# Neurofeedback on children with Attention-Deficit/Hyperactivity Disorder: what factors influence its efficacy?

April 27, 2018

Aurore Bussalb[a], Richard Delorme[b], Eric Acquaviva[b], Marco Congedo[c], Jean-Arthur Micoulaud-Franchi[d], Louis Mayaud[a]

[a]: Mensia Technolgies, Paris France
[b]: Hôpital Robert Debré, Paris France
[c]: GIPSA-Lab, Grenoble France
[d]: Univ. Bordeaux, SANPSY, USR 3413, F-33000 Bordeaux, France

**Full postal address**
Mensia Technologies
Plateforme d'innovation Boucicaut
130 rue de Lourmel
75015 Paris - France
Aurore Bussalb's e-mail: aurore.bussalb@mensiatech.com
Louis Mayaud's e-mail: lm@mensiatech.com

**Abstract**

Neurofeedback is a noninvasive technique that aims to reduce the ADHD symptoms. Given the impact of meta-analysis on that subject, we first proposed to replicate and update the last one. Then, we tried to identify factors with an influence on Neurofeedback based on the heterogeneity of studies using three multivariate approaches which associated factors with the within subject effect size. The replication and update of the latest meta-analysis confirm the results obtained by the authors: effect sizes are not significant when probably blind ratings are the outcome whereas they are for most proximal raters. Analysis of factors identify 3 elements which may have an impact on Neurofeedback efficacy: the length of the treatment, the quality of the acquisition of the signals and the person assessing the evolution of the symptoms. Besides these results, we introduce here a new way to look into the heterogeneity of clinical trials.

# 1 Introduction

Attention deficit/hyperactivity disorder (ADHD) is a common psychiatric disorder of childhood with an estimated prevalence of about 5% in school-aged children yielding to an estimated 3.6 millions of children in Europe [Association et al., 2013].

This neurodevelopmental disorder is characterized by impaired attention and/or hyperactivity/impulsivity, symptoms which may persist in adulthood with clinical significance which makes ADHD a life-long problem for many patients [Faraone et al., 2006]. There is mounting evidence that objective neurophysiological measurements also identify specific phenotypes: this was particularly reported with electroencephalogram (EEG) recordings [Loo et al., 2017]. In case of ADHD, more theta waves (4-7Hz) are present in the frontal area whereas there are less beta waves (12-32Hz) and sensorimotor rhythm (SMR) (13-15Hz) in the central area [Monastra, 2005; Matoušek et al., 1984; Janzen et al., 1995].

Neurofeedback (NFB) is a noninvasive technique based on behavioral therapy that aims to reduce the ADHD symptoms. The brain is trained to improve its own regulation by providing real-time video/audio information about its electrical activity measured from scalp electrodes [Arns et al., 2015; Steffert and Steffert, 2010]. In case of ADHD, several NFB protocols have been proposed and investigated to decrease the symptoms:

Protocols based on frequency band training: a child can be asked to enhance his SMR while suppressing theta or beta [Lubar and Shouse, 1976], or he can have to enhance beta while suppressing theta (this scenario is known as theta beta ratio (TBR)) [Arns et al., 2013];

Protocol based on the slow cortical potentials (SCP) training which consists in the regulation of cortical excitation thresholds by focusing on activity generated by external cues (similar to event-related potentials (ERPs)) [Heinrich et al., 2004; Banaschewski and Brandeis, 2007];

Protocol based on ERPs (P300) [Fouillen et al., 2017]: ADHD children have a reduced P300 amplitude so it can be considered as a specific neurophysiological marker of selective attention.

Thus, during NFB sessions, the child has to concentrate in order to modify his waves: if he manages to change correctly his EEG patterns, he will be rewarded thank to a positive visual or auditive feedback. The operant conditioning principle will enable the child to repeat more and more easily this task and thanks to the natural neuronal plasticity, a neuronal reorganization is observed [Van Doren et al., 2017].

Since the early 1970s, NFB has been investigating as a potential treatment for ADHD resulting in a large body of scientific literature [Lubar and Shouse, 1976; Rossiter and La Vaque, 1995; Linden et al., 1996; Maurizio et al., 2014]. Besides, the efficacy of NFB on the core symptoms of ADHD (inattention, hyperactivity and impulsivity) has been subject to several meta-analytic studies [Loo and Barkley, 2005; Lofthouse et al., 2012; Arns et al., 2009; Micoulaud-Franchi et al., 2014; Sonuga-Barke et al., 2013].

The most recent meta-analysis solely on the efficacy of NFB has been conducted by Cortese et al. [2016] in which 13 studies are included. Although only randomized controlled trials (RCTs) are selected, the authors of this meta-analysis have made some choices which have been debated by the community in particular by Micoulaud-Franchi et al. [2016] who has criticized the use of a uncommon behavioral scale provided by Steiner et al. [2014] for the teachers' assessments and the inclusion of a pilot study carried out by Arnold et al. [2014] in the meta-analysis.

Meta-analysis have a particularly important impact on the NFB domain which is characterized by a clinical literature that is tremendously heterogeneous in terms of methods. Therefore, when a meta-analysis seems to present loopholes, it is important to make sure the choices made are wise and that the study was well conducted. While we mostly agree with Cortese et al. [2016], we decided to replicate the methods he suggested to a few exceptions. Indeed, we have investigated the limitations of his work and amended as detailed below. So we replicated and updated this existing meta-analysis with new-found studies matching Cortese et al. [2016] inclusion criteria.

This first step underlines the fact that performing a meta-analysis is complex because of the heterogeneity of the studies published on NFB. Indeed, they differ on many points such as for instance trial methodology and NFB implementation. So even if all included studies are conformed to an inclusion criteria, they remain different from each other. Since we supposed that the choices made by authors may lead to various NFB results, we extended the replication and updating of Cortese et al. [2016] to a broader set of studies and used adequate statistical tools to take advantage of the heterogeneity of the methodological implementations (both clinical and technical) in order to identify which of the factors independently influences the reported effect size (ES).

# 2 Materials and Methods

## 2.1 Select the studies

A systematic review was conducted in order to identify studies to include in the update of Cortese et al. [2016] and in the analysis of factors. On one hand, search terms were entered in Pubmed and on the other hand all articles previously included in meta-analysis were identified. Then duplicates were removed and only studies available in English, German or French describing trials on NFB treatment for ADHD were selected. Subjects had to be diagnosed ADHD based on DSM-4 [Association, 2000], DSM-5 [Association, 2013], ICD-10 [Organization, 1993] criteria or according to an expert psychiatrist. Studies that included more than eight subjects in each study group as well as subjects younger than 25 years old were kept. Eventually, if the remaining articles provided enough data to compute required metrics for the following analysis, they were included. The last step was to apply the inclusion criteria of Cortese et al. [2016] in order to replicate and update this meta-analysis.

## 2.2 Extract data

In included studies, the severity of ADHD symptoms have been assessed by parents and, when available, by teachers. Cortese et al. [2016] and Micoulaud-Franchi et al. [2014] defined parents as most proximal (MProx) raters who are not blind to the treatment of their child, as opposed to teachers who are considered as probably blind (Pblind) raters. We decided to follow these definitions for the work on the meta-analysis and for the analysis of factors. Efficacy of NFB was given for the following outcomes on clinical scales when available: inattention, hyperactivity/impulsivity and total scores. Nevertheless, the factors analysis was performed only with the total score reported on clinical scales.

## 2.3 Perform a meta-analysis

Meta-analysis is a powerful tool that enables to quantify the effect of a treatment by combining results from several studies thanks to an ES.ES represents the difference between two groups and since it is a standardized measure, it can be calculated for different studies and then combined into an overall summary as detailed below and described precisely in the Supplemental material.

To conduct meta-analysis, different software exist: for instance Cortese et al. [2016] used RevMan 5.3 [Cochrane Collaboration, 2011] which computes the ES and its variance of each included study by applying the formula presented in Morris [2008]. However, in order to compute the variance of the ES, the pooled within-group Pearson correlation $\rho$ (i.e the pre-post correlation) was required [James et al., 2013]. In our case, this correlation was not known and the raw data were not available so we took an approximation: Balk et al. [2012] found that a value of 0.5 yields values closer to those computed with the right value of the correlation. In this replication of the work of Cortese et al., the same formulas are used [Borenstein et al., 2009] but instead of using RevMan, a Python code was developed in order to perform the meta-analysis. To increase replicability and transparency and promote open science, we provide the full raw data used for this research as well as the Python code developed available on a Github repository; it is tested with Cortese et al. [2016] raw data to show that same results were found and could be used for replication and expansion of this work. The toolbox could also be used to run any similar meta-analysis.

## 2.4 Replicate and update a meta-analysis

Cortese et al. work was replicated following the previously described steps implemented in the Python code. Besides using the Python code instead of RevMan, we choose to bring two major changes:

the ES of Arnold et al. study is computed from the post-test clinical values taken after the 40 sessions were completed contrary to Cortese et al. [2016] who had used the results after 12 sessions of NFB because final values were not available;

the ES computed from teachers' assessments for Steiner et al. [2014] rely on the BOSS Classroom Observation [Shapiro, 2010] whereas another scale more often used [Christiansen et al., 2014; Bluschke et al., 2016] and which is the revision of the Conners Rating Scale Revised [Conners et al., 1998] whose reliability has been studied [Collett et al., 2003]. Thus we decided to compute the ES based on the results from the Conners.

