Maynooth University
National University of Ireland Maynooth

P300-Based Neurofeedback and Adaptive Task Difficulty Using Iterative Learning Control: A Novel Approach to Cognitive Training in Healthy Adults

Sandra-Carina Noble

A thesis submitted in fulfillment of the
requirements for the degree of
Doctor of Philosophy

Maynooth University
Faculty of Science and Engineering
Department of Electronic Engineering

Primary Supervisor	*Secondary Supervisor*	*Head of Department*
Prof John V. Ringwood	Prof Tomás Ward	Prof Gerard Lacey

August 2024

COLOPHON

This thesis was typeset with LaTeX using the `classicthesis` class developed by André Miede, with minor modifications.

ABSTRACT

The rising prevalence of neurodegenerative diseases such as dementia and Parkinson's disease poses a critical challenge as the global population continues to age. Enhancing cognitive reserve through cognitive training, particularly via neurofeedback (NFB), has become a promising strategy to counteract cognitive decline. This thesis presents a comprehensive study on the development and evaluation of a novel NFB training system designed to enhance attention in healthy adults. The system leverages event-related potentials (ERPs) and iterative learning control (ILC) to dynamically personalise task difficulty, thereby optimising training efficiency and engagement.

The research is underpinned by extensive data collection, involving a large-scale clinical trial with a significant sample size of healthy adult participants. The trial rigorously tested the system efficacy, providing robust evidence of its effectiveness. Participants were divided into groups, with one group receiving ILC-adapted training and others following traditional or random difficulty protocols. The results demonstrate that the ILC group not only completed the training more rapidly but also achieved substantial improvements in attention, validated by both behavioural metrics and neurophysiological markers.

Further investigations within this thesis address the system practicality, including studies on reducing the number of EEG electrodes to improve usability. The potential transferability of attentional improvements to motor skill acquisition in surgical training is also explored, revealing insights that guide future research in this domain.

In conclusion, this thesis contributes significantly to the field of cognitive training by showcasing the potential of ERP-based NFB systems in enhancing attention through large-scale, real-world clinical trials. The findings open new avenues for applying such systems in broader cognitive training and rehabilitation contexts, with recommendations for future studies to explore long-term impacts and cross-domain applicability.

DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education.

Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Maynooth, August 2024

Sandra-Carina Noble

ACKNOWLEDGEMENTS

PERSONAL

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor John Ringwood and Professor Tomás Ward. Their unwavering support, invaluable guidance, and insightful feedback have been instrumental in shaping this project. Their encouragement and expertise have greatly influenced my research journey, and I am incredibly grateful for their mentorship.

I would also like to extend my sincere thanks to Eva Woods, whose assistance, thought-provoking discussions, and encouragement have been invaluable. Her hard work and dedication significantly aided the progress of this research.

I am grateful to my collaborators, Madalena Rente and Seamus Morris, for their critical role in facilitating the surgical training study. Their collaboration and expertise have significantly contributed to the progress and success of this project.

My heartfelt thanks go to Carrie Anne Barry, Ann Dempsey, John Maloco, and Denis Buckley for their continuous administrative and technical support. Their dedication and efficiency have ensured the smooth running of the project, and I am thankful for their assistance.

I would like to acknowledge all past and current members of the Centre for Ocean Energy Research (COER) for their support and camaraderie during my time as an honorary member. Even though my project was quite different from COER's primary focus on wave energy, I always felt welcome and appreciated. Thank you for "volunteering" as my guinea pigs, for your insightful feedback, and for making the work environment enjoyable and stimulating. Special thanks to Sonal, Ahsan, and Peter for their help and support with the EEG system, and for their company at the BCI meeting.

I am deeply indebted to my family. To my parents, Anke and Christoph, thank you for instilling a sense of curiosity in me from a young age and for all your support throughout my academic journey. Your encouragement and belief in my abilities have been a constant source of strength. To my sisters, Anna and Johanna, thank you for always being there to listen to my rants and ramblings, and for providing me with love and encouragement. I am eternally grateful to my husband, Joshua, for his unwavering support, patience, and love. Thank you for always being there to lift me up during challenging times, for keeping me fed, and for making sure that I see the sunlight every now and then. I also want to thank my cat, Stinker, for being by

my side (or on my lap) during the thesis write-up, even if that slowed me down.

The following is a translation into my mother tongue, German:

Ich bin meiner Familie unendlich dankbar. Meinen Eltern, Anke und Christoph, möchte ich von Herzen danken, dass sie von klein auf meine Neugier geweckt und mich in all den Jahren auf meinem akademischen Weg unterstützt haben. Eure Ermutigung und euer Vertrauen in mich waren immer eine große Quelle der Kraft. Meinen Schwestern, Anna und Johanna, danke ich dafür, dass sie mir immer zuhören und mir mit Liebe und Unterstützung zur Seite stehen. Meinem Mann Joshua bin ich unendlich dankbar für seine bedingungslose Unterstützung, Geduld und Liebe. Danke, dass du immer für mich da bist, mich in schwierigen Zeiten aufmunterst, und dafür sorgst, dass ich immer etwas zu essen habe und ab und zu das Tageslicht sehe. Ich möchte auch meinem Kater Stinker danken, der immer an meiner Seite (oder auf meinem Schoß) war, während ich meine Arbeit geschrieben habe, auch wenn mich das manchmal langsamer gemacht hat.

INSTITUTIONAL

I would like to express my sincere gratitude to the Irish Research Council for funding this work through the Government of Ireland Postgraduate Scholarship.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ADHD attention deficit hyperactivity disorder
ANOVA analysis of variance
ART aligned rank transformation
BCI brain-computer interface
DAN dorsal attention network
DC direct current
ECoG electrocorticography
EEG electroencephalography
EMG electromyography
EP evoked potential
ERN error-related negativity
ERP event-related potential
FES functional electrical stimulation
fMRI functional magnetic resonance imaging
fNIRS functional near-infrared spectroscopy
ICC intraclass correlation coefficient
ILC iterative learning control
ISI inter-stimulus interval
LDA linear discriminant analysis
MCI mild cognitive impairment
MEG magnetoencephalography
MSE mean-squared error
NFB neurofeedback
NOILC norm-optimal iterative learning control
NSR noise-to-signal ratio
P_e error positivity
POILC parameter-optimal iterative learning control
RDM random dot motion
RT response time
SCP slow cortical potential
SNR signal-to-noise ratio
TBR theta-beta ratio

TLX task load index

TOST two one-sided tests

TTI target-to-target interval

VAN ventral attention network

VEP visual evoked potential

Part I

INTRODUCTION AND BACKGROUND

1 INTRODUCTION

1.1 MOTIVATION AND OBJECTIVES

The global population is aging rapidly. In 2022, nearly 19% of people living in Europe and North America were 65 years old or older, a figure projected to rise to almost 27% by 2050 [1]. This demographic shift is not confined to specific regions but is a global trend, with over 16% of the world population expected to be over 65 years old by 2050, up from less than 10% in 2022 [1]. Aging is often accompanied by an increased prevalence of diseases that impair cognitive abilities, such as dementia and Parkinson's disease. The incidence of these conditions, and the associated burden on individuals, families, and society, has already increased and is expected to continue growing in tandem with the aging population [2, 3]. While significant research efforts are focused on disease-modifying treatments that aim to slow, halt, or reverse the progression of such diseases, most current therapies primarily focus on symptom management [4, 5].

Given the challenges of treating these progressive diseases, and the anticipated rise in their occurrence, there is a growing emphasis on prevention strategies, aimed at promoting healthy aging. These strategies include lifestyle factors like physical activity, smoking abstinence, alcohol consumption reduction, and social engagement [6]. A key concept in healthy aging and the prevention of cognitive decline is cognitive reserve, which refers to the ability of the brain to adapt and maintain function, despite damage or disease [7]. Cognitive reserve is influenced by both genetic and environmental factors, including education and lifestyle choices [7]. Regular participation in cognitively stimulating activities has shown promise in enhancing cognitive reserve and reducing the risk of developing dementia later in life [6].

One method for intentionally boosting cognitive reserve is cognitive training. This involves a series of tasks designed to enhance specific cognitive functions such as memory, attention, and problem-solving skills. Research has demonstrated the effectiveness of cognitive training in both preventing and treating diseases like dementia [8] and Parkinson's [9]. Beyond healthy aging, cognitive training can be beneficial for individuals in professions or activities requiring high levels of cognitive functioning, such as surgery [10], military [11], and athletics [12].

