were recorded using two active electrodes situated 2.5cm below the skinfold on each wrists. Electrooculographic (EOG) activity of one eye was recorded using three active electrodes. Two, situated below and above the eye, aimed at recording vertical movements of the eye and one on the side aimed at recording horizontal movements. Physiological signals were measured using a g.USBAmp (g.tec, Austria), sampled at 256 Hz, and processed online using OpenviBE 2.1.0 [25].

To classify the two MI tasks from EEG data, we used participant-specific spectral and spatial filters. First, from the EEG signals recorded during the calibration runs, we identified a participant-specific discriminant frequency band using the heuristic algorithm proposed by Blankertz et al. in [3] (Algorithm 1 in that paper). Roughly, this algorithm selects the frequency band whose power in the sensorimotor channels maximally correlates with the class labels. Here we used channels C3 & C4 after spatial filtering with a Laplacian filter as sensorimotor channels, as recommended in [**blankertz08**]. We selected a discriminant frequency band in the interval from 5 Hz to 35 Hz, with 0.5Hz large bins. Once this discriminant frequency band identified, we filtered EEG signals in that band using a butterworth filter of order 5.

Then, we used the Common Spatial Pattern (CSP) algorithm [24] to optimize 3 pairs of spatial filters, still using the data from the two calibration runs. Such spatially filtered EEG signals should thus have a band power which is maximally different between the two MI conditions. We then computed the band power of these spatially filtered signals by squaring the EEG signals, averaging them over a 1 second sliding window (with 1/16th second between consecutive windows), and log-transforming the results. This led to 6 different features per time window, which were used as input to a Linear Discriminant Analysis (LDA) classifier [18]. As mentioned above, this LDA was calibrated on the data from the two calibration runs. These filters and classifier were then applied on the subsequent runs to provide online feedback.

Variables & Factors:

The aim was to evaluate the influence of the gender of the experimenters and participants on the MI-BCI performances of the participants over a series of 4 runs with online BCI use. Two measures were used to assess the performance of the participants.

The first is the mean classification accuracy which is traditionally used by the community. This measure represents the percentage of time windows from the feedback periods that were correctly classified. However, this metric only considers whether the classification was correct, but not the quality of this classification, i.e., it does not take into account the classifier output. Since our participants were instructed to train to obtain not only a correct classification, but also a feedback bar as long as possible, we also studied a metric considering the feedback bar length, i.e., the classifier output.

Thus, we also used the Quality-Weighted Accuracy (QWA), the standard performance metric provided in

OpenViBE MI-BCI applications, which is inspired by the SensoriMotor Rhythm quality score in [7]. To compute it, we first summed the (signed) LDA classifier outputs (distance to the separating hyperplane) over all time windows during a trial feedback period. If this sum sign matched the required trial label, i.e. negative for left hand MI and positive for right hand MI, then the trial was considered as correctly classified, otherwise it was not. Finally, a run QWA was estimated as the percentage of trials considered as correctly classified using this approach.

RESULTS

Comparability of groups:

Among 59 participants, 3 outperformed the others (by more than two SDs) both in term of mean classification accuracy (Outliers $\bar{X}1=88.94$, $\bar{X}2=90.36$, $\bar{X}3=94.51$; $\bar{X}$grp=59.33%; SDgrp=12.3) and QWA (Respectively, outliers $\bar{X}1=98.13$, $\bar{X}2=98.13$, $\bar{X}3=99.38$; $\bar{X}$grp=62.78%; SDgrp=16.2). Thus, the following analyzes are based on the results of 56 participants (27 women).

Before it all, we verified if the distribution of the data collected was normal using Shapiro-Wilk tests. The variables describing the mental rotation scores (p=0.34), anxiety (p=0.06) and autonomy (p=0.14) of our participants could be considered as having a normal distribution. Though, the mean classification accuracy of the runs did not have a normal distribution ($p \leq 10^{-3}$) and neither did the QWA metrics for the different runs ($p \leq 10^{-3}$).