As suggested in Cortese et al. [2016] we performed two subgroups analysis: first, summary effect (Se), the weighted average of all the ES, was calculated with only studies following standard protocol as defined by Arns et al. [2014] and second with studies whose participants take low-dose or no medication during the trial. These analysis were performed with the choices described above.

Eventually, the new studies conformed to the inclusion criteria defined by Cortese et al. were added to the replication of the meta-analysis.

## 2.5 Detect factors influencing the Neurofeedback

The second step of this work consisted in determining the factors that possibly influence the efficacy of NFB using various statistical methods described below.

### 2.5.1 Identify and pre-process factors

There are arguably 3 types of factors influencing the measured efficacy of an intervention: methodological, technical and linked to the quality of acquisition. Factors were chosen based on what was typically reported in the literature and presumed to influence ES.

*methodological biases*: the presence of a control group, the psychostimulants intake during NFB treatment, the age bounds of children, the blinding of assessors, the randomization of subjects, and the approval by an Institutional Review Board (IRB);

*technical factors*: the protocol used (SCP, SMR, theta up, beta up in frontal areas, theta down), the presence of a transfer phase during NFB training, the possibility to train at home or at school with a transfer card reminding of the NFB session, the ocular artifacts correction, the artifact correction based on amplitude, the type of thresholding reward, the number of NFB sessions, the sessions frequency during a week, the session length and the treatment length;

*quality of acquisition* the presence of one or more active electrode and the EEG quality. This last factor was an indicator between 1 and 3: if EEG was recorded and processed in poor conditions then the indicator would be 1. Besides, if the article didn't precise the recording conditions, the factor would be set to 1. To get an indicator bigger than 1, several points had to be satisfied:

*the type of electrodes used*: AgCl/Gel and Gold/Gel are preferable;

*check of the electrode contact quality trough the amplifier impedance acquisition mode*: impedance must be $< 40\text{k}\Omega$.

*the amplifier used*: those that are conformed to European norms (such as Thera Prax ®Neuroconn and Eemagine EE-430) are preferable or whose reliability is known.

To prevent any bias, the names of these factors were hidden during the analysis and were only revealed once the data analysis and results were accepted as valid: this included choice of variable normalization and validation of model hypothesis as detailed below.

The pre-processing of factors for the analysis included the following steps: factors for which there were too many missing observations, arbitrarily set to more than 20% of the total of observations, were removed from the analysis. Furthermore, if a factor have more than 80% similar observations it was removed as well. Since some of the independent variables were categorical, they were coded in dummies meaning that the presence of the factor is represented by a 1 and the its absence by 0. Independent variables are standardized, except when the decision tree is performed.

### 2.5.2 Associate independent factors to effect sizes

To compute this ES, means of total ADHD score given by parents and teachers were used. Besides, in case studies provided results for more than one behavioral scale, ES were computed for each one. The ES computed in this analysis was different from the one used previously for the replication and updating of Cortese et al. [2016]. Indeed, here we focused on the effect of the treatment within a group as defined by Cohen [1988], definition of the ES that was already used in the literature [Arns et al., 2009; Maurizio et al., 2014; Strehl et al., 2017]. This ES enables to quantify the efficacy of NFB inside the treatment group. as presented in eq. (1).

$$\text{ES} = \frac{M_{\text{post},T} - M_{\text{pre},T}}{\sqrt{\frac{\text{SD}_{\text{pre},T}+\text{SD}_{\text{post},T}}{2}}}. \tag{1}$$

The ES was then considered as a dependent variable to be explained by the factors identified using the following three methods, which were implemented in the Scikit-Learn Python [Pedregosa et al., 2011] and the Statsmodels Python[Seabold and Perktold, 2010] libraries:

weighted multiple linear regression Weighted Least Squares (WLS) [Montgomery et al., 2012];

sparsity-regularized linear regression with Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani, 1996];

decision tree [Quinlan, 1986].

The most often used linear regression analysis is the Ordinary Least Squares (OLS) but here we applied the WLS as described in eq. (2): a weight was assigned to each observation in order to take into account the fact that some observations came from the same study because studies may provide several scales. Besides, the weight is function of the sample size as well: because of their different sample sizes, studies are not equivalent and should be analyzed accordingly. That's why the weight corresponds to the ratio between the experiment group's sample size of the study and the number of behavioral scales available in the study. We also run the analysis with OLS method to assess the impact of the weights on the results.

$$\mathsf{W}y = \mathsf{WX}\beta + \epsilon. \tag{2}$$

$\mathsf{X}$ is a $(n \times p)$ full rank matrix and represents $n$ observations on each $p-1$ independent variables and an intercept term, $\beta$ is a $(p \times 1)$ vector of associated regression coefficients, $\mathsf{W}$ is a $(n \times n)$ diagonal matrix with weights, $y$ is a $(n \times 1)$ vector of dependent variables and $\epsilon$ is a $(n \times 1)$ vector of errors.

The aim of the WLS is to estimate the vector of coefficients $\beta$ by minimizing the Weighted Residual Sum of Squares (WRSS) as presented in eq. (3):

$$\mathsf{WRSS} = \sum_{i=1}^{n} w_i \Big( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2. \tag{3}$$

Thanks to a t-test, we could conclude if coefficients were significantly different from 0 (the p-value had inferior to 0.05). If this was the case, then the associated factor had an influence on NFB results. Moreover, if the coefficient was negative then its associated factor improve the effect of NFB.

However, before interpreting the results of the WLS, the assumptions of this model had to be checked (distribution of the residuals are normal, the moment matrix $\mathbf{X^T W^T W X}$ must be full rank, the fit had to be significant and the independent variables had to be uncorrelated).

The second method applied was the sparsity-regularized linear regression: LASSO. This method is able to perform variable selection in the linear model thanks to $\ell - 1$-norm applied on the coefficients. The coefficients $\beta_j$ are obtained by minimizing the term presented in eq. (4):

$$\sum_{i=1}^{n} \Big( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4}$$

$\lambda$ is the tuning parameter: as it increases, more coefficients are set to zero. The optimal tuning parameter was determined by a leave-one-out cross-validation. This method retains $n$ - 1 observations as the validation data for testing the model and the remaining observation is used as training data. The cross-validation process is then repeated $n$ times with each of the observation used exactly once as the testing data. For each fold, the Mean Square Error (MSE) on the test set was computed and eventually, the $n$ results can be averaged to produce a single observation that enables to find the optimal $\lambda$: it corresponds to the abscissa of the minimum value of the MSE on the mean fold computed for a large range of $\lambda$ [James et al., 2013].

Eventually, the last method used to determine factors influencing NFB was the decision tree. It broke down a dataset into smaller and smaller subsets based on the MSE. The final tree was a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node (terminal node) represents a decision on the dependent target (the ES). The topmost decision node (root node) in a tree corresponds to the best predictor. Besides the lower a factor is situated in the tree, the less reliable it is because the split is run on a smaller subset.

No weights are applied when running the LASSO and the decision tree because the option is not implemented in the Python methods used in our analysis.

# 3 Results

## 3.1 Systematic review

Search terms entered in Pubmed returned 152 results during the last check on December 14, 2017 and 28 articles included in previous meta-analysis on NFB were identified. After the selection process illustrated in fig. 1, 31 studies were included in the factors analysis and 15 in the update of Cortese et al. [2016].

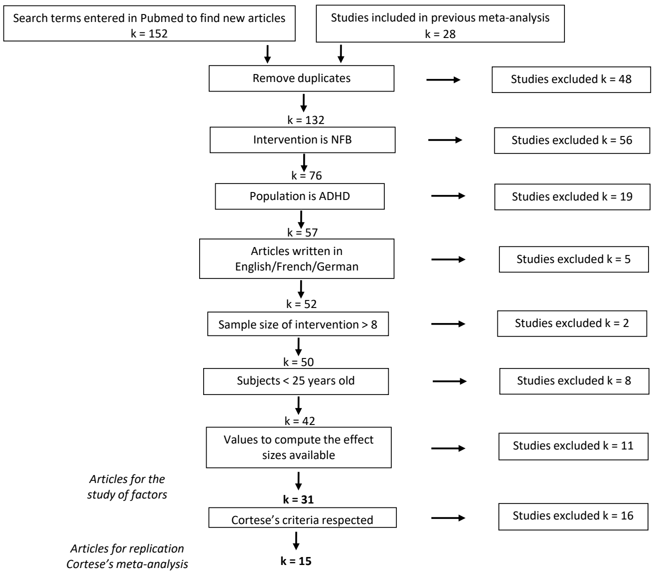
Search terms entered in Pubmed to find new articles k = 152
Studies included in previous meta-analysis k = 28
Remove duplicates
Studies excluded k = 48
k = 132
Intervention is NFB
Studies excluded k = 56
k = 76
Population is ADHD
Studies excluded k = 19
k = 57
Articles written in English/French/German
Studies excluded k = 5
k = 52
Sample size of intervention > 8
Studies excluded k = 2
k = 50
Subjects < 25 years old
Studies excluded k = 8
k = 42
Values to compute the effect sizes available
Studies excluded k = 11
Articles for the study of factors
k = 31
Cortese's criteria respected
Studies excluded k = 16
Articles for replication Cortese's meta-analysis
k = 15

Figure 1: Flow diagram of selection of studies (last search on December 14, 2017).