Neurofeedback (NFB) training, a form of cognitive training using brain-computer interfaces (BCIs), where users receive real-time feedback based on their brain activity, allowing them to self-regulate this

activity, has shown promising results [13, 14]. However, there are concerns about the efficacy of NFB training due to a lack of large-scale and well-controlled clinical trials [15]. Furthermore, the use of NFB training is mostly limited to lab environments [16]. Factors such as long setup and training times, along with high associated costs, present significant barriers to its practical application in real-world settings [17].

Building on the potential of NFB-based cognitive training, and addressing current challenges in this field, this research aims to develop and evaluate an accessible, and effective, NFB training system. The objectives of this project are to:

- Demonstrate that NFB training can effectively improve cognitive function in healthy adults through rigorously designed clinical trials, providing robust evidence of its efficacy.
- Enhance the efficiency of training, enabling faster cognitive improvements, through the use of iterative learning control (ILC) to dynamically adapt task difficulty based on user performance.
- Ensure practicality and usability by minimising the number of electrodes required, thereby reducing setup time and making the system more scalable and user-friendly.
- Demonstrate the applicability of the system in real-world environments.

Ultimately, the goal is to create an NFB training system that can be easily deployed as an intervention in real-world settings.

1.2 MAIN CONTRIBUTIONS

This thesis makes the following key contributions, some of which are tied to specific research questions:

- **Development of a novel NFB training system:** A system designed to enhance attention in healthy adults. This system uses event-related potentials (ERPs), instead of the typical frequency bands for NFB, and employs an ILC controller to optimise task difficulty. The controller automatically adjusts the task difficulty based on user performance, accommodating changes due to factors such as learning effects or fatigue.
- **Rigorous evaluation of ILC-controlled training:** The effectiveness of ILC-controlled training is rigorously evaluated through simulations and a large-scale clinical trial involving a substantial cohort of healthy adults. Although the trial did not involve patients, it was registered as a clinical trial because it adhered to the strict protocols of intervention-based research. The study

systematically tested neurofeedback training interventions to assess their effects on cognitive function, aligning with formal clinical trial criteria. This registration ensured transparency, ethical oversight, and adherence to trial design standards, thereby reinforcing the reliability and generalisability of the findings.

  - **Research Question 1:** Can ILC enhance the efficiency of NFB training to improve attention in healthy adults?

- **Reduction of electrodes:** Several studies are conducted to explore and enhance the system practicability by reducing the number of electrodes required for training. This reduction not only minimises setup time and cost but also enhances user comfort, making the system more accessible and feasible for use in real-world contexts beyond laboratory settings.

  - **Research Question 2:** How does the number of electrodes affect the usability, accuracy, and effectiveness of the NFB system?

- **Exploration of training transfer effects:** The impact of the training on motor skill acquisition and retention in surgical trainees is investigated, successfully demonstrating the deployment of the NFB system in a large scale, real-world setting.

  - **Research Question 3:** Can attentional improvements gained from NFB training transfer to motor skill learning?

- **Contributions to open science:** The studies described in this thesis generated substantial datasets comprising 151 NFB sessions with 132 participants. These sessions include EEG signals recorded from a wide variety of settings, with electrode configurations ranging from 4 to 32 electrodes, and environments spanning from shielded rooms to typical office settings. Where participants consented to further use of their data, these datasets will be made available in online repositories to facilitate future research. Additionally, the code used for the experiments is freely available to support future research and foster transparency.

1.2.1 *List of Publications*

This section lists the publications that resulted from the work described in this thesis.

1.2.1.1 *Journal Publications*

- In Preparation:
  - **Noble, S.C.**, Rente, M.N., Ward, T., Morris, S., Ringwood, J.V., "Evaluating the effect of P300-based neurofeedback on

surgical training", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*

- Rente, M.N., **Noble, S.C.**, Ward, T., Ringwood, J.V., Morris, S., "Investigating the use of cognitive simulation and neurofeedback training for improving retention of motor skills in surgical training", *Surgical Innovation*
- **Noble, S.C.**, Ward, T., Ringwood, J.V., "Investigating the efficacy of P300-based neurofeedback training with minimal electrode sets", *IEEE Transactions on Biomedical Engineering*
- **Noble, S.C.**, Ward, T., Ringwood, J.V., "Multi-Condition EEG Dataset from Neurofeedback Training with a P300 Speller", *Data in Brief*

- Published:
  - **Noble, S.C.**, Woods, E., Ward, T., Ringwood, J.V., "Accelerating P300-based neurofeedback training for attention enhancement using iterative learning control: A randomized controlled trial," *Journal of Neural Engineering*, vol. 21, no. 2, 026006, 2024
  - **Noble, S.C.**, Woods, E., Ward, T., Ringwood, J.V., "Adaptive P300-based brain-computer interface for attention training: Protocol for a randomized controlled trial," *JMIR Research Protocols*, vol. 12, e46135, 2023

1.2.1.2 *Conference Publications*

- In Press:
  - **Noble, S.C.**, Ward, T., Ringwood, J.V., "Assessing the impact of environment and electrode configuration on P300 speller performance and EEG signal quality," *Proc. 2024 IEEE International Conference on Engineering in Medicine and Biology (EMBC), Orlando, FL, USA*, 2024

- Published:
  - **Noble, S.C.**, Ward, T., Ringwood, J.V., "Comparing the effect of electrode selection on P300 speller performance," *Proc. 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, HI, USA*, 2023
  - **Noble, S.C.**, Ward, T., Ringwood, J.V., "A phenomenological model of cognitive performance as a measure of attention in a P300-speller task," *Proc. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic*, 2022

1.3 THESIS OUTLINE

This thesis is organised into four parts. The first part includes this introduction, along with a literature review and technical background, providing the necessary context for the rest of the thesis. In Chapter 2, attention as a cognitive function is explored, along with how it can be trained using NFB. Different types of NFB and methods for task difficulty adaptation are also discussed. Chapter 3 introduces ILC, which is used in this thesis to adapt task difficulty in NFB training.

The second part of the thesis details the core developmental work, including system development, modelling, and control design, as described in Chapter 4.

The third part presents the experimental work, encompassing data analysis and experimental studies, which are discussed in Chapters 5, 6, and 7.

The fourth and final part of the thesis provides the conclusions in Chapter 8.

2

NEUROFEEDBACK TRAINING FOR ATTENTION ENHANCEMENT

2.1 INTRODUCTION

NFB training is a method of closed-loop brain training where individuals learn to modulate their brain activity through real-time feedback based on that activity. It relies on neural plasticity, the ability of the brain to change its structure, as well as operant conditioning, a learning process driven by (positive or negative) reinforcement [18].

Specifically, in NFB training, a person's brain activity of interest is measured with one or more of the neuroimaging methods explained in the following paragraph. This measured brain activity is used to provide feedback to the person, who attempts to control it by self-regulating their brain activity, thus creating a feedback loop. Over time, this self-regulation can lead to a change in brain patterns and/or behaviour [18]. An illustration of this feedback loop, showing different neuroimaging methods and feedback modalities, is provided in Figure 2.1.

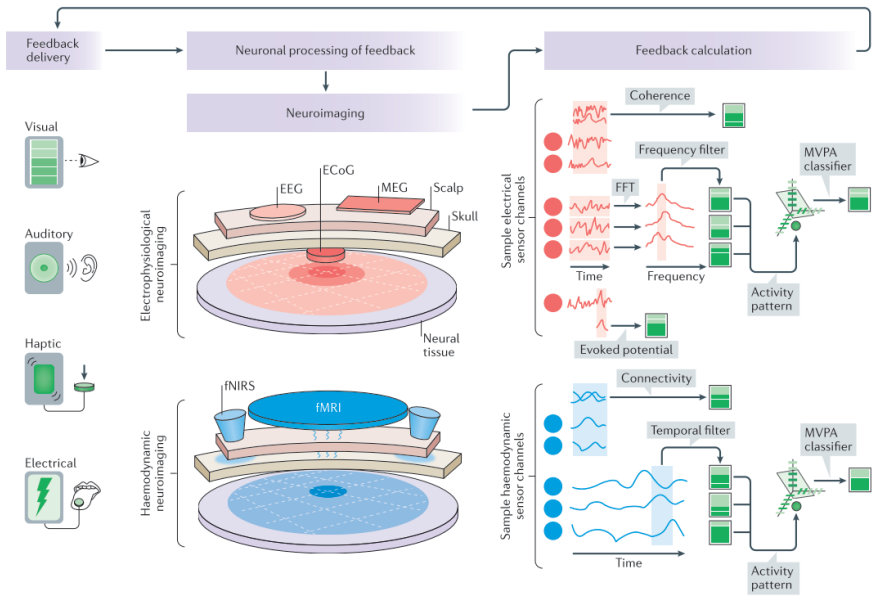


Figure 2.1: Neurofeedback (NFB) loop with possible feedback modalities, neuroimaging methods and brain activity measurements. Taken from [18].