We also checked that groups formed by participants' gender, i.e., *"ParGender"*, and experimenters' gender, i.e., *"ExpGender"*, were comparable. We focused on mental rotation scores (MRS), anxiety or autonomy, which were shown to influence on MI-BCI performances [10]. Participants with low MRS [30], anxious or non-autonomous (both measured using the 16PF5 questionnaire [4]) were shown to have lower MI-BCI performances than the others [10, 28]. To check that groups were comparable, we ran 2-way ANOVAs with *"ExpGender*ParGender"* as independent variables and either mental rotation scores, anxiety or autonomy as dependent variable.

Results indicate that groups are comparable in terms of anxiety. Though, participants' gender influence their MRS [$F(1,52)=17.47$; $p \leq 10^{-3}$, $\eta^2=0.25$]. Men ($\bar{X}$men=0.072; SD=0.024) had higher MRS than women ($\bar{X}$women=0.045; SD=0.023), which is in accordance with the literature [15]. Furthermore, participants training with men or women experimenters did not have the same level of autonomy [$F(1,52)=4.01$; p=0.05, $\eta^2=0.07$]. Participants training with men experimenters ($\bar{X}$menExp=6.35; SD=1.74) were more autonomous than participants training with women experimenters ($\bar{X}$womenExp=5.67; SD=1.66). Therefore, we controlled for the influence of these variables in our subsequent analyses by using them as covariates in ANCOVAs (see paragraph *Checking for confounding factors*).

Participants' and experimenters' gender:

Then, we analyzed the influence of the gender of the

experimenters and participants on the MI-BCI performances of the participants over the runs, i.e., *"Run"*. To do so, we performed a 3-way repeated measures mixed ANOVAs with *"ExpGender*ParGender*Run"* as independent variables and the repeated measures of performance over the runs, i.e., mean classification accuracy or QWA, as dependent variable. Even though the normality of the data is a pre-requisite of an ANOVA, the ANOVA is considered as robust against the normality assumption and to the best of our knowledge no other non parametric test enabled to perform such analysis.

First, we performed such ANOVA using the mean classification accuracy. Results revealed no simple effect of *"Run"* [F(3,156)=1.53; p=0.22, η^2=0.03], *"ExpGender"* [F(1,52)=0.26; p=0.61, $\eta^2 \leq 0.01$] and *"ParGender"* [F(1,52)=0.23; p=0.64, $\eta^2 \leq 0.01$]. They revealed no interaction of *"Run*ParGender"* [F(3,156)=1.92; p=0.13, η^2=0.04], *"Run*ExpGender"* [F(3,156)=0.23; p=0.87, η^2=0.01] nor *"ParGender*ExpGender"* [F(1,52)=0.92; p=0.34, η^2=0.02]. Finally, the interaction of *"Run*ParGender*ExpGender"* was not significant either [F(3,156)=1.38; p=0.25, η^2=0.03].

Next, we performed this same analysis using QWA. Results revealed no simple effect of *"Run"* [F(3,16)=1.81; p=0.15, η^2=0.03], *"ExpGender"* [F(1,52)=0.54; p=0.47, η^2=0.01] nor *"ParGender"* [F(1,52)=0.09; p=0.76, η^2=0.01]. They also revealed no interaction of *"Run*ExpGender"* [F(3,16)=0.08; p=0.97, η^2=10^{-2}] nor *"ParGender*ExpGender"* [F(1,52)=0.60; p=0.44, η^2=0.01]. Though, the *"Run*ParGender"* interaction was significant [F(3,156)=5.98; p=0.001, η^2=0.1]. Figure 1 represents the evolution of the participants' QWA depending on their gender.
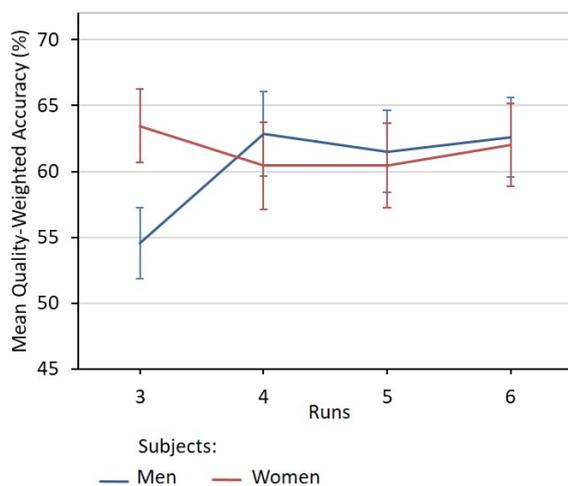

Figure 1: QWA evolution depending on participants' gender with standard errors.