## 3.2 Replicate a meta-analysis

Contrary to Cortese et al. [2016], the Python code returns a negative ES when the clinical trial is in favor of NFB. The Se whose p-value $< 0.05$ was considered statistically significant. A significant Se means that there is a significant difference between treatment and control groups' effects.

First, when using the ES found by Cortese et al. [2016] thanks to RevMan [Cochrane Collaboration, 2011], and then performed the following steps of meta-analysis with the Python code, we observe no major differences between these results and those obtained with RevMan [Cochrane Collaboration, 2011] as listed in table 1. The minor discrepancies, especially observed at the p-values level, are due to our choice to always use a pre-post correlation of 0.5 when computing the variance of each ES. Moreover, a sensitivity analysis was conducted to ensure the minor impact of the pre-post correlation value: when it varies between 0.2 and 0.8 the significance of the Se does not change.

Table 1: Comparison between Cortese et al. [2016] results obtained with RevMan [Cochrane Collaboration, 2011] and those obtained with the Python code. Summary effects and their corresponding p-value in parenthesis are presented. With the Python program, a negative summary effect is in favor of NFB.

| Input data | | Results from Cortese et al. [2016] | Effect sizes from Cortese et al. [2016] |
|---|---|---|---|
| Implementation | | RevMan Cochrane Collaboration [2011] | Python program |
| *Parents* | Total | $0.35\ (0.004)$ | $-0.34\ (0.004)$ |
| | Inattention | $0.36\ (0.009)$ | $-0.35\ (0.011)$ |
| | Hyperactivity | $0.26\ (0.004)$ | $-0.24\ (0.02)$ |
| *Teachers* | Total | $0.15\ (0.20)$ | $-0.13\ (0.25)$ |
| | Inattention | $0.06\ (0.70)$ | $-0.09\ (0.50)$ |
| | Hyperactivity | $0.17\ (0.13)$ | $-0.15\ (0.21)$ |

Thanks to the previous step, we can conclude that the Python code yields results close to those returned by RevMan Cochrane Collaboration [2011], so all the following results were computed with the Python code.

The replication of Cortese et al. [2016] was conducted by applying the following choices and the results obtained are presented in table 2:

To compute the ES for Arnold et al. [2014], Cortese et al. [2016] took as post-test values the assessments after 12 NFB sessions because results at post-test were not available. In our case, we used the values at post-test (i.e after the 40 sessions). With these values, we find smaller ES than Cortese et al. [2016].

Different results for teachers' assessments were found for Steiner et al. [2014] because we decided not to use the same scale as Cortese et al. [2016]. Indeed, Cortese et al. [2016] relied on the BOSS Classroom Observation [Shapiro, 2010] to compute ES for teachers' ratings even if this scale is not as used as other scales provided in this study. That's why we decided to conduct our analysis with a more common scale which has been part of studies assessing the pros and cons of different ADHD scales [Epstein and Weiss, 2012; Collett et al., 2003]: The Conners. Besides, this scale was already used in this study to compute the ES for the parents' ratings. Using this scale leads to higher ES in attention but not in total and hyperactivity. Moreover, this different choice of scale does not affect the significance of the summary effects.

Eventually, there is a slight overall difference between ES computed by Cortese et al. [2016] and those yielded by our Python code maybe due to the fact that the correcting factor for small sample $cp$ is not applied in our work. Besides, as mentioned earlier, there is a little difference in some ES' standard error explained by the use of a pre-post correlation value of 0.5 while computing the variance of the ES. These two discrepancies does not change the significance of the summary effect.

Table 2: Comparison between Cortese et al. [2016] results obtained with RevMan [Cochrane Collaboration, 2011] and those obtained with the Python code with our choices applied. Summary effects and their corresponding p-value (in parenthesis) are presented. With the Python program, a negative summary effect is in favor of NFB.

| Input data | | Results from Cortese et al. [2016] | Means and standard deviations from articles included in Cortese et al. [2016] |
|---|---|---|---|
| Implementation | | RevMan Cochrane Collaboration [2011] | Python program |
| Hypothesis | | Same as Cortese et al. [2016] | Our choices |
| *Parents* | Total | $0.35$ $(0.004)$ | $-0.32$ $(0.013)$ |
| | Inattention | $0.36$ $(0.009)$ | $-0.31$ $(0.036)$ |
| | Hyperactivity | $0.26$ $(0.004)$ | $-0.24$ $(0.02)$ |
| *Teachers* | Total | $0.15$ $(0.20)$ | $-0.11$ $(0.37)$ |
| | Inattention | $0.06$ $(0.70)$ | $-0.17$ $(0.16)$ |
| | Hyperactivity | $0.17$ $(0.13)$ | $-0.022$ $(0.85)$ |

## 3.3 Update a meta-analysis

The next step consisted in extend Cortese et al. [2016] meta-analysis by adding the two new articles [Strehl et al., 2017; Baumeister et al., 2016] found during the systematic review. Baumeister et al. [2016] provided results only for parents total outcome whereas Strehl et al. [2017] gave teachers and parents' assessments for all outcomes. In spite of favorable results for NFB, particularly on parents' assessments, adding these two new studies does not change either the magnitude or the significance of the summary effect for any outcome whatever the raters or the choices made as summed up in table 3.

Table 3: Comparison between Cortese et al. [2016] results obtained with RevMan [Cochrane Collaboration, 2011] with and without the two new articles and those obtained with the Python code with our choices applied plus the two new articles. Summary effects and their corresponding p-value (in parenthesis) are presented. With the Python program, a negative summary effect is in favor of NFB.

| Input data | | Results from Cortese et al. [2016] | Results from Cortese et al. [2016] and means and standard deviations from the new two articles | Means and standard deviations from articles included in Cortese et al. [2016] and the two new studies |
|---|---|---|---|---|
| Implementation | | RevMan Cochrane Collaboration [2011] | Effect sizes: RevMan Cochrane Collaboration [2011] and then Python program | Python program |
| Hypothesis | | Same as Cortese et al. [2016] | Same as Cortese et al. [2016] | Our choices |
| *Parents* | Total | 0.35 (0.004) | 0.37 (0.0004) | −0.34 (0.0017) |
| | Inattention | 0.36 (0.009) | 0.37 (0.0015) | −0.33 (0.011) |
| | Hyperactivity | 0.26 (0.004) | 0.24 (0.002) | −0.23 (0.0094) |
| *Teachers* | Total | 0.15 (0.20) | 0.14 (0.15) | −0.11 (0.27) |
| | Inattention | 0.06 (0.70) | 0.06 (0.66) | −0.14 (0.17) |
| | Hyperactivity | 0.17 (0.13) | 0.16 (0.96) | −0.05 (0.6) |

Next, we ran the analysis on two specific subgroups: on the one hand only studies following standard protocol defined by Arns et al. [2014] are selected and on the other hand only studies forbidding participants to take medication during the clinical trial are included.

Regarding the standard protocol subgroup, Cortese et al. [2016] found all the outcomes significant except for the hyperactivity symptoms rated by teachers. However, when adding [Strehl et al., 2017] results, we find no significance for the inattention symptoms assessed by teachers as well (p-value = 0.10 when following Cortese et al.'s choices and 0.11 with our choices). As for the low/no medication subgroup, summary effects are not significant except for the inattention symptoms assessed by parents (p-value = 0.013). Besides, the differences in Arnold et al. [2014] values causes a loss of significance in hyperactivity outcome for parents (p-value = 0.066) compared to Cortese et al. [2016] (p-value = 0.016). The two new studies were not included in this subgroup because subjects are taking psychostimulants during the trial.

All the scales used to compute the effect sizes following our choices are summarized in Supplemental material.

## 3.4 Detect factors influencing the Neurofeedback

This analysis was performed on 31 trials assessing the efficacy of NFB as presented in table 4. Among the 25 factors selected, 6 were removed because there were either too many missing observations or they were too homogeneous: beta up frontal, the use of a transfer card, the type of threshold for the rewards (incremental or fixed), the EEG quality equal to 3 and presence of a control group.

To assess the variability of each factor, box plots of their standardized values were displayed in fig. 2: treatment length, session length and number of sessions are more variable across studies than session pace, minimum and maximum age.