There are generally five different neuroimaging methods used for NFB, which are used on their own, or combined for multi-modal NFB systems [18]:

- **Electroencephalography (EEG) and Magnetoencephalography (MEG):** Non-invasive methods that measure the cortical electrical and magnetic activity of the brain, respectively, through the skull. These methods have high temporal resolution but relatively low spatial resolution.
- **Electrocorticography (ECoG):** An invasive method that improves spatial resolution by measuring electrical activity directly from the brain's cortex.
- **Functional magnetic resonance imaging (fMRI) and Functional near-infrared spectroscopy (fNIRS):** Methods that detect oxygenated and deoxygenated blood within the brain, offering greater spatial resolution but low temporal resolution.

EEG-NFB is the most commonly used modality and was the first modality to be used for NFB. The emergence of EEG-NFB began in the 1960s and 1970s. Kamiya [19] was the first to demonstrate the learned control of brain waves through reward. Concurrently, Sterman [20] applied NFB training to treat seizure disorders in cats, and later humans. Despite facing criticism and a decline in research during the 1970s, interest in EEG-NFB has resurged due to advancements in BCI technologies [15]. Today, EEG-NFB is utilised for treating various mental and mood disorders, such as schizophrenia, depression, dementia, ADHD, and in cognitive rehabilitation after brain injuries or stroke [13].

However, EEG-NFB is not limited to individuals with cognitive deficits; it also offers potential benefits for healthy individuals, such as enhancing cognitive skills [14] or motor performance [21].

This thesis focuses on using EEG-NFB for attention enhancement. The fundamentals of EEG are introduced in Section 2.2. Attention is further described in Section 2.3, with explanations of its importance, neural substrates, and correlates. The different types of EEG-NFB are described in Section 2.4, focusing on how they modulate the neural substrates and correlates of attention. An overview of common personalisation methods in EEG-NFB is provided in Section 2.5, before concluding the chapter in Section 2.7.

2.2 ELECTROENCEPHALOGRAPHY

EEG is a method to measure the electrical activity of the brain. This is done by attaching electrodes to certain locations on the scalp [22]. The International 10-20 system was previously the standard for electrode

placement. However, it has been modified to accommodate more electrodes and is now referred to as the 10-10 system [23]. This system can be seen in Figure 2.2. Electrodes are usually not in direct contact with the scalp, so an electroconductive gel is applied to bridge the gap. The measured electrical activity is then amplified by the EEG amplifier. The output of this amplifier is a time-series of voltage for each electrode [22].

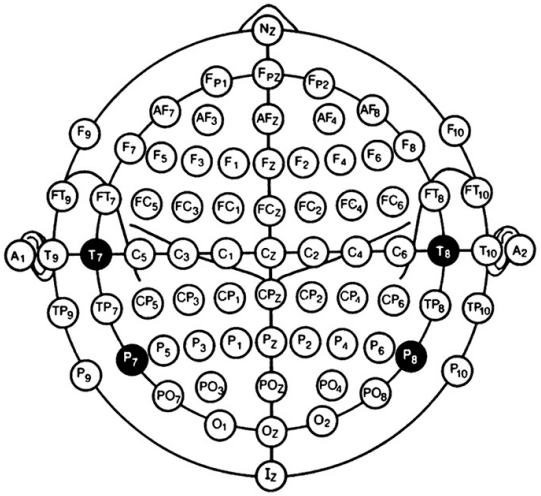


Figure 2.2: Electrode placement according to the 10-10 system. Taken from [23].

Various features are extracted from these time-series. The most important for EEG-NFB are spectral features (the amplitude of EEG signals at different frequencies), slow cortical potentials (SCPs), which are direct current (DC) shifts in EEG signals, and patterns occurring in response to stimuli or events, known as ERPs. These features will be discussed in more detail in Section 2.2.1, Section 2.2.2, and Section 2.2.3, respectively.

2.2.1 *Spectral Features*

EEG signals are often analysed in the frequency domain, where different frequency bands and their associated behaviours or psychological states are defined. These bands are called EEG rhythms.

The most well-known rhythm is the alpha rhythm, observed in relaxed but awake states, particularly with eyes closed. The low-frequency delta rhythm is associated with deep sleep, but also cognitive processing, and is often observed in ERP studies. The theta rhythm is typically associated with drowsiness, and high activity in this rhythm is considered abnormal when resting. However, theta activity is also associated with attention and working memory. The beta rhythm is most commonly observed in frontal regions and is associated with

cognitive processing and the sensorimotor system. The high-frequency gamma rhythm in temporal regions is associated with memory [24].

An overview of EEG rhythms with their frequency bands, location, and associated state is given in Table 2.1.

Table 2.1: EEG rhythms and their associated behaviour or psychological state.

Rhythm	Band (Hz)	Location	Behaviour / Psychological State
Delta	<4	variable	deep sleep, cognitive processing
Theta	4-7	frontal and temporal	drowsiness, attentional processing
Alpha	8-13	occipital and parietal	mentally inactive but awake
Beta	13-25	frontal and central	sensorimotor functions, cognitive processing
Gamma	>25	temporal	memory processes

2.2.2 *Slow Cortical Potentials*

SCPs are relatively slow DC shifts in EEG signals that occur in response to stimuli. They can last from several hundred milliseconds to several seconds. Negative SCPs are associated with cognitive processing, while positive SCPs reflect an attenuation of cortical excitability, which is often observed in behavioural inhibition [25].

2.2.3 *Event-Related Potentials*

ERPs, formerly known as evoked potentials (EPs), are brain signals that occur in response to stimuli or events. They are commonly used as a tool to study the mind and brain in cognitive psychology and neuroscience, as well as in BCIs [26].

An ERP signal usually consists of several (overlapping) ERP components, which are in the form of a positive or negative voltage peak happening a certain time after the stimulus onset, although there are some anticipative ERP components that occur pre-stimulus. The naming convention for ERPs is P for a positive peak and N for a negative peak, followed by either the number of the peak in the waveform, i.e. P1 for the first peak and P2 for the second peak, or the latency in mil-

liseconds, i.e. the P300 is a positive peak that occurs approximately 300 ms after the stimulus onset [26].

Different components, which are distinguished by their polarity, latency, and source in the brain, are associated with different cognitive processes. It should be noted that the same label for a component can refer to different cognitive processes, depending on the source of the component and the sensory modality, i.e. visual or auditory. However, this mostly affects early components [26].

Since all brain activity that is not in response to the given stimulus or event is considered noise in the ERP signal, it is important to improve the signal-to-noise ratio (SNR) to obtain a clear ERP wave. This is typically achieved by averaging the ERP signal over several trials [22]. Machine learning algorithms are commonly used for ERP classification and can nowadays achieve over 90% accuracy even with single trials [27, 28], although classification performance is highly dependent on EEG signal quality.

An overview of common (visual) ERP components is presented in the following paragraphs.

An early visual ERP component is the P100. It usually occurs 100 ms to 130 ms after a visual stimulus and is modulated by selective attention and arousal [29].

The P100 is often followed by the N100 component, which reflects spatial attention and discriminative processing in the brain. It consists of several subcomponents that typically happen between 100 ms and 200 ms post-stimulus [29]. One of the subcomponents is the N170, labelled as such due to its latency of usually around 170 ms. The N170 is evoked by face perception [30].

Later visual components are often observed due to the so-called oddball paradigm. In the oddball paradigm, a frequent stream of common stimuli is interspersed with infrequent, uncommon stimuli, or target stimuli, that the participant was told to attend to. While the P200 with a typical latency of approximately 200 ms is not as well understood as some other components, it is believed to occur in response to the oddball paradigm with simple targets [31].

Several components are grouped together and commonly referred to as the N200 component. The anterior N200, also known as N2b, reflects response inhibition in the go/no-go paradigm, where a button press is required for one type of stimulus (go) but not for another (no-go) [32]. It also reflects the detection of a mismatch of attended stimuli. The posterior N200, which is also called N2c, usually occurs in response to the oddball paradigm [33].

Similarly to the N2, the P3 can also refer to several components depending on the source of the component. While the so-called P3a, which originates in frontal brain regions, and the P3b, which is mostly observable in parietal brain regions, both respond to stimulus changes, i.e. the oddball paradigm, the P3b is only elicited if these changes are

task-relevant. The general labels P3 and P300 usually refer to the P3b [33].