Finally, a significant *"Run*ParGender*ExpGender"* interaction was found [F(3,156)=3.46; p=0.02, η^2=0.06]. Figure 2 represents the participants' QWA evolution depending on the participants' and experimenters' gender.
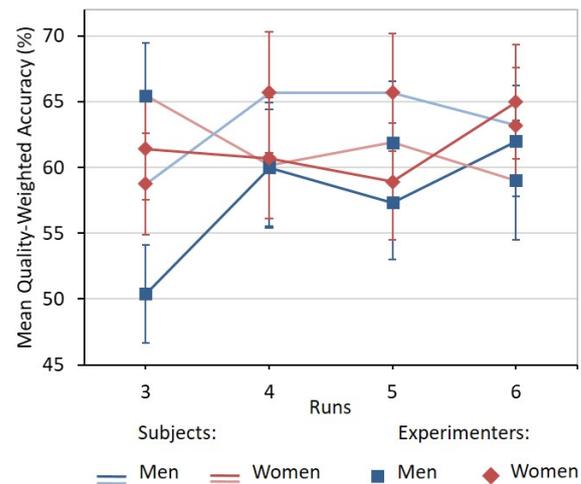

Figure 2: Participants' QWA evolution depending on the participants' and experimenters' gender with standard errors.

Checking for confounding factors:
As stated before, the groups of participants formed using the participants' and experimenters' gender had differences in terms of mental rotation scores and autonomy. Therefore, we studied the potential impact these differences could have had on our results. First, we checked if a correlation could be found between our metrics of performances and these variables. No significant correlation was found between the autonomy and the mean classification accuracy ($r = -0.11$, $p = 0.40$) nor the QWA ($r = -0.07$, $p = 0.62$). The correlations between the mental rotation score and the mean classification accuracy ($r = -0.13$, $p = 0.36$) or the QWA ($r = -0.24$, $p = 0.08$) was not significant either.

Second, we ran a 3-way repeated measures mixed ANCOVA with *"ExpGender, ParGender, Run"* as independent variables and one of the measures of performance, i.e., mean classification accuracy or QWA, as dependent variable, with the autonomy, i.e., *"Aut"*, or the mental rotation score, i.e., *"MRs"*, of the participants as covariate. When performing the analysis on the QWA we did not find any single effect or interaction of the autonomy (*"Aut"* [F(1,51)=0.26; p=0.61, η^2=10^{-2}], *"Aut*Run"* [F(3,15)=0.81; p=0.49, η^2=0.02]) or the mental rotation score (*"MRs"* [F(1,51)=1.75; p=0.19, η^2=0.03], *"MRs*Run"* [F(3,15)=1.52; p=0.21, η^2=0.03]).

When investigating the mean classification accuracy we found as well no impact of the autonomy (*"Aut"* [F(1,51)=0.44; p=0.51, η^2=10^{-2}], *"Aut*Run"* [F(3,15)=1.46; p=0.23, η^2=0.03]) or the mental rotation score (*"MRs"* [F(1,51)=1.05; p=0.31, η^2=0.02], *"MRs*Run"* [F(3,15)=1.35; p=0.26, η^2=0.03]).

DISCUSSION

We analyzed results using two metrics of performances: QWA which represented what the participants were instructed to improve during training, and the mean classification accuracy, a traditional measure of BCI perfor-

mances. Initial differences in mental rotation scores and autonomy between groups did not seem to bias results.

No influence of the experimenters' and/or participants' gender on the mean accuracy performance was found. Though, we found a significantly different evolution across runs of QWA between men and women participants (see Figure 1). Women participants seemed to start the training with already good QWA, which decreased on the second run and increased again during the last run. Men participants however, started with rather low QWA and then drastically improved on the second run and then stagnated to reach slightly higher final QWA performances than women.

In addition, experimenters' gender seemed to have an influence on this previous interaction. Indeed, the evolution of the QWA appear to depend on participants' and experimenters' gender (see Figure 2). On the one hand, we found the same tendency for men participants to start with lower QWA at the beginning of the session independently of the experimenter's gender. However, men seemed to start with drastically lower QWA performances when they were training with men experimenters. They also seemed to have higher QWA performances throughout the session when they were training with women experimenters. On the other hand, women participants seemed to start with higher QWA when training with men experimenters, though their QWA performances tended to drop throughout the session. However, when training with women experimenters they seemed to have a great increase in QWA during the last run.