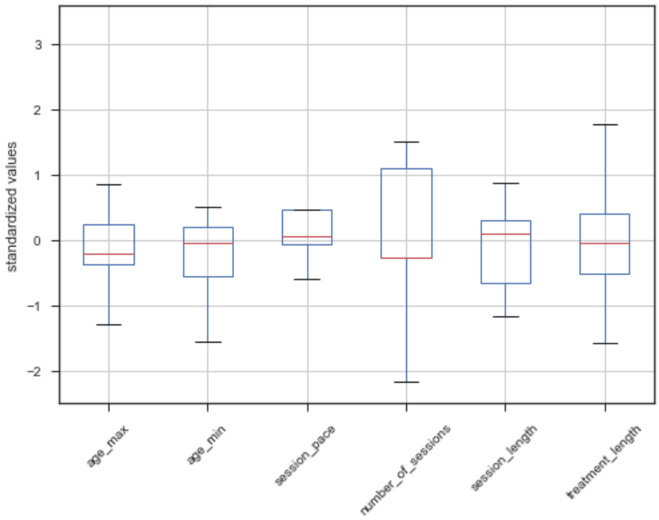

Figure 2: Boxplots of the standardized values of each continuous factor.

Table 4: List of included studies to perform the factors analysis.

| Trial | Year | Size of the NFB group | Weight |
|---|---|---|---|
| Arnold et al. | 2014 | 26 | 6.5 |
| Bakhshayesh et al. | 2011 | 18 | 9 |
| Baumeister et al. | 2016 | 8 | 8 |
| Bakhshayesh et al. | 2011 | 18 | 9 |
| Beauregard and Levesque | 2006 | 15 | 15 |
| Bink et al. | 2014 | 45 | 15 |
| Bluschke et al. | 2016 | 19 | 19 |
| Christiansen et al. | 2014 | 14 | 7 |
| Deilami et al. | 2016 | 12 | 12 |
| Drechsler et al. | 2007 | 17 | 5.67 |
| Duric et al. | 2012 | 23 | 23 |
| Escolano et al. | 2014 | 20 | 20 |
| Gevensleben et al. | 2009 | 59 | 14.75 |
| Heinrich et al. | 2004 | 13 | 13 |
| Holtmann et al. | 2009 | 20 | 20 |
| Kropotov et al. | 2005 | 86 | 86 |
| Lee and Jung | 2017 | 18 | 6 |
| Leins et al. | 2007 | 19 | 19 |
| Li et al. | 2013 | 20 | 10 |
| Linden et al. | 1996 | 9 | 9 |
| Maurizio et al. | 2014 | 13 | 3.25 |
| Meisel et al. | 2014 | 12 | 6 |
| Mohagheghi et al. | 2017 | 30 | 30 |
| Mohammadi et al. | 2015 | 16 | 16 |
| Monastra et al. | 2002 | 51 | 51 |
| Ogrim and Hestad | 2013 | 13 | 6.50 |
| Steiner et al. | 2011 | 9 | 2.25 |
| Steiner et al. | 2014 | 34 | 6.80 |
| Strehl et al. | 2006 | 23 | 5.75 |
| Strehl et al. | 2017 | 72 | 16.25 |
| van Dongen-Boomsma et al. | 2013 | 22 | 11 |

The first method used to detect the influencing factors was the WLS. First the assumptions inherent to this method were checked:

the moment matrix $\mathbf{X^T W^T W X}$ was invertible;

no apparent correlation between the continuous independent variables was found;

the fit was significant as shown by the F-statistic (prob(F-statistic) = 2.21e-10);

the residuals were normally distributed as demonstrated by the skew (-0.166), kurtosis (2.85) and the Omnibus test (prob(Omnibus) = 0.84).

Since these assumptions are satisfied, we can be rather confident in the WLS results presented in table 5. We find that 8 factors are significantly different from zero for an adjusted R-squared of 0.74. When applying the OLS the assumptions of the model are satisfied too and the same factors are returned as significant except the transfer phase, the protocol theta down and the artifact correction based on amplitude with an adjusted R-squared of 0.42.

A negative coefficient meant that the factor was in favor of the NFB. Here, among the factors whose coefficient is significantly different from 0, a long treatment, a blind rater, including a transfer phase, and correcting artifact thanks to the amplitude appear to have a negative influence on the NFB performance. Conversely, an IRB approval,

Table 5: Results of the WLS. A p-value $< 0.05$ means that the coefficient of the corresponding factor is significantly different from $0$ (in bold). When the value of the coefficient is negative, the corresponding factor may lead to better NFB results.

| Independent variables (factors) | | Value of the coefficient | p-value |
|---|---|---|---|
| *Methodological* | age max | $-0.087$ | $0.17$ |
| | age min | $-0.050$ | $0.43$ |
| | **Pblind** | $0.11$ | $0.038$ |
| | on drugs | $0.065$ | $0.45$ |
| | randomization | $0.0099$ | $0.89$ |
| | **IRB** | $-0.30$ | $0.00$ |
| *Technical* | number of sessions | $-0.015$ | $0.84$ |
| | session length | $0.17$ | $0.18$ |
| | **session pace** | $-0.25$ | $0.00$ |
| | **treatment length** | $0.57$ | $0.00$ |
| | SMR | $-0.067$ | $0.38$ |
| | beta up central | $-0.025$ | $0.73$ |
| | **theta down** | $-0.29$ | $0.019$ |
| | SCP | $-0.90$ | $0.54$ |
| | **transfer phase** | $0.27$ | $0.030$ |
| | electro-oculogram (EOG) correction | $-0.081$ | $0.40$ |
| | **artifact correction based on amplitude** | $0.15$ | $0.039$ |
| *Quality of acquisition* | more than one active electrode | $0.0639$ | $0.359$ |
| | **EEG quality 2** | $-0.36$ | $0.00$ |

a high number of sessions per week, a theta down protocol and an EEG quality of 2 seemed to lead to more efficacy.

Next, a LASSO regression was conducted in order to perform variables selection among the factors. As the first step, the tuning parameter $\lambda$ was obtained by leave one out cross validation. $\lambda$ corresponds to the abscissa of the minimum of the MSE on the mean fold computed on $\lambda$ values ranging from very small to very big, covering the full range of scenarios from the null model to the least square fit.

The penalty applied leads to the selection of 12 factors as summed up in table 6.

With this method, a negative coefficient corresponds also to favorable factors. Thus, having several sessions per week, being approved by an IRB, a protocol theta down, and an EEG quality of 2 seem to improve the results of the NFB. On the contrary, a blind rater, a long treatment, randomizing the groups, a SCP and SMR protocols, being on drugs during the trial, correcting artifact based on the amplitude of the signal, and including a transfer phase during the session appear to be factors

Table 6: Results of the LASSO. Factors different from 0 are selected (in bold). When the value of the coefficient is negative, the corresponding factor may lead to better NFB results.

| Independent variables (factors) | | Value of the coefficient |
|---|---|---|
| *Methodological* | age max | $0.00$ |
| | age min | $0.00$ |
| | **Pblind** | $0.11$ |
| | **on drugs** | $0.032$ |
| | **randomization** | $0.032$ |
| | **IRB** | $-0.15$ |
| *Technical* | number of sessions | $0.00$ |
| | session length | $0.00$ |
| | **session pace** | $-0.14$ |
| | **treatment length** | $0.33$ |
| | **SMR** | $0.061$ |
| | beta up central | $0.00$ |
| | **theta down** | $-0.051$ |
| | **SCP** | $0.10$ |
| | **transfer phase** | $0.11$ |
| | EOG correction | $0.00$ |
| | **artifact correction based on amplitude** | $0.047$ |
| *Quality of acquisition* | more than one active electrode | $0.00$ |
| | **EEG quality 2** | $-0.23$ |

with a negative influence.

Eventually, the decision tree presented in fig. 3 splits the dataset based on the factor leading to the smallest MSE. The best predictor is the one at the top of the tree: in our case it is the Pblind. So, the dataset is divided in two subsets: 43 samples where the assessments were reported by non-blind raters and 19 samples based on blind ratings. This last subset, where the Se is smaller than the one obtained for non-blind raters, is split on the age min criteria: it appeares that the lower the age, the better the results. Regarding the side of the tree where the raters are non-blind, the next factor leading to the smallest MSE is the treatment length. A smaller treatment length seems to lead to better NFB results. In that case, the EEG quality 2 enables to break down the subset and according to these results if this factor is respected in studies it seems to lead to higher Se. The lower we get into the tree, the less samples are available, so results were less and less reliable.