A subset of ERPs are steady-state visual evoked potentials (VEPs). They occur when stimuli are presented at a fast rate and the brain activity starts to synchronise with the stimulus frequency. Thus, the steady-state response consists of two sine waves, one at the stimulus frequency and another, less dominant, one at twice the stimulus frequency [34]. Due to their robustness against noise, steady-state VEPs are particularly valuable for BCI applications.

Another type of VEP commonly used in BCI systems is the code-modulated VEP. Code-modulated VEPs use uniquely coded visual sequences that allow for rapid and accurate detection of user intentions [35]. For diagnostic purposes, motion-onset VEPs are often employed; these occur in response to the onset of motion within a visual stimulus and typically include components such as the P100, N200, and P200 [36].

The error-related ERP is a negative peak following an incorrect response, also known as the error-related negativity (ERN). Notably, this component is response-locked, occurring a specific time after an incorrect response, rather than being stimulus-locked. The ERN is usually followed by a positive wave, which peaks approximately 400 ms after the incorrect response. This is called the error positivity (P_e). It is only present when the participant is aware of the error, whereas the ERN occurs regardless of the participant's awareness [37].

2.3 ATTENTION

2.3.1 *Importance of Attention in Cognitive Functioning*

This thesis focuses on the cognitive ability of attention. Attention is a fundamental cognitive ability, crucial for effective functioning in daily life. It underlies the ability to focus on specific tasks, filter out irrelevant information, and manage multiple tasks simultaneously. This ability is fundamental for learning, memory, and overall cognitive performance. Attention is particularly important in various daily activities, such as driving, where the ability to selectively focus on the road and ignore distractions is vital for safety. In academic and professional settings, attention facilitates absorbing and applying new information, problem-solving, and critical thinking.

Attention is defined as the selectivity of processing and can be classified into selective and divided attention. Selective attention is the ability to focus on one stimulus while ignoring other stimuli, whereas divided attention is the ability to focus on more than one stimulus at once, in other words, multi-tasking [38].

Attention can be bottom-up, driven by external stimuli, such as a loud noise that demands immediate attention, or top-down, where

attention is controlled voluntarily based on the goals of the individual [39].

However, attention is not a static ability and can be affected by both normal aging and various diseases and conditions. Age-related cognitive decline often includes a reduction in attentional capacity, making it harder for older adults to concentrate and filter out distractions. Diseases and conditions such as dementia, stroke, attention deficit hyperactivity disorder (ADHD), and autism spectrum disorders can significantly impair attentional processes [40]. For instance, individuals with ADHD may struggle with sustained selective attention [41], while those with autism might find it challenging to shift attention between tasks [42]. Stroke and dementia patients may experience deficits in both selective and divided attention, affecting their ability to process information and respond to environmental stimuli effectively [40].

Given its crucial role and the widespread impact of attentional deficits, enhancing attention through targeted interventions is essential. This thesis focuses on attention enhancement because improving this cognitive ability can significantly benefit individuals at all stages of life.

2.3.2 *Neural Substrates and Correlates of Attention*

Attention is a complex cognitive function supported by various brain regions and networks. Understanding its neural substrates and correlates aids in developing targeted interventions to enhance attentional capacities. This section explores networks of brain regions that underlie attention (neural substrates) and neurophysiological markers that are associated with attention (neural correlates).

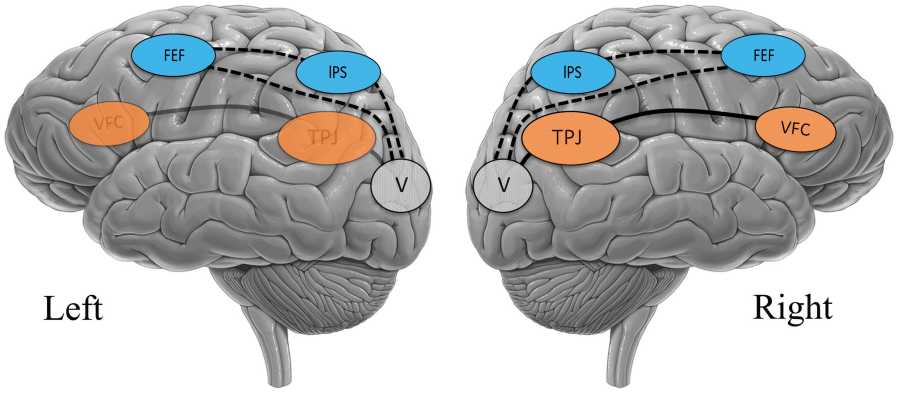


Figure 2.3: Brain regions involved in the dorsal attention network (DAN), shown in blue, and the ventral attention network (VAN), shown in orange. Adapted from [43].

2.3.2.1 *Dorsal Attention Network*

The dorsal attention network (DAN) is involved in all forms of selective attention. It mainly includes the frontal eye field (FEF), a brain region responsible for voluntary eye movement, and the intraparietal sulcus (IPS), a brain region responsible for perceptual-motor coordination, which are shown in blue in Figure 2.3 [44].

2.3.2.2 *Ventral Attention Network*

The ventral attention network (VAN) is another network associated with attention. While the DAN is always active in selective (visual) attention, the VAN is suppressed during top-down attention and active in bottom-up attention. It includes the temporoparietal junction (TPJ), involved in information processing, and the ventral frontal cortex (VFC), associated with decision-making. These regions are illustrated in orange in Figure 2.3 [44].

2.3.2.3 *EEG Rhythms*

The main EEG rhythm associated with attention is the alpha rhythm. Decreased alpha activity is hypothesised to reflect attention directed towards external stimuli [45]. This suppression of alpha activity has been found to correlate with the amplitude and latency of the P300 ERP [46]. Conversely, increased alpha activity has been observed in internally directed attention, such as during mental imagery or arithmetic [45].

High theta rhythm activity is associated with inattentiveness, while high beta activity reflects focused attention [47].

High-frequency EEG rhythms (beta and gamma) are believed to be related to stimulus selection, which is necessary for selective attention [48].

2.3.2.4 *Event-Related Potentials*

While there are no ERPs that are directly elicited by attention, several are modulated by it. These include the P100, N100, N200, and P300, previously discussed in Section 2.2.3. Specifically, the amplitude of these components is amplified with increased attention [49].

2.4 TYPES OF EEG-NEUROFEEDBACK

This section provides an overview of the different types of EEG-NFB reported in the literature. While EEG-NFB is applied across a range of areas, including treatments for mental health conditions such as schizophrenia, depression, and anxiety, as well as for pain management and epilepsy [50], this section focuses specifically on applications designed to enhance attention. The majority of literature in this

area focuses on using NFB to treat ADHD, e.g. [51–54]. However, it has also been applied to other clinical populations, such as those with mild cognitive impairment (MCI) [55, 56] or stroke [57], as well as healthy populations, e.g. [58–60].

The most popular type of EEG-NFB is rhythm-based. This means that the target for self-modulation is at least one EEG rhythm, discussed in Section 2.2.1. This type is explained further in Section 2.4.1.

Another type of EEG-NFB used to enhance attention, though less frequently than rhythm-based NFB, is SCP-based NFB. SCPs have been explained previously in Section 2.2.2. This type of NFB is discussed in Section 2.4.2.

Lastly, an emerging type of EEG-NFB is ERP-based, where the target of self-modulation is a specific ERP component instead of a rhythm (see Section 2.2.3 for details on ERPs). This type of EEG-NFB is discussed in Section 2.4.3.

2.4.1 *Rhythm-Based*

The most commonly used NFB target for attention enhancement is the downregulation of the theta-beta ratio (TBR), aiming to increase beta rhythm activity while simultaneously decreasing theta activity. As discussed in Section 2.3.2.3, low theta and high beta activity have been associated with attention, making the TBR an obvious target. The TBR protocol is sometimes accompanied by the downregulation of higher frequencies to improve the specificity of the NFB protocol [51, 61, 62].

Other targets for rhythm-based NFB include the up- or downregulation of activity in the alpha rhythm, treated in [63–65] and [66–68], respectively. Upregulation of the beta rhythm is also a popular protocol for attention training [56, 69, 70], sometimes accompanied by the simultaneous downregulation of alpha activity [55, 71]. The theta rhythm on its own has also been targeted, with specific protocols for both upregulation [72, 73] and downregulation [74].

Numerous studies have applied the TBR protocol to treat ADHD in both children and adults, e.g. [51–53]. These studies have reported positive effects on symptoms that were maintained up to 25 months later [51, 52], as well as task performance in cognitive/attentional tests [53, 54]. Improved performance in cognitive tasks has also been observed following a single session of alpha downregulation NFB in adults, with and without ADHD [68]; and in children with ADHD following a course of alpha upregulation [75], as well as with beta upregulation [69, 72], and theta upregulation [72].