Interestingly enough, this result does not match those of a recently published neurofeedback study [31], in which the combination woman experimenter–woman participant appeared to hamper the training outcomes of the last, so that no learning effect was observed in this group.

CONCLUSION

We investigated the potential influence of the experimenters' gender depending on the participants' gender on MI-BCI performances and progression throughout one MI-BCI session. Six experimenters (3 men; 3 women) trained 59 participants (30 men; 29 women). The general observation emerging from this study is that women experimenters seemed to induce better QWA performance progress for both men and women participants. Men participants seemed to start with substantially lower performances when they were training with men experimenters compared to when they were training with women experimenters. Also, even though women participants started with higher performances when training with men experimenters, their performances decreased throughout the session when they overall increased when training with women experimenters.

These results naturally need to be confirmed with larger populations. Further analysis are also needed regarding other variables that might influence or provide insights on our results. This includes inter-experimenter variables

(e.g., traits or teaching competence), intra-experimenters variables (e.g., appearance or states), inter and intra-participants variables (e.g., traits or motivation) and interaction related variables (e.g., quantity and quality of interaction between the participant and the experimenter). There might also be other analysis to perform based on different performance metrics reflecting user performances independently of the classifier output [16].

Further formal studies investigating the role of the BCI experimenter are needed. The need for research methods that explicit larger amounts of influencing factors (such as the experimenter) emerging from experimental protocol and context is equally important. In BCI research, the instructions (i.e., what participants are instructed to do during mental-imagery tasks) are rarely formalized, or in any case they are not taken into account and mentioned in papers. Similarly, protocols rarely evoke demonstrations [17] (i.e. showing a demonstration of a successful BCI use to the participant, together with a demonstration of feedback during (in)correctly performed mental tasks). It is common practice for studies in the BCI field not to report experimenter gender, though the literature as well as our results indicate that the influence of experimenters should be considered carefully while designing and reporting experimental protocols.

Literature suggests several solutions to limit the potential bias arising from the experimenter [19, 26]. These methods include: monitoring participant-experimenter interaction; increasing the number and diversity of data collectors; pre-testing the method and controlling expectancy; providing an extensive training for administrators/ data collectors; monitoring and standardizing the behavior of experimenters with detailed protocol and pre-written instructions for the participant; and statistically controlling for bias.

Beyond the potential bias that could arise from the experimenters' presence, the social and emotional feedback that experimenters provide could benefit MI-BCI. Indeed, the use of social feedback in BCI has been encouraged [27]. Social presence and trust relationship between the user and the experimenter are essential for maintaining training motivation, which has been shown to facilitate the BCI learning process [11].

During MI-BCI training, using a learning companion to provide the participants with social and emotional feedback have proven effective in improving the user experience [23]. An advanced conversational agent could also be used to supplement the role of the experimenter. It would represent yet another interesting method to control and/or enhance the experimenter influence. Taking experimenter-related factors into account might lead to a conjoint progress of the global BCI performance and the validity and understanding of BCI experimental results.

ACKNOWLEDGEMENTS

REFERENCES

[1] Allison BZ, Neuper C. Could anyone use a BCI? In: Brain-computer interfaces, 2010, 35–54.

[2] Aslaksen PM, Myrbakk IN, Høifødt RS, Flaten MA. The effect of experimenter gender on autonomic and subjective responses to pain stimuli. Pain. 2007;129(3):260–268.

[3] Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller KR. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal processing magazine. 2008;25(1):41–56.

[4] Cattell RB, P. Cattell HE. Personality structure and the new fifth edition of the 16PF. Educational and Psychological Measurement. 1995;55(6):926–937.

[5] Chapman CD, Benedict C, Schiöth HB. Experimenter gender and replicability in science. Science advances. 2018;4(1):e1701427.

[6] Clerc M, Bougrain L, Lotte F. Brain-Computer Interfaces 2: Technology and Applications. John Wiley & Sons, 2016.

[7] Grosse-Wentrup M, Schölkopf B. High gamma-power predicts performance in sensorimotor-rhythm brain–computer interfaces. Journal of neural engineering. 2012;9(4):046001.