The different methods do not detect exactly the same factors but many are common as presented in table 7, in particular the treatment length, the assessment by a blind rater and and EEG quality of 2 are returned by the three methods. Besides, the

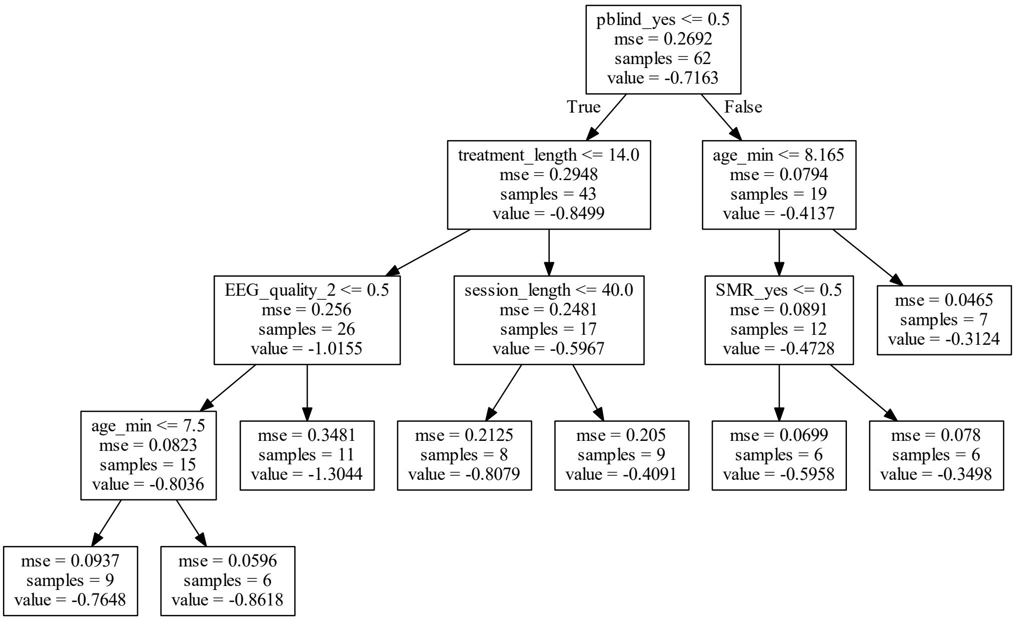

Figure 3: Decision Tree obtained with the factors. The criteria to minimize was the MSE. For the dummy variables, a value of 1 meant that the factor was observed in the study. The value corresponded to the value of the dependent variable.

methods agree on the direction of the influence of these factors. However, it is more difficult to interpret the influence of the factors returned only by one or two methods. For instance, both WLS and LASSO find that relying on the amplitude of the signal to correct artifacts, and including a transfer phase seem not to improve the ADHD symptoms. However, the IRB approval, a theta down protocol, and a high number of sessions per week appear to positively influence the results. The decision tree and LASSO have in common the protocol SMR: it is associated with lower ES. Five factors are returned by only one of the methods: the minimal age of the children, being on drugs during NFB treatment, randomizing the groups and the SCP protocol.

# 4 Discussion

## 4.1 Replicate and update a meta-analysis

In this replication and extension of Cortese et al. [2016], we investigated choices made by Cortese et al., which proved controversial: the computation of ES based on an unusual scale [Steiner et al., 2014] and the inclusion of a pilot study [Arnold et al., 2014] whose final results weren't available at the time Cortese et al. conducted his meta-analysis. We review here the list of changes, their justification, and their impact on the analysis.

First, relying on the Conners-3 [Conners, 2008] instead of the BOSS Classroom

Table 7: Summary of the results obtained with the three methods. The type of sign describes the direction of the influence on the NFB results (+ for a positive effect and a − for a negative one). The number of signs illustrates the number of methods who finds the factor as influencing. Factors returned by 3, 2 and 1 methods influence the NFB with respectively high, moderate and poor confidence.

| Independent variables (factors) | | Influence on the NFB |
|---|---|---|
| *Methodological* | age max | 0 |
| | age min | + |
| | **Pblind** | - - - |
| | on drugs | - |
| | randomization | + |
| | IRB | ++ |
| *Technical* | number of sessions | 0 |
| | session length | - |
| | **treatment length** | - - - |
| | session pace | ++ |
| | SMR | - - |
| | beta up central | 0 |
| | theta down | ++ |
| | SCP | - |
| | transfer phase | - - |
| | EOG correction | 0 |
| | artifact correction based on amplitude | - - |
| *Quality of acquisition* | more than one active electrode | 0 |
| | **EEG quality 2** | +++ |

Observation [Shapiro, 2010] for teachers ratings seems preferable because this scale is more often used [Christiansen et al., 2014; Bluschke et al., 2016] and is the revision of the Conners Rating Scale Revised [Conners et al., 1998] whose reliability has been studied [Collett et al., 2003]. However, relying on one or the other scale did not change the significance of the sES whatever the outcome studied.

Second, to compute the ES of Arnold et al. [2014] we used the values at post-test whereas Cortese et al. calculated it with ratings after only 12 sessions of NFB. Some studies suggest that the number of sessions positively correlates with the changes in the EEG [Vernon et al., 2004] so a lower number of sessions would lead to artificially smaller ES. Here, the ES computed with the values at post test of Arnold et al. [2014] are smaller than those obtained after 12 sessions but these differences do not lead to a change of significance of the Se.

Finally, we found that some of the studies included in Cortese et al.'s work [Arnold et al., 2014] and [Steiner et al., 2011] were described by their respective authors as *pilot* studies and it was disclosed in van Dongen-Boomsma et al. [2013]; van Dongen-Boomsma M et al. [2015]

> [they were] unable to recruit a sufficient number of participants to meet [their] planned sample size .

Nevertheless not including them would introduce bias.

To conclude on the replication of Cortese et al. [2016], though some of the choices made by authors were controversial and the fact that - for the reasons mentioned earlier - different choices could reasonably be made, it turns out that the impact on the meta-analysis results are minimal and do not change the statistical significance of any outcome. Consequently, the completion of the meta-analysis with studies published since the publication of his work were done with the choices:

To compute the ES of Arnold et al. [2014] the values at post-test were used.

The scores reported by teachers on the Conners-3 in Steiner et al.'s study were taken into account instead of these of the BOSS Classroom Observation.

Results were obtained with the Python code.

The addition of the two new studies [Strehl et al., 2017; Baumeister et al., 2016] further confirms those results. Indeed, the significance does not change for any outcome: Se found remains significant for parents' scores and non-significant for teachers.

Adding two more studies increases the significance of the sensitivity analysis ran by Cortese et al.. Most interestingly, the ES from the subset of studies corresponding to standard protocols of NFB as defined by Arns et al. [2014]. While Cortese et al. found that this subset tend to perform better, particularly on the Pblind outcome, adding two studies confirms this result on the total score with a p-value of 0.043. Despite the obvious heterogeneity of the studies included in this subset (particularly in terms of protocol used), this result suggests a positive relation between the features of this *standard* design and NFB performance.

Eventually, concerning the raters, we considered teachers as Pblind raters as Cortese et al. and Micoulaud-Franchi et al. did although they may be aware of the treatment followed thanks to the parents. Besides, the amplitude of the clinical scale at baseline suggests that teachers do not capture the full picture of the condition and are therefore less likely to see a change (prone to type II error).

Along with this article, the Python code and raw data are provided in order to facilitate a potential replication of this work (available on the Github repository).

## 4.2 Identify factors influencing the Neurofeedback

Description and analysis of NFB implementation was subject to several studies [Arns et al., 2014; Enriquez-Geppert et al., 2017; Vernon et al., 2004] but to our knowledge none used statistical tools to detect the influence of methodological, clinical and technical factors on such a wide range of studies.

A somewhat puzzling result is the fact that the three methods which offer to identify factors contributing to the NFB performance do not lead to the exact same results. These discrepancies are clearly explained by the varying hypothesis of these models and actually offer interesting insight into the results and their significance. For instance, the decision tree method is non linear and accounts for variables interaction which is not the case for the two others methods. Moreover, the decision tree is unstable [Dwyer and Holte, 2007], meaning that a small change in the data can cause an important change in the structure of the optimal decision tree.

Nevertheless, despite these differences between the methods, 3 factors are consistently identified by all the methods with the same influence direction: if the rater is probably blind to the treatment, the treatment length, and the EEG quality.

As expected, the assessment of symptoms by non-blind raters leads to more favorable results than by blind raters, result observed in several meta-analysis [Cortese et al., 2016; Micoulaud-Franchi et al., 2014]. If the differences observed between blind and non-blind raters are due to the placebo effect, the part of the decision tree where there is only observations of non-blind raters may enable us to detect the factors affected by the placebo effect. Thus, two factors linked to the perception of the treatment (treatment length and session length) are present in this part of the tree whereas no such factors are returned for the Pblind part of the tree.

The treatment length varied more than the session pace and the age bounds of included children between the included studies as shown by the boxplot of standardized values fig. 2, so detecting it as an influencing factor may have been easier. It appears here that the longer the treatment the less efficient it becomes. Arguably, the treatment length is a proxy for treatment intensity, which means that a treatment that is short in length (and consequently intense in pace) is more likely to succeed. This hypothesis is back-up by the fact that the variable *session pace* (number of sessions per week) is also associated with larger ES according to the WLS and LASSO.

Eventually, this analysis points out the fact that recording EEG in good conditions seems to lead to better results, which can be explained by the fact that better signal quality enables to extract more correctly the EEG patterns linked to ADHD and henceforth leads to better learning and therapeutic efficacy. However, it remains difficult to really assess the quality of the hardware because little information is provided in the studies.

The interpretation of factors returned only by two methods was less reliable. Keeping that in mind, it was interesting to study the direction of their effect though: an IRB approval seems preferable implying that a well-conducted study seems lead to favorable results; a high session pace, so a more intensive training, appears to lead to better results, and the artifact correction based on amplitude do not seem to be an appropriate method to remove the artifacts and the SMR protocol appears not to be the best approach.