The TBR protocol has also been applied to child epilepsy [61], as well as to stroke patients [57], both with improved task performance in cognitive tests. Cognitive functioning, including attention, has been successfully improved in the elderly, with and without MCI, using

beta upregulation [56, 70] and theta upregulation [73]. All discussed rhythm-based protocols have also been used to enhance cognitive abilities in healthy children and adults, e.g. [58, 59, 64, 67, 73, 76].

Studies using these protocols often report EEG changes in one or more of the targeted rhythms, e.g. [51, 57, 74], with some observing changes in non-targeted rhythms [63, 77, 78]. These changes are often only observed in experimental groups, and not control groups that receive sham-NFB, where feedback is not based on the participant's own brain activity as they believe, but rather based on recorded or simulated signals (e.g. [51]). However, participants in these sham-control groups often experience the same behavioural outcomes as the experimental groups (e.g. [52]), which has led to the efficacy of NFB being questioned and criticised as potentially being a placebo effect [79].

To address this, researchers commonly employ rigorous study designs to distinguish genuine NFB effects from placebo responses. These include using active control tasks, blinding participants to group allocation, and incorporating objective outcome measures (both brain activity and behavioural measures). Such strategies enhance the reliability of NFB findings by helping to rule out placebo-driven improvements and strengthen the case for NFB efficacy [80].

One older study also observed an increased P300 amplitude following TBR-NFB [60], while another reported increased P300 and N100 amplitudes following alpha downregulation [68]. An fMRI study observed changes in gray and white matter, in brain regions associated with the DAN, following NFB training with a beta upregulation protocol [76]. Similarly, Ros et al. [67] observed changes in the connectivity of brain regions associated with the VAN following just a single half-hour alpha downregulation NFB session.

Rhythm-based NFB protocols have several notable strengths that contribute to their widespread use and established presence in the literature. One major strength is their extensive validation through numerous studies, which have demonstrated their efficacy in enhancing cognitive performance and attention across diverse populations. This extensive research base provides a solid foundation for the continued application and refinement of these protocols.

Another key strength is the inherent flexibility of rhythm-based NFB. Since EEG rhythms are always present in brain activity, training can be adapted to various tasks and feedback mechanisms. This flexibility allows practitioners to tailor interventions to specific cognitive goals or individual needs, enhancing the practical applicability of NFB in both clinical and non-clinical settings.

However, rhythm-based NFB also has its weakness. One significant limitation is the fixed nature of the rhythms typically used in these protocols. These standard rhythms may not be optimal for every individual, as individual differences in EEG spectra can influence the

effectiveness of the training. Research has shown that individualising the target rhythms for NFB can lead to better outcomes, suggesting that more personalised approaches may be necessary to maximise efficacy [69, 81].

2.4.2 *Slow Cortical Potential-Based*

SCP-based NFB targets the up- and downregulation of SCPs, typically with a 50/50 split. SCP has not been used as extensively as rhythm-based NFB, with some applications in epilepsy and migraines [82]. For attention, it has been used to treat people with ADHD, with some studies reporting improved symptoms and performance in cognitive tasks [83, 84], while others did not find improvements to be superior to control groups [85–87].

Some studies report a change in SCPs following NFB training [84, 88], while some observed a decrease in P300 amplitude which is believed to reflect the adaptation to the given task, leading to reduced attentional processing being required [85, 88]. One study observed increased activation in the prefrontal cortex, following SCP-based NFB training, but also saw this change in the control group, who received electromyography (EMG)-based NFB training [83].

2.4.3 *Event-Related Potential-Based*

Given the relevance of ERPs like the P300 and N200 to attention, as discussed in Section 2.3.2.4, their potential as biomarkers for conditions such as ADHD [89], and evidence suggesting that ERPs are modifiable via NFB training as mentioned in Section 2.4.1, it is logical to consider targeting ERPs for self-modulation directly. However, while rhythm-based EEG-NFB is well established, there has been less research on the use of ERP-based NFB training. Rieger et al. [90] investigated the use of the N100, an auditory ERP component associated with attention, for NFB treatment of hallucinations in individuals with schizophrenia, but did not find the training to be effective. Musso et al. [91] successfully used NFB based on the auditory P300 for language training in patients with aphasia. Mismatch negativity (MMN), a subcomponent of the auditory N200, was used as the target in NFB for working memory training in patients with subjective cognitive decline [92].

Fouillen [93] used P300-based video games for attention training in children with ADHD. While she found improved symptoms that were maintained 2 months post-training, this was also the case in the control group who played the same game with gaze-based feedback. However, only the NFB group maintained improved task performance in cognitive tests at the 2-month follow-up. Li et al. [94] used a P300-based video game for cognitive training in healthy adults and

reported improved P300 amplitude and latency post-training, as well as decreased alpha activity. Both Jacoby [95] and Arvaneh et al. [96] used a P300 speller, traditionally a BCI used for communication, for cognitive training in healthy adults. These studies reported improved performance in cognitive tasks [96], as well as increased P300 amplitudes [95, 96] and decreased alpha activity [96]. Arvaneh et al. [96] did not observe these training effects in the control group, which completed the P300 speller training without receiving feedback.

Another study later reported that a single session of P300-based NFB training with a variation of the P300 speller leads to changes in gray matter in brain regions that are associated with the DAN and sustained visual attention [97], further supporting the use of ERP-based NFB training for attention enhancement.

ERP-based NFB has several notable strengths. One major strength is the well-researched nature of ERPs and their close association with specific cognitive processes, such as attention and working memory. This allows for highly targeted interventions that can specifically address deficits in these cognitive areas. The personalisation aspect is another significant advantage, as ERP-based NFB typically involves creating a personalised classifier for each individual, which could increase the specificity and effectiveness of the training.

However, there are also weaknesses associated with ERP-based NFB. A significant limitation is the fixed nature of the scenarios required to elicit ERPs, such as the oddball paradigm for P300, which makes the training less flexible compared to rhythm-based NFB. Additionally, while ERP-based NFB shows promise, there is relatively less research on its use compared to rhythm-based NFB, resulting in fewer established protocols and less evidence to support its widespread application. Furthermore, the high specificity of ERP-based NFB, while a strength, can also be a weakness as the training may become too narrowly focused, potentially overlooking broader cognitive processes that are important for overall cognitive function. Lastly, due to the repetitions required to accurately measure ERPs, the feedback frequency is slower compared to rhythm-based or SCP-based NFB. The importance of feedback frequency has been demonstrated in [98], showing that more frequent feedback leads to greater training effects.

2.5 PERSONALISATION OF EEG-NEUROFEEDBACK

Personalisation of NFB can be achieved in several ways. As discussed in Section 2.4.1, targeted rhythms can be personalised based on an individual's baseline brain activity and needs. This section, however, focuses on the personalisation of task difficulty.

Adjusting the task difficulty ensures that participants continuously work harder to receive positive feedback. This can be achieved by

directly adjusting the threshold needed for positive feedback or by modifying task conditions.

Personalising feedback thresholds or task difficulty is crucial because participants' confidence in their ability to control the signal and their motivation significantly influence NFB outcomes [18]. Both overly easy (consistently good feedback) and overly difficult (consistently bad feedback) training can negatively impact motivation and the overall effectiveness of the training.

In rhythm-based and SCP-based NFB, threshold adjustments are typically made according to specific rules, such as adjusting the threshold to maintain an 80% reward rate [61, 85], or by a certain percentage after two successful trials [66, 99]. These threshold adjustments are often done manually and sometimes not reported in studies, even though they might significantly impact training efficacy.

In ERP-based training, task difficulty adjustments, similar to these rule-based threshold adjustments, have been reported [93].

Jacoby [95] reported a performance plateau after three sessions of P300 speller training, hypothesising that the task might not be engaging enough to maintain participant motivation. Arvaneh et al. [96] addressed this issue by progressively reducing the number of repetitions based on participants' performance, to maintain engagement. Notably, reducing the number of repetitions in ERP-based training increases feedback frequency, aligning the training more closely with the real-time nature of rhythm-based NFB, which is beneficial for learning and maintaining participant engagement [98].

Outside the context of NFB training, several methods adapt the P300 speller to each individual, often referred to as early stopping. These methods continuously assess the probability of each stimulus being the target for each repetition and terminate once this probability surpasses a specific threshold [100, 101] or stabilises, i.e. when additional repetitions no longer provide new information [102]. Early stopping adapts the speller to the user's brain signals in real-time, which can reduce both task duration and mental fatigue by avoiding unnecessary repetitions. However, in contrast to the approach used by Arvaneh et al. [96], early stopping methods prioritise operational efficiency rather than fostering cognitive engagement, as Arvaneh et al.'s approach does by intentionally reducing repetitions to challenge participants. This deliberate increase in task difficulty, even at the cost of initial performance dips, is designed to drive attention enhancement and sustained learning over time.