[8] Grysman A, Denney A. Gender, experimenter gender and medium of report influence the content of autobiographical memory report. Memory. 2017;25(1):132–145.

[9] Iachini T, Coello Y, Frassinetti F, Senese VP, Galante F, Ruggiero G. Peripersonal and interpersonal space in virtual and real environments: Effects of gender and age. Journal of Environmental Psychology. 2016;45:154–164.

[10] Jeunet C, N'Kaoua B, Subramanian S, Hachet M, Lotte F. Predicting Mental Imagery-Based BCI Performance from Personality, Cognitive Profile and Neurophysiological Patterns. PLoS ONE. 2015;10(12):20.

[11] Kleih SC et al. Motivation influences performance in SMR-BCI. na, 2011.

[12] Kline JP, Blackhart GC, Joiner TE. Sex, lie scales, and electrode caps: An interpersonal context for defensiveness and anterior electroencephalographic asymmetry. Personality and Individual Differences. 2002;33(3):459–478.

[13] Kober SE, Witte M, Ninaus M, Neuper C, Wood G. Learning to modulate one's own brain activity: the effect of spontaneous mental strategies. Frontiers in human neuroscience. 2013;7:695.

[14] Levine FM, De Simone LL. The effects of experimenter gender on pain report in male and female subjects. Pain. 1991;44(1):69–72.

[15] Linn MC, Petersen AC. Emergence and characterization of sex differences in spatial ability: A meta-analysis. Child development. 1985;1479–1498.

[16] Lotte F, Jeunet C. Defining and quantifying users' mental imagery-based BCI skills: a first step. Journal of neural engineering. 2018;15(4):046030.

[17] Lotte F, Larrue F, Mühl C. Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. Frontiers in Human Neuroscience. 2013;7:568.

[18] Lotte F et al. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. Journal of neural engineering. 2018;15(3):031005.

[19] Miyazaki AD, Taylor KA. Researcher interaction biases and business ethics research: Respondent reactions to researcher characteristics. Journal of Business Ethics. 2008;81(4):779–795.

[20] Neuper C, Scherer R, Reiner M, Pfurtscheller G. Imagery of motor actions: Differential effects of kinesthetic and visual–motor mode of imagery in single-trial EEG. Cognitive brain research. 2005;25(3):668–677.

[21] Nichols AL, Maner JK. The good-subject effect: Investigating participant demand characteristics. The Journal of general psychology. 2008;135(2):151–166.

[22] Pfurtscheller G, Neuper C. Motor imagery and direct brain-computer communication. Proceedings of the IEEE. 2001;89(7):1123–1134.

[23] Pillette L, Jeunet C, Mansencal B, N'Kambou R, N'Kaoua B, Lotte F. Peanut: Personalised emotional agent for neurotechnology user-training. In: 7th International BCI Conference. 2017.

[24] Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE transactions on rehabilitation engineering. 2000;8(4):441–446.

[25] Renard Y et al. OpenViBE: An Open-Source Software Platform to Design, Test and Use Brain-Computer Interfaces in Real and Virtual Environments. Presence. 2010;19(1):35–53.

[26] Rosnow R, Rosenthal R. People studying people: Artifacts and ethics in behavioral research. WH Freeman, 1997.

[27] Sexton CA. The overlooked potential for social factors to improve effectiveness of brain-computer interfaces. Frontiers in systems neuroscience. 2015;9:70.

[28] Tan LF, Dienes Z, Jansari A, Goh SY. Effect of mindfulness meditation on brain–computer interface performance. Consciousness and cognition. 2014;23:12–21.

[29] Uzzell D, Horne N. The influence of biological sex, sexuality and gender role on interpersonal distance. British Journal of Social Psychology. 2006;45(3):579–597.

[30] Vandenberg SG, Kuse AR. Mental rotations, a group test of three-dimensional spatial visualization. Perceptual and motor skills. 1978;47(2):599–604.

[31] Wood G, Kober SE. EEG Neurofeedback Is Under Strong Control of Psychosocial Factors. Applied psychophysiology and biofeedback. 2018;43(4):293–300.

[32] Zizzo DJ. Experimenter demand effects in economic experiments. Experimental Economics. 2010;13(1):75–98.