46. Bachtiar V, Near J, Johansen-Berg H, Stagg CJ. Modulation of GABA and resting state functional connectivity by transcranial direct current stimulation. Elife. 2015;4:e08789.

47. Paquette C, Sidel M, Radinska BA, Soucy JP, Thiel A. Bilateral transcranial direct current stimulation modulates activation-induced regional blood flow changes during voluntary movement. Journal of Cerebral Blood Flow & Metabolism. 2011;31(10):2086–2095.

48. Meinzer M, Jähnigen S, Copland DA, Darkow R, Grittner U, Avirame K, et al. Transcranial direct current stimulation over multiple days improves learning and maintenance of a novel vocabulary. Cortex. 2014;50:137–147.

49. Fertonani A, Rosini S, Cotelli M, Rossini PM, Miniussi C. Naming facilitation induced by transcranial direct current stimulation. Behavioural brain research. 2010;208(2):311–318.

50. Fregni F, Boggio PS, Nitsche M, Bermpohl F, Antal A, Feredoes E, et al. Anodal transcranial direct current stimulation of prefrontal cortex enhances working memory. Experimental brain research. 2005;166(1):23–30.

51. Olsen RK, Moses SN, Riggs L, Ryan JD. The hippocampus supports multiple cognitive processes through relational binding and comparison. Frontiers in human neuroscience. 2012;6:146.

52. Staresina BP, Davachi L. Mind the gap: binding experiences across space and time in the human hippocampus. Neuron. 2009;63(2):267–276.

53. Gray SJ, Brookshire G, Casasanto D, Gallo DA. Electrically stimulating prefrontal cortex at retrieval improves recollection accuracy. Cortex. 2015;73:188–194.

54. Matzen LE, Trumbo MC, Leach RC, Leshikar ED. Effects of non-invasive brain stimulation on associative memory. Brain research. 2015;1624:286–296.

55. Smirni D, Turriziani P, Mangano GR, Cipolotti L, Oliveri M. Modulating memory performance in healthy subjects with transcranial direct current stimulation over the right dorsolateral prefrontal cortex. PloS one. 2015;10(12).

56. Cesario J. Priming, replication, and the hardest science. Perspectives on psychological science. 2014;9(1):40–48.

57. Jones KT, Gözenman F, Berryhill ME. Enhanced long-term memory encoding after parietal neurostimulation. Experimental brain research. 2014;232(12):4043–4054.

58. Minarik T, Berger B, Althaus L, Bader V, Biebl B, Brotzeller F, et al. The importance of sample size for reproducibility of tDCS effects. Frontiers in Human Neuroscience. 2016;10:453.