2.6 CHALLENGES AND LIMITATIONS OF EEG-NEUROFEEDBACK

Despite the promising results demonstrated by various NFB training methods, as discussed in previous sections, the practical use of

NFB remains limited [16]. Several factors contribute to the lack of widespread, real-world applications of NFB.

Firstly, there are concerns about the efficacy and NFB-specific treatment effects due to a lack of large-scale, well-controlled studies. Specifically, the absence of double-blind, sham-controlled clinical trials is a significant issue [15]. Most NFB research comprises small exploratory studies rather than rigorously designed randomised controlled trials, limiting the strength of evidence for NFB interventions [21]. To address this, recent guidelines and recommendations within the NFB community have called for well-designed, rigorously analysed studies to adequately evaluate the efficacy of NFB interventions [80].

While the lack of double-blind, sham-controlled studies is a significant issue in NFB research, it is also one that poses considerable practical and ethical challenges. In the case of this thesis, double-blinding was not feasible as the project was conducted primarily by a single researcher, making it difficult to implement a double-blind design. Additionally, a sham-control group was not included, as obtaining ethical approval for a sham condition involving deception would have been particularly challenging. These factors often contribute to the limited implementation of such rigorous designs.

Secondly, there is a lack of research investigating the transition from laboratory research to real-world application of NFB interventions [16]. While some studies have shown that NFB can be applied in real-world contexts [103, 104], these studies are few, and comprehensive investigations into real-world deployment are needed.

Significant barriers to the real-world applicability of BCIs, including NFB, are the long training times required before achieving adequate control or good performance, which increases the cost of interventions and can negatively impact user motivation [17]. Motivation is a critical factor in the success of NFB training, as it directly influences engagement and outcomes [18]. Additionally, the time-consuming setup of wet electrode EEG systems, along with the typically user-unfriendly software, requiring expert operation, further limit the usability and applicability of NFB in non-laboratory settings [17].

2.7 SUMMARY

This chapter explores NFB, EEG, attention, and its neural substrates. Initially, it introduces the concepts and history of NFB. The section on EEG explains the technology and the significance of different brain wave patterns in relation to cognitive functions.

Attention is then discussed, emphasising its importance in cognitive functioning and detailing its neural substrates and correlates. Key brain networks involved in attention, such as the DAN and VAN, are examined along with neurophysiological markers of attention.

The chapter reviews different types of EEG-NFB, focusing on rhythm-based and ERP-based NFB methods. Rhythm-based protocols, targeting the TBR, alpha, beta and theta rhythms, are well-studied and have been used to enhance cognitive performance and attention across various populations, including individuals with ADHD, stroke patients, and healthy adults. Despite their efficacy, limitations such as the fixed nature of rhythms are noted.

ERP-based NFB, targeting specific ERP components such as the P300, offers a more personalised approach. Although promising, with studies showing cognitive improvements and neural changes, ERP-based methods are less flexible and not as extensively researched as rhythm-based protocols. The need for specific scenarios to elicit ERPs and slower feedback compared to rhythm-based methods are among the limitations.

Personalisation of NFB is also discussed, highlighting the importance of adjusting feedback thresholds and task difficulty to maintain participant motivation and engagement.

In conclusion, while ERP-based NFB holds potential for enhancing cognitive functions, particularly attention, current evidence is insufficient for its widespread application, and practical use is limited. The challenges identified include the lack of well-controlled, large-scale studies, limited research on real-world applications, and the practical issues of system setup and training duration.

To address these challenges, this thesis focuses on demonstrating the applicability of ERP-based NFB in practical settings. By conducting rigorously controlled studies, this research aims to provide robust evidence for the efficacy of ERP-based NFB in enhancing attention. The use of ILC is explored to optimise training efficiency and reduce the time required to achieve meaningful results, making the system more suitable for real-world use. Additionally, the thesis investigates ways to streamline the system setup by reducing the number of electrodes, thereby enhancing usability and scalability. These efforts aim to bridge the gap between laboratory research and real-world applications.

By directly addressing these limitations, this thesis seeks to establish a robust foundation for the practical application of ERP-based NFB, paving the way for its broader adoption in various real-world contexts.

3

ITERATIVE LEARNING CONTROL

3.1 INTRODUCTION

As discussed in Chapter 2, ERP-based NFB training has shown promise. However, a significant usability issue with NFB in general is the long training time to see results. This issue is exacerbated in ERP-based NFB due to the trial averaging needed for robust detection of ERPs. Not only does trial averaging slow down the training, but it also significantly reduces the feedback frequency, possibly making it challenging for users to recognise and adjust their brain activity effectively. As mentioned in Section 2.5, the number of trials in ERP-based systems can modulate task difficulty. Personalising task difficulty in NFB training, which can significantly improve motivation and user engagement, by adapting the number of trials may therefore accelerate the training and increase feedback frequency.

In this thesis, ILC is explored as a means to adapt task difficulty in ERP-based NFB training, potentially enhancing its efficiency and usability. ILC is a control technique designed to learn from past experiences, specifically applied to repetitive systems, where the same task is repeated multiple times, with each repetition referred to as a run. Its inherent feedback structure allows it to naturally adapt to small system changes. NFB training similarly involves task repetition within a system that is inherently complex and time-varying due to factors such as learning and fatigue, changes to which ILC can effectively adapt. Due to its learning capabilities, ILC can automate task difficulty adaptation, reducing the need for human intervention to manually adjust task conditions, thereby further enhancing system efficiency and usability.

Inspired by the need for precision in industrial robots performing repetitive pick-and-place tasks, ILC was independently developed by Uchiyama [105] and Arimoto et al. [106] in the late 1970s and 1980s. ILC assumes perfect repetition of tasks and aims to eliminate tracking errors over multiple runs by considering past inputs and errors to determine the new input for the next run, thus learning from previous experiences. Initially used in industrial robotics and semiconductor manufacturing [107], ILC has found applications in the biomedical world, including exoskeleton control [108] and stroke rehabilitation [109].

ILC can be generally described by

$$u_{k+1}(\cdot) = f(e_{k+1}(\cdot), ..., e_{k-s}(\cdot), u_k(\cdot), ..., u_{k-v}(\cdot)), \tag{3.1}$$

where $k = 1, 2, ..., N$ is the run index, u_k is the input in run k and e_k is the error between the reference r and the output y_k [110], i.e.

$$e_k(\cdot) = r(\cdot) - y_k(\cdot). \tag{3.2}$$

s and v in Equation (3.1) determine the order of the ILC algorithm. For instance, in a first-order algorithm, both s and v are unity, meaning that the algorithm relies solely on information from the most recent run [110]. On the other hand, higher-order ILC algorithms incorporate data from multiple previous runs [107]. The control law can be based exclusively on historical data, known as previous-cycle learning, or it might include the error from the current run, referred to as current-cycle learning. Some approaches even combine past and current run data [107], or use predictions of future errors [110].

First-order, previous-cycle, ILC is the most common and often takes the form of Equation (3.3) [111]:

$$u_{k+1}(t) = u_k(t) + f(e_k(t+1)), \tag{3.3}$$

where t is discrete time and k is the run number.

In comparison, typical feedback control can be expressed like this:

$$u_k(t) = f(e_k(t-1)). \tag{3.4}$$

The primary aim of ILC design is to eliminate the tracking error $e_k(t)$, meaning the control input u_k should be adjusted so that $\lim_{k\to\infty} e_k = 0$ [110]. This process is referred to as convergence of the tracking error. For ILC to successfully converge, it typically requires that the reference signal remains constant across all runs, each run has the same finite length, and the initial conditions are reset identically at the beginning of each run [110]. However, some algorithms have been developed that relax one or more of these constraints, offering more flexibility in certain applications [112–114].

There are a number of ways to determine the function f in Equations (3.1) and (3.3). ILC algorithms can use a model of the system, or can be data-driven, meaning they do not require (accurate) knowledge of the underlying system. An overview of these algorithms is presented in Section 3.2. Common application areas of ILC, both in the industrial and biomedical domains, are discussed in Section 3.3.

3.2 TYPES OF ITERATIVE LEARNING CONTROL ALGORITHMS

This section provides a broad overview of various ILC update laws, distinguishing between classes of algorithms.