Surprisingly, the number of sessions is not found as an significant influencing factor by any method, which is somewhat in contradiction with existing literature. For instance, Enriquez-Geppert et al. [2017] insisted on the fact that the number of sessions should be chosen carefully to avoid "overtraining". Moreover, Arns et al. [2014] stated that performing less than 20 NFB sessions lead to smaller effects. Indeed Vernon2004 observed that positive changes in the EEG and behavioral performance occurred after a minimum of 20 sessions, but he also points out the fact that the location of the NFB training may had an important influence. Nevertheless, in our study, regardless of the significance of the number of sessions, the coefficient found by the WLS is negative, meaning that as expected, the more sessions performed the more efficient the NFB seem to be.

According to our analysis, the type of protocol does not seem to influence the NFB results except for the the theta down which appeared more efficient and SMR which conversely seems associated with lower ES. We expected more precised results on the protocols criteria because this point is central in NFB as pointed out by Vernon et al. [2004]. A possible explanation is that all these protocols are equally efficacious to the populations they were offered to and thereby do not constitute a significant explanatory factor. This result, however, does not preclude a combined and personalized strategy (offer the right protocol to the right kid) to further improve performance.

It would have been interesting to study the influence of some other factors such as the delay between brain state and feedback signal as well as the type of NFB game used but these information are rarely available in studies and also . Besides, to add more reliability to these results it should be preferable to add more studies, particularly studies with teachers assessments (considered as Pblind).

# 5 Disclosure statement

We report no potential conflicts of interest.

# 6 Acknowledgments

We would like to thank Dr. Quentin Barthelemy and Dr. David Ojeda for their helpful comments and ideas on that work.

# References

L. E. Arnold, N. Lofthouse, S. Hersch, X. Pan, E. Hurt, B. Bates, K. Kassouf, S. Moone, and C. Grantier. Eeg neurofeedback for attention-deficit/hyperactivity disorder: Double-blind sham-controlled randomized pilot feasibility trial. *Journal of Atention Disorder*, 2014.

M. Arns, S. de Ridder, U. Strehl, M. Breteler, and A. Coenen. Efficacy of neurofeedback treatment in adhd: the effects on inattention, impulsivity and hyperactivity: a meta-analysis. *Clinical EEG and neuroscience*, 40(3):180–189, 2009.

M. Arns, C. K. Conners, and H. C. Kraemer. A decade of eeg theta/beta ratio research in adhd: a meta-analysis. *Journal of attention disorders*, 17(5):374–383, 2013.

M. Arns, H. Heinrich, and U. Strehl. Evaluation of neurofeedback in adhd: the long and winding road. *Biological psychology*, 95:108–115, 2014.

M. Arns, H. Heinrich, T. Ros, A. Rothenberger, and U. Strehl. Neurofeedback in adhd. *Frontiers in human neuroscience*, 9:602, 2015.

A. P. Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

T. A. P. Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-4)*, 4 edition, 2000.

T. A. P. Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*, 5 edition, 2013.

A. R. Bakhshayesh, S. Hänsch, A. Wyschkon, M. J. Rezai, and G. Esser. Neurofeedback in adhd: a single-blind randomized controlled trial. *European child & adolescent psychiatry*, 20(9):481, 2011.

E. M. Balk, A. Earley, K. Patel, T. A. Trikalinos, and I. J. Dahabreh. Empirical assessment of within-arm correlation imputation in trials of continuous outcomes. 2012.

T. Banaschewski and D. Brandeis. Annotation: what electrical brain activity tells us about brain function that other techniques cannot tell us–a child psychiatric perspective. *Journal of Child Psychology and Psychiatry*, 48(5):415–435, 2007.

S. Baumeister, I. Wolf, N. Holz, R. Boecker-Schlier, N. Adamo, M. Holtmann, M. Ruf, T. Banaschewski, S. Hohmann, and D. Brandeis. Neurofeedback training effects on inhibitory brain activation in adhd: A matter of learning? *Neuroscience*, 2016.

M. Beauregard and J. Levesque. Functional magnetic resonance imaging investigation of the effects of neurofeedback training on the neural bases of selective attention and response inhibition in children with attention-deficit/hyperactivity disorder. *Applied psychophysiology and biofeedback*, 31(1):3–20, 2006.

M. Bink, C. van Nieuwenhuizen, A. Popma, I. L. Bongers, and G. J. M. van Boxtel. Behavioral effects of neurofeedback in adolescents with adhd: a randomized controlled trial. *European Child and Adolescent Psychiatry*, 2014.

A. Bluschke, F. Broschwitz, S. Kohl, V. Roessner, and C. Beste. The neuronal mechanisms underlying improvement of impulsivity in adhd by theta/beta neurofeedback. *Scientific reports*, 6:31178, 2016.

M. Borenstein, L. V. Hedges, J. Higgins, and H. R. Rothstein. *Introduction to meta-analysis*. Wiley, 2009.

H. Christiansen, V. Reh, M. H. Schmidt, and W. Rief. Slow cortical potential neurofeedback and self-management training in outpatient care for children with adhd: study protocol and first preliminary results of a randomized controlled trial. *Frontiers in human neuroscience*, 8, 2014.

Cochrane Collaboration. Revman 5.1, 2011.

J. Cohen. *Statistical power analysis for the behavioral sciences 2nd edn*. Erlbaum Associates, Hillsdale, 1988.

B. R. Collett, J. L. Ohan, and K. M. Myers. Ten-year review of rating scales. v: scales assessing attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42(9):1015–1037, 2003.

C. K. Conners. Conners 3. MHS, 2008.

C. K. Conners, G. Sitarenios, J. D. Parker, and J. N. Epstein. The revised conners' parent rating scale (cprs-r): factor structure, reliability, and criterion validity. *Journal of abnormal child psychology*, 26(4):257–268, 1998.

S. Cortese, M. Ferrin, D. Brandeis, M. Holtmann, P. Aggensteiner, D. Daley, P. Santosh, E. Simonoff, J. Stevenson, A. Stringaris, et al. Neurofeedback for attention-deficit/hyperactivity disorder: meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials. *Journal of the American Academy of Child & Adolescent Psychiatry*, 55(6):444–455, 2016.

M. Deilami, A. Jahandideh, Y. Kazemnejad, Y. Fakour, S. Alipoor, F. Rabiee, G. S. Pournesaie, R. N. Heidari, and S. A. Mosavi. The effect of neurofeedback therapy on reducing symptoms associated with attention deficit hyperactivity disorder: A case series study. *Basic and clinical neuroscience*, 7(2):167, 2016.

R. Drechsler, M. Straub, M. Doehnert, H. Heinrich, H.-C. Steinhausen, and D. Brandeis. Controlled evaluation of a neurofeedback training of slow cortical potentials in children with attention deficit/hyperactivity disorder (adhd). *Behavioral and brain functions*, 3(1):35, 2007.

N. S. Duric, J. Assmus, D. Gundersen, and I. B. Elgen. Neurofeedback for the treatment of children and adolescents with adhd: a randomized and controlled clinical trial using parental reports. *BMC psychiatry*, 12(1):107, 2012.

K. Dwyer and R. Holte. Decision tree instability and active learning. In *European Conference on Machine Learning*, pages 128–139. Springer, 2007.

S. Enriquez-Geppert, R. J. Huster, and C. S. Herrmann. Eeg-neurofeedback as a tool to modulate cognition and behavior: a review tutorial. *Frontiers in human neuroscience*, 11:51, 2017.

J. N. Epstein and M. D. Weiss. Assessing treatment outcomes in attention-deficit/hyperactivity disorder: a narrative review. *The primary care companion for CNS disorders*, 14(6), 2012.

C. Escolano, M. Navarro-Gil, J. Garcia-Campayo, M. Congedo, and J. Minguez. The effects of individual upper alpha neurofeedback in adhd: an open-label pilot study. *Applied psychophysiology and biofeedback*, 39(3-4):193–202, 2014.

S. V. Faraone, J. Biederman, and E. Mick. The age-dependent decline of attention deficit hyperactivity disorder: a meta-analysis of follow-up studies. *Psychological medicine*, 36(2):159–165, 2006.

M. Fouillen, E. Maby, L. Le Career, V. Herbillon, and J. Mattout. Erp-based bci for children with adhd. 2017.

H. Gevensleben, B. Holl, B. Albrecht, C. Vogel, D. Schlamp, O. Kratz, P. Studer, A. Rothenberger, G. H. Moll, and H. Heinrich. Is neurofeedback an efficacious treatment for adhd? a randomised controlled clinical trial. *Journal of Child Psychology and Psychiatry*, 50(7):780–789, 2009.

H. Heinrich, H. Gevensleben, F. J. Freisleder, G. H. Moll, and A. Rothenberger. Training of slow cortical potentials in attention-deficit/hyperactivity disorder: evidence for positive behavioral and neurophysiological effects. *Biological psychiatry*, 55(7):772–775, 2004.

M. Holtmann, D. Grasmann, E. Cionek-Szpak, V. Hager, N. Panzner, A. Beyer, F. Poustka, and C. Stadler. Spezifische wirksamkeit von neurofeedback auf die impulsivität bei adhs. *Kindheit und Entwicklung*, 18(2):95–104, 2009.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

T. Janzen, K. Graap, S. Stephanson, W. Marshall, and G. Fitzsimmons. Differences in baseline eeg measures for add and normally achieving preadolescent males. *Biofeedback and self-Regulation*, 20(1):65–82, 1995.

J. D. Kropotov, V. A. Grin-Yatsenko, V. A. Ponomarev, L. S. Chutko, E. A. Yakovenko, and I. S. Nikishena. Erps correlates of eeg relative beta training in adhd children. *International journal of psychophysiology*, 55(1):23–34, 2005.

E.-J. Lee and C.-H. Jung. Additive effects of neurofeedback on the treatment of adhd: A randomized controlled study. *Asian Journal of Psychiatry*, 25:16–21, 2017.

U. Leins, G. Goth, T. Hinterberger, C. Klinger, N. Rumpf, and U. Strehl. Neurofeedback for children with adhd: a comparison of scp and theta/beta protocols. *Applied psychophysiology and biofeedback*, 32(2):73–88, 2007.

L. Li, L. Yang, C. Zhuo, and Y.-F. Wang. A randomised controlled trial of combined eeg feedback and methylphenidate therapy for the treatment of adhd. *Swiss Med. Wkly*, 143:w13838, 2013.

M. Linden, T. Habib, and V. Radojevic. A controlled study of the effects of eeg biofeedback on cognition and behavior of children with attention deficit disorder and learning disabilities. *Applied Psychophysiology and Biofeedback*, 21(1):35–49, 1996.

N. Lofthouse, L. E. Arnold, S. Hersch, E. Hurt, and R. DeBeus. A review of neurofeedback treatment for pediatric adhd. *Journal of Attention Disorders*, 2012.

S. K. Loo and R. A. Barkley. Clinical utility of eeg in attention deficit hyperactivity disorder. *Applied Neuropsychology*, 12:64–76, 2005.

S. K. Loo, J. J. McGough, J. T. McCracken, and S. L. Smalley. Parsing heterogeneity in attention-deficit hyperactivity disorder using eeg-based subgroups. *Journal of Child Psychology and Psychiatry*, 2017.

J. F. Lubar and M. N. Shouse. Eeg and behavioral changes in a hyperkinetic child concurrent with training of the sensorimotor rhythm (smr). *Biofeedback and Self-regulation*, 1(3):293–306, 1976.

M. Matoušek, P. Rasmussen, and C. Gillberg. Eeg frequency analysis in children with so-called minimal brain dysfunction and related disorders. In *Neurophysiological Correlates of Mental Disorders*, volume 15, pages 102–108. Karger Publishers, 1984.

S. Maurizio, M. D. Liechti, H. Heinrich, L. Jäncke, H.-C. Steinhausen, S. Walitza, D. Brandeis, and R. Drechsler. Comparing tomographic eeg neurofeedback and emg biofeedback in children with attention-deficit/hyperactivity disorder. *Biological psychology*, 95:31–44, 2014.

V. Meisel, M. Servera, G. Garcia-Banda, E. Cardo, and I. Moreno. Reprint of "neurofeedback and standard pharmacological intervention in adhd: a randomized controlled trial with six-month follow-up". *Biological psychology*, 95:116–125, 2014.

J.-A. Micoulaud-Franchi, P. A. Geoffroy, G. Fond, R. Lopez, S. Bioulac, and P. Philip. Eeg neurofeedback treatments in children with adhd: an updated meta-analysis of randomized controlled trials. *Frontiers in human neuroscience*, 8, 2014.

J.-A. Micoulaud-Franchi, F. Salvo, S. Bioulac, and T. Fovet. Neurofeedbach in attention-deficit/hyperactivity disorder: Efficacy. *Journal of the American Academy of Child & Adolescent Psychiatry*, 2016.

A. Mohagheghi, S. Amiri, N. Moghaddasi Bonab, G. Chalabianloo, S. G. Noorazar, S. M. Tabatabaei, and S. Farhang. A randomized trial of comparing the efficacy of two neurofeedback protocols for treatment of clinical and cognitive symptoms of adhd: Theta suppression/beta enhancement and theta suppression/alpha enhancement. *BioMed Research International*, 2017, 2017.

M. R. Mohammadi, N. Malmir, A. Khaleghi, and M. Aminiorani. Comparison of sensorimotor rhythm (smr) and beta training on selective attention and symptoms in children with attention deficit/hyperactivity disorder (adhd): A trend report. *Iranian journal of psychiatry*, 10(3):165, 2015.

V. J. Monastra. Electroencephalographic biofeedback (neurotherapy) as a treatment for attention deficit hyperactivity disorder: rationale and empirical foundation. *Child and Adolescent Psychiatric Clinics of North America*, 14(1):55–82, 2005.

V. J. Monastra, D. M. Monastra, and S. George. The effects of stimulant therapy, eeg biofeedback, and parenting style on the primary symptoms of attention-deficit/hyperactivity disorder. *Applied psychophysiology and biofeedback*, 27(4): 231–249, 2002.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

S. B. Morris. Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11(2):364–386, 2008.

G. Ogrim and K. A. Hestad. Effects of neurofeedback versus stimulant medication in attention-deficit/hyperactivity disorder: a randomized pilot study. *Journal of child and adolescent psychopharmacology*, 23(7):448–457, 2013.

W. H. Organization. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, volume 2. World Health Organization, 1993.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

D. T. R. Rossiter and T. J. La Vaque. A comparison of eeg biofeedback and psychostimulants in treating attention deficit/hyperactivity disorders. *Journal of Neurotherapy*, 1(1):48–59, 1995.

S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. SciPy society Austin, 2010.

E. S. Shapiro. *Academic skills problems fourth edition workbook*. Guilford Press, 2010.

E. J. Sonuga-Barke, D. Brandeis, S. Cortese, D. Daley, M. Ferrin, M. Holtmann, J. Stevenson, M. Danckaerts, S. van der Oord, M. Döpfner, R. W. Dittmann, E. Simonoff, A. Zuddas, T. Banaschewski, J. Buitelaar, D. Coghill, C. Hollis, E. Konofal, M. Lecendreux, I. C. Wong, and J. Sergeant. Nonpharmacological interventions for adhd: Systematic review and meta-analyses of randomized controlled trials of dietary and psychological treatments. *American Journal of Psychiatry*, 2013.

B. Steffert and T. Steffert. Neurofeedback and adhd. *ADHD in practice*, 2(1):16–19, 2010.

N. J. Steiner, R. C. Sheldrick, D. Gotthelf, and E. C. Perrin. Computer-based attention training in the schools for children with attention deficit/hyperactivity disorder: a preliminary trial. *Clinical pediatrics*, 50(7):615–622, 2011.

N. J. Steiner, E. C. Frenette, K. M. Rene, R. T. Brennan, and E. C. Perrin. Neurofeedback and cognitive attention training for children with attention-deficit hyperactivity disorder in schools. *Journal of Developmental & Behavioral Pediatrics*, 35(1):18–27, 2014.

U. Strehl, U. Leins, G. Goth, C. Klinger, T. Hinterberger, and N. Birbaumer. Self-regulation of slow cortical potentials: a new treatment for children with attention-deficit/hyperactivity disorder. *Pediatrics*, 118(5):e1530–e1540, 2006.

U. Strehl, P. Aggensteiner, D. Wachtlin, D. Brandeis, B. Albrecht, M. Arana, C. Bach, T. Banaschewski, T. Bogen, A. Flaig-Rohr, et al. Neurofeedback of slow cortical potentials in children with attention-deficit/hyperactivity disorder: A multicenter randomized trial controlling for unspecific effects. *Frontiers in human neuroscience*, 11, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

M. van Dongen-Boomsma, M. A. Vollebregt, D. Slaats-Willemse, and J. K. Buitelaar. A randomized placebo-controlled trial of electroencephalographic (eeg) neurofeedback in children with attention-deficit/hyperactivity disorder. *J Clin Psychiatry*, 74(8): 821–827, 2013.

van Dongen-Boomsma M, V. MA, S.-W. D, and B. JK. Efficacy of frequency-neurofeedback and cogmed jm-working memory training in children with adhd. *Tijdschrift voor Psychiatrie*, 2015.

J. Van Doren, H. Heinrich, M. Bezold, N. Reuter, O. Kratz, S. Horndasch, M. Berking, T. Ros, H. Gevensleben, G. H. Moll, et al. Theta/beta neurofeedback in children with adhd: Feasibility of a short-term setting and plasticity effects. *International Journal of Psychophysiology*, 112:80–88, 2017.

D. Vernon, A. Frick, and J. Gruzelier. Neurofeedback as a treatment for adhd: A methodological review with implications for future research. *Journal of Neurotherapy*, 8(2):53–82, 2004.

# 7 Supplemental material

## 7.1 Perform a meta analysis

To perform the meta-analysi several steps must be followed. First the choice of the model: this analysis is based on either one of the following statistical models [Borenstein et al., 2009]:

*The fixed-effect model*: the true ES (i.e the ES that would be observed with an infinitely large sample size) is the same for all the studies in the analysis. The differences between the actually observed ESs are due to sampling errors;

*The random-effects model*: the true ES could vary from study to study. The differences between the observed ESs are due to sampling errors but also to the various designs of the studies (for instance the number of participants or the implementation).

In the present case, although the studies included into the meta-analysis met the same criteria, they remained different from each other on various points, so the random effects model was more appropriate than the fixed-effect model.

### 7.1.1 Compute the effect size of each study

First, the scores presented in the articles were extracted and the ES of each study as defined in Morris [2008] was computed as in eq. (5):

$$\mathsf{ES} = c_p \left[ \frac{(M_{\mathsf{post},T} - M_{\mathsf{pre},T}) - (M_{\mathsf{post},C} - M_{\mathsf{pre},C})}{\mathsf{SD}_{\mathsf{pre}}} \right]. \tag{5}$$

An ES is exactly equivalent to a z-score of a standard normal distribution, it is computed as mean pre- to post-treatment score change in the NFB group ($M_{\mathsf{pre},T}$, $M_{\mathsf{post},T}$) minus the mean pre- to post- treatment score change in the control group ($M_{\mathsf{pre},C}$, $M_{\mathsf{post},C}$), divided by the pooled pretest standard deviation ($SD_{\mathsf{pre}}$) defined as eq. (6):

$$\mathsf{SD}_{\mathsf{pre}} = \sqrt{\frac{(n_T - 1)\mathsf{SD}^2_{\mathsf{pre},T} + (n_C - 1)\mathsf{SD}^2_{\mathsf{pre},C}}{n_T + n_C - 2}}, \tag{6}$$

where $\mathsf{SD}_{t,G}$ indicates the standard deviation for group $G$ at time $t$ and $n_G$ defines the sample size of each group; $c_p$ is a bias adjustment typically used for small sample sizes and defined as eq. (7):

$$\mathsf{cp} = 1 - \frac{3}{4(n_T + n_C - 2) - 1}. \tag{7}$$

The means (first statistical moments) correspond to the mean average score over all scores given by raters to assess the ADHD symptoms. The standard deviations of the means correspond to the squared root of the second statistical moment, the variance. The variance measures how far a set of numbers are spread out from their average value.

### 7.1.2 Compute the variance of each effect size

Then, the variance of each ES was computed as described in eq. (8) [Morris, 2008]:

$$\sigma^2(\mathsf{ES}) = c_p^2 \left( \frac{n_T + n_C - 2}{n_T + n_C - 4} \right) \left[ \frac{2(1-\rho)(n_T + n_C)}{n_T n_C} + \mathsf{ES}^2 \right] - \mathsf{ES}^2. \tag{8}$$

To compute the variance of the ES, the pooled within-group Pearson correlation $\rho$ (i.e the pre-post correlation) was required as described in eq. (9) [James et al., 2013]:

$$\rho = \frac{\sum_{i=1}^{n} (\mathsf{pre}_i - \mu_{\mathsf{pre}})(\mathsf{post}_i - \mu_{\mathsf{post}})}{\sqrt{\sum_{i=1}^{n} (\mathsf{pre}_i - \mu_{\mathsf{pre}})^2} \sqrt{\sum_{i=1}^{n} (\mathsf{post}_i - \mu_{\mathsf{post}})^2}}, \tag{9}$$

where $n$ is the number of patients included in a study, $pre_i$, $post_i$ are score values for patient $i$ at pre- and post-test respectively, and $\mu_{pre}$, $\mu_{post}$ the mean scores over all patients. It is a measure of linear correlation between two variables. A value of 1 means that there is a positive correlation whereas a value of -1 means a negative correlation. When $\rho = 0$, there is no linear correlation [James et al., 2013]. In our case, this correlation was not known and the raw data were not available so we took an approximation: Balk et al. [2012] found that a value of 0.5 yields values closer to those computed with the right value of the correlation.

Once variances were obtained with eq. (8), we could compute the standard error and the 95% confidence interval of each ES.

### 7.1.3 Compute the weight of each study

To compute the Se a weight must be assigned to each study. To obtain them several steps must be followed. At first, the fixed-effects model weight $w_{fixed}$ of each study $k$ was computed as defined in Borenstein et al. [2009] described in eq. (10):

$$w_{\mathsf{fixed}_k} = \frac{1}{\sigma^2(\mathsf{ES}_k)}. \tag{10}$$

Nevertheless, we chose to use the random effects model, so the weights associated to this model are different. To compute them, the between-studies variance $\tau^2$ is required. It was calculated in three steps described in eq. (11), eq. (12) and eq. (13) [Borenstein et al., 2009]:

$$Q = \sum_{k=1}^{K} (w_{\mathsf{fixed}_k} \mathsf{ES}_k^2), \tag{11}$$

$$C = \sum_{k=1}^{K} (w_{\mathsf{fixed}_k} - \frac{\sum_{k=1}^{K} (w_{\mathsf{fixed}_k})^2}{\sum_{k=1}^{K} (w_{\mathsf{fixed}_k})}, \tag{12}$$

with $K$ the total number of included studies.

$$\tau^2 = \frac{Q - \mathsf{df}}{C}, \tag{13}$$

with df $= K - 1$ the degrees of freedom.

The random-effects model takes into account the differences between the studies, so the weights are equal to the inverse of the addition between the within-study variance (the variance of the ES) and the between-studies variance as presented in eq. (14):

$$w_k = \frac{1}{\sigma^2(\mathsf{ES}_k) + \tau^2}. \tag{14}$$

### 7.1.4 Compute the summary effect

Eventually, the weighted average of the $K$ ES was computed to obtain the Se as described in eq. (15) [Borenstein et al., 2009]:

$$\mathsf{Se} = \frac{\sum_{k=1}^{K} w_k \mathsf{ES}_k}{\sum_{k=1}^{K} w_k}. \tag{15}$$

Once the Se is obtained, we can compute its variance, its standard error, its 95% confidence interval, its p-value, and $I^2$ estimating effects size's between studies heterogeneity.

## 7.2 Scales used for replication

Table 8: Clinical scales used to update Cortese et al. [2016] with our choices and the two new articles.

| Study | Outcome | Score Names - Parents ratings | Score Names - Teachers ratings |
|---|---|---|---|
| Arnold et al. [2014] | Total | SNAP IV | SNAP IV |
| | Inattention | SNAP IV | SNAP IV |
| | Hyperactivity | SNAP IV | SNAP IV |
| Bakhshayesh et al. [2011] | Total | German ADHD-RS | German ADHD-RS |
| | Inattention | German ADHD-RS | German ADHD-RS |
| | Hyperactivity | German ADHD-RS | German ADHD-RS |
| Baumeister et al. [2016] | Total | DISYPS | - |
| Beauregard and Levesque [2006] | Total | CPRS | - |
| | Inattention | CPRS | - |
| | Hyperactivity | CPRS | - |
| Bink et al. [2014] | Total | ADHD-RS self report | - |
| | Inattention | ADHD-RS self report | - |
| | Hyperactivity | ADHD-RS self report | - |
| Christiansen et al. [2014] | Total | Conners-3 Parents | Conners-3 Teachers |
| Gevensleben et al. [2009] | Total | German ADHD-RS | German ADHD-RS |
| | Inattention | German ADHD-RS | German ADHD-RS |
| | Hyperactivity | German ADHD-RS | German ADHD-RS |
| Heinrich et al. [2004] | Total | German ADHD-RS | - |
| Holtmann et al. [2009] | Total | German ADHD-RS | - |
| | Inattention | German ADHD-RS | - |
| | Hyperactivity | German ADHD-RS | - |
| Linden et al. [1996] | Total | IOWA Conners | - |
| | Inattention | IOWA Conners | - |
| Maurizio et al. [2014] | Total | CPRS | CTRS |
| | Inattention | CPRS | CTRS |
| | Hyperactivity | CPRS | CTRS |
| Steiner et al. [2011] | Total | Conners Rating Scales Revised | Conners Rating Scales Revised |
| | Inattention | Conners Rating Scales Revised | Conners Rating Scales Revised |
| | Hyperactivity | Conners Rating Scales Revised | Conners Rating Scales Revised |
| Steiner et al. [2014] | Total | Conners-3 Parents | Conners-3 Teachers |
| | Inattention | Conners-3 Parents | Conners-3 Teachers |
| | Hyperactivity | Conners-3 Parents | Conners-3 Teachers |
| Strehl et al. [2017] | Total | German ADHD-RS | German ADHD-RS |
| | Inattention | German ADHD-RS | German ADHD-RS |
| | Hyperactivity | German ADHD-RS | German ADHD-RS |
| van Dongen-Boomsma et al. [2013] | Total | ADHD RS | ADHD RS |
| | Inattention | ADHD RS | ADHD RS |
| | Hyperactivity | ADHD RS | ADHD RS |

SNAP: Wanson, Nolan and Pelham Questionnaire, ADHD-RS: ADHD Rating Scale, CPRS: Conners Parent Rating Scale, CTRS: Conners Teacher Rating Scale, BOSS Classroom Observation: Behavioral Observation of Students in Schools, DISYPS: Diagnostic System of Mental Disorders in Children and Adolescents