As the name suggests, in model-based ILC, the update law is derived from the system model. Considering the equation

$$y = Gu, \tag{3.5}$$

where G is a linear model of the system, this means that the update law makes use of G in some way. In Equation (3.5), G is a general system mapping that transforms the input u to the output y. While G could be a specific transfer function or a super-vector form of a state-space model [110], here it is used as a generic placeholder to illustrate the concept of input-output transformation. The exact nature of G is not specified as it is meant to represent a variety of possible system dynamics.

Fundamentally, model-based ILC algorithms use the inverse of G or an approximation thereof [111]:

$$u_{k+1}(t) = u_k(t) + \gamma G^{-1} e_k(t+1), \tag{3.6}$$

where γ is a constant learning gain. Since G transforms u from input to output space, G^{-1} transforms the error e from output to input space. Assuming that the system model is perfect, G is invertible, and $\gamma = 1$, model inverse-based ILC algorithms would lead to the elimination of the tracking error from the first run. However, in practice, this ideal set of conditions is rarely met [111]. Model inverse-based ILC can achieve fast convergence with an accurate system model, although determining the model inverse is not always possible [111].

On the other hand, model-free algorithms do not require an explicit system model, making them more flexible but sometimes less precise. One of the simplest, and first, model-free ILC algorithms is the Arimoto algorithm:

$$u_{k+1}(t) = u_k(t) + \gamma e_k(t+1), \tag{3.7}$$

where γ is a constant [106]. This P-type algorithm was popular due to its simplicity, however, monotonic convergence of the tracking error is not guaranteed [115]. This means that the tracking error may increase before it decreases. Other algorithms may also use the derivative and/or integral of the error to create PD- and PID-type algorithms [107]. The performance of these model-free algorithms is highly dependent on the choice of γ.

The most popular class of ILC algorithms is based on optimisation. In optimal ILC, u_{k+1} is determined by minimising a cost function that uses predictions of the error for different inputs or controller parameters. This means that these algorithms do not directly use a system model in the update law, but require a model to make predictions. One popular optimal ILC algorithm is norm-optimal iterative learning control (NOILC):

$$J_{k+1}(u_{k+1}) = \|e_{k+1}\|^2 + \epsilon^2 \|u_{k+1} - u_k\|^2, \tag{3.8}$$

where ϵ is a constant [111]. Similarly, in parameter-optimal iterative learning control (POILC), a cost function is minimised to determine γ in Equation (3.7) for each run [110],

$$J_{k+1}(\gamma_{k+1}) = \|e_{k+1}\|^2 + \epsilon^2 \|\gamma_{k+1}\|^2. \tag{3.9}$$

POILC solves the problem of asymptotic convergence in the classic Arimoto algorithm [110].

In data-driven ILC, a system model is estimated based on input-output relationships observed in previous runs [116]. This estimated model can then be employed in optimal ILC algorithms (e.g. [117]), leveraging the convergence properties of these algorithms without requiring a pre-defined system model.

The choice of algorithm depends on whether a system model is available or can be determined, the level of uncertainty, the presence of disturbances, and the specific requirements of the control system.

Model-based ILC algorithms typically achieve strong performance with monotonic convergence, making them particularly valuable in applications requiring high precision. Although ILC can handle some degree of model uncertainty, significant uncertainties may substantially impair the controller effectiveness. In contrast, model-free ILC, while potentially less accurate than model-based approaches, offers the advantages of reduced complexity and easier implementation.

3.3 TYPICAL APPLICATIONS OF ITERATIVE LEARNING CONTROL

While ILC was inspired by industrial robots carrying out pick-and-place tasks and initially found application in manufacturing, its application areas today are widespread. This section will give an overview of how ILC has been used in industrial and biomedical domains.

3.3.1 *Industrial Applications*

In the industrial sector, ILC has been used extensively in robotics [107]. Robotic systems often have very complex dynamics that are difficult to model accurately, making ILC an attractive choice for repetitive tracking tasks. ILC has been applied to control actuators in various robot types, including robotic arms [118], wheeled robots [119], and gantry robots [120].

Another common application area is batch processes, where high precision is required in the presence of model uncertainties and disturbances. Examples of batch processes where ILC has been applied include injection moulding, where a specific velocity profile must be maintained [121, 122], and the motion control of wafer scanners [123].

ILC has also been applied to urban traffic management, specifically for controlling traffic signals [124, 125]. Another example of ILC applications are the control of autonomous vehicles [126] and trains [127, 128].

3.3.2 *Biomedical Applications*

ILC is also widely used in the biomedical domain, particularly in training and rehabilitation, where repetitive tasks are common. These systems, which include medical devices interacting with human users, often change over time due to factors like learning and fatigue. ILC can naturally adapt to these variations. Additionally, model-free and data-driven ILC does not require detailed system knowledge, an important feature in the biomedical domain where system models can be highly complex and difficult to identify.

The main application of ILC in the biomedical domain is the control of functional electrical stimulation (FES) for rehabilitation. In FES, the nerves in the body are stimulated with electrical impulses to trigger muscle movements. This stimulation has been shown to be effective for rehabilitation of paralysis or weakness and gait assistance in patients following stroke or spinal cord injury [129]. ILC algorithms have been used to control the intensity of FES for upper [130, 131] and lower [132] limbs.

In a related field and drawing on the initial application area of industrial robotics, ILC has been used for robotic rehabilitation [133, 134].

3.4 SUMMARY

ILC has traditionally been applied to perfectly repetitive tracking problems, where it learns from past runs to eliminate tracking errors over time while adapting to minor system changes. Due to its learning abilities, ILC can achieve good control performance without requiring an accurate system model. Although initially used in industrial processes, ILC has also found applications in the biomedical field, particularly in rehabilitation. These application areas share the challenge of dealing with highly complex, repetitive, systems where developing precise models is difficult, and where dynamics often vary between runs.

As discussed in Chapter 2, the efficacy of NFB training may be significantly influenced by task difficulty, which should therefore be carefully controlled. Given the parallels between NFB training and applications that are typically targeted for ILC, it makes sense to use ILC to adapt task difficulty, even though NFB is not a trajectory tracking problem. This approach leverages the feedback mechanism inherent in ILC to adaptively adjust training parameters, ensuring that the training remains challenging yet achievable.

While model-based ILC algorithms typically lead to better control performance, the complexity involved in modelling an NFB training system makes a model-free approach more suitable for this novel application. A model-free algorithm is also less complex, which can im-

prove the usability and acceptability of the system. The development and evaluation of an ILC controller for task difficulty adaptation are discussed in the next chapter.

Part II

SYSTEM DEVELOPMENT

4

DEVELOPMENT OF AN ADAPTIVE P300-BASED NEUROFEEDBACK TRAINING SYSTEM

4.1 SYSTEM OVERVIEW

As discussed in Chapter 2, ERP-based NFB training remains underexplored, despite promising evidence of its effectiveness. The P300 speller, in particular, appears to be an efficient tool for P300-based NFB, enhancing attention in healthy adults [95, 96].

In this thesis, the use of ERP-based NFB for attention enhancement in healthy adults, using a P300 speller and ILC for the personalisation of task difficulty, is investigated. This chapter provides an overview of the NFB system, as illustrated in Figure 4.1, and describes its development.

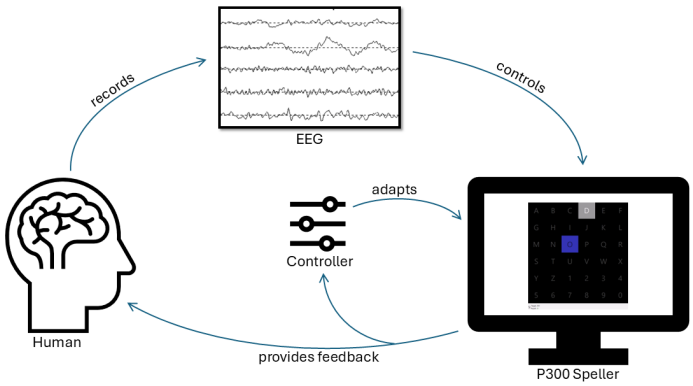


Figure 4.1: Overview of the neurofeedback (NFB) training system.

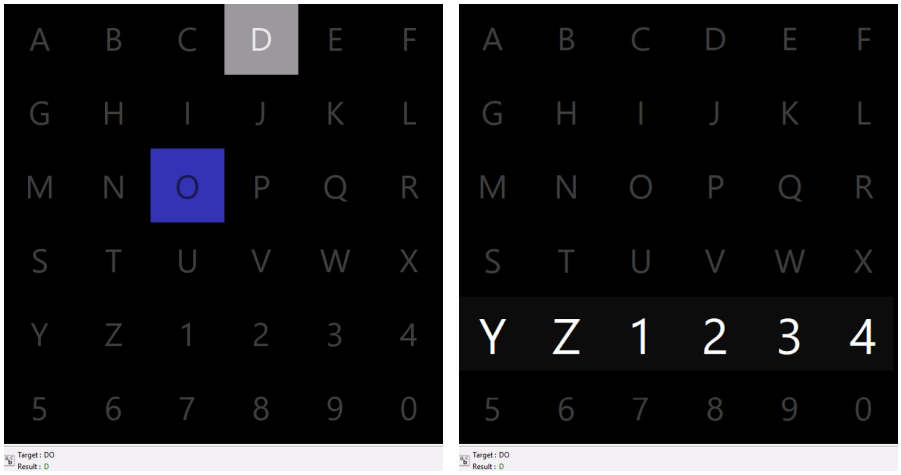
The system comprises three main components:

- **The human:** The person undergoing NFB training is an integral part of the system. They perform a task while their EEG is recorded, and feedback is provided based on this data.
- **The P300 speller:** The P300 speller serves simultaneously as the training task and feedback mechanism. A brief explanation of how a P300 speller works and how it can be used for NFB training, along with an overview of possible design choices, is provided in Section 4.2. This section also outlines the development of the speller.
- **The controller:** The ILC controller adapts the training difficulty based on the user performance. A model of performance in the P300 speller is developed to facilitate controller design, enabling

simulation testing. The model and controller development are detailed in Section 4.3.

4.2 P300 SPELLER DESIGN AND IMPLEMENTATION

The P300 speller is a widely used BCI application that uses the P300 component. Farwell and Donchin [135] developed the P300 speller in the 1980s as a communication tool for individuals with severe paralysis, including patients with locked-in syndrome. The speller functions like an on-screen keyboard, presenting the user with a grid of symbols, such as letters and numbers, where each symbol in the grid is highlighted or flashed. These flashes typically occur per row and column in the grid. The user focuses on the symbol they wish to select, generating a P300 wave when the target symbol is flashed. In this thesis, the user is instructed to select a specific letter, which is highlighted in blue, by counting every time the letter is flashed. This method is known as copy-spelling. Figure 4.2 shows what the P300 speller looks like.



(a) (b)

Figure 4.2: Screenshots of P300 speller used in this thesis. (a) The previously selected letter is highlighted in grey, regardless of whether it was correct or not, while the next target letter is highlighted in blue. (b) Row 5 is being flashed, indicated by an increase in font size and a change in font colour to white. The previous and current target letters, as well as previously selected letters, are displayed at the bottom of the window.

To enhance the SNR of the ERP, each symbol in the grid is flashed multiple times, allowing the EEG signals to be averaged across all flashes. This averaging process improves the reliability of the recorded signals, facilitating a more accurate evaluation of the P300 response. EEG signals, in response to each flash, are called trials. Target trials refer to EEG signals associated with the flashes of the row and col-

umn containing the symbol the user wants to select. This means that for each flash, per row and column, there are 2 target trials. Nontarget trials refer to EEG signals corresponding to the flashes of all other rows and columns, resulting in 10 nontarget trials for each flash per row and column.

Figure 4.3 shows the mean EEG response to target and nontarget trials for four different individuals. Both types of trials result in oscillations, representing steady-state VEPs due to the flashes. However, target trials result in higher amplitude oscillations. These higher amplitudes correspond to the P300 (positive amplitudes) and N200 (negative amplitudes) components.

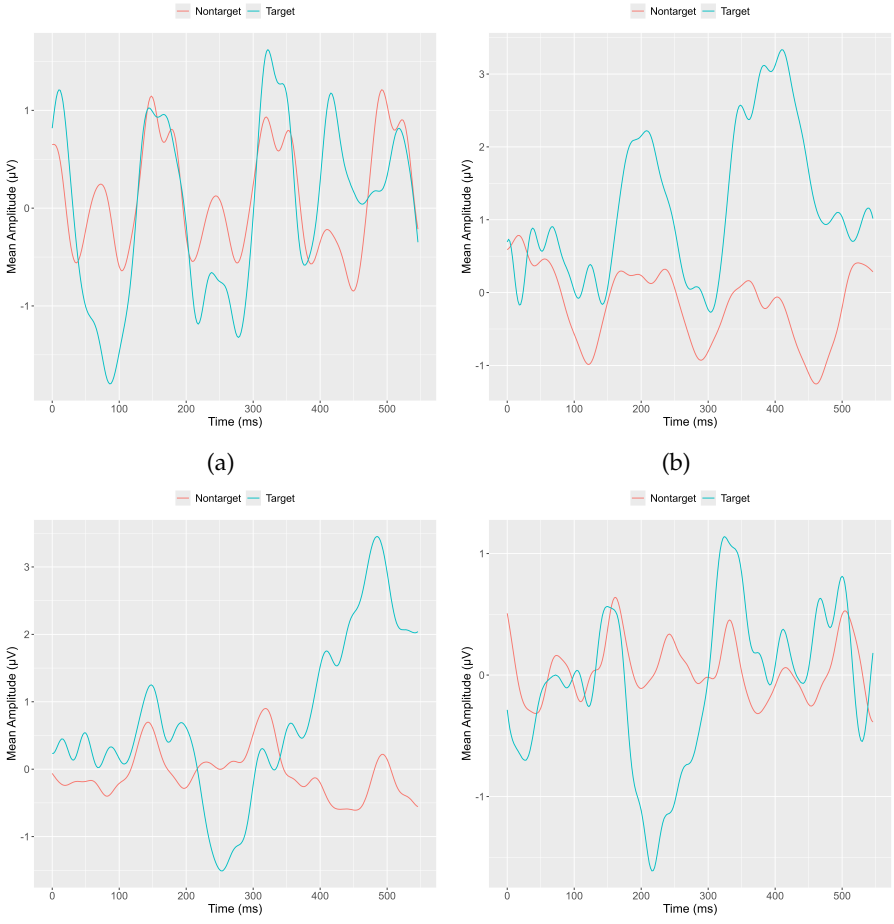


Figure 4.3: Example of EEG responses to target and nontarget trials in the P300 speller for four different individuals from the study described in Chapter 5. The x-axis represents time from stimulus onset (0 ms) to 550 ms.

Traditionally, the number of flashes per row and column is fixed. However, as discussed in Section 2.4.3, the P300 speller has been used as an NFB training tool, with the number of flashes serving as the task difficulty parameter of the speller [96, 98]. This section discusses the

possibilities of turning a P300 speller into an NFB tool, and how it was implemented.

4.2.1 *Design Choices*

By definition, in NFB training, the user is provided with feedback based on their real-time brain activity. While not necessarily a requirement of NFB training, as seen in Section 2.5, training difficulty is often adapted. Therefore, to turn a P300 speller into an NFB training tool, it is necessary to extract some output from the P300 speller that reflects user brain activity. This output is then used to provide feedback to the user and to adapt the conditions of the spelling task to modulate its difficulty. Several possible outputs and task conditions are considered, and the benefits and drawbacks of each are discussed in the following sections.

4.2.1.1 *Possible Outputs*

There are two possible types of metrics to use for feedback: the user performance and their brain signals. The performance in the P300 speller is measured by spelling accuracy, which is the percentage of correctly identified letters in a run. A run in the P300 speller is defined as any sequence of letters or symbols spelled in one go, which could be a few letters, a word, or a sentence. In the remainder of this thesis, a run is always a single word. The benefits of using spelling accuracy are its ease of calculation and interpretability. As a standard metric in the P300 speller, spelling accuracy is well-documented in the literature. However, spelling accuracy only indirectly reflects the user P300 wave and may be influenced by external factors, such as movement artifacts. For example, a user might subconsciously move their head every time the target letter flashes, which could be detected by the EEG system.

Regarding brain signals, there are four main metrics considered: P300 amplitude, P300 latency, total power, and alpha power. P300 amplitude is positively correlated with cognitive abilities [46]. While P300 amplitude is commonly used as a strength metric, P300 amplitude calculation is not well-defined. Some approaches measure the voltage difference between the baseline and peak within a specified time window, while others use the voltage difference between the positive and negative peaks. Variations in the definition of the baseline and the chosen time window can greatly affect the results [26]. Total power, which quantifies EEG signal strength across all rhythms within a specified time window, provides an alternative that can indirectly indicate P300 strength without requiring peak definition. However, like P300 amplitude, total power remains sensitive to specific signal processing choices.

P300 latency is the time between stimulus onset and the point at which the maximum positive amplitude peak occurs within a specified time window. Again, the choice of an appropriate time window is crucial but not well-defined. Similar to P300 amplitude, latency is known to correlate with cognitive abilities [46].

The final brain signal metric considered is alpha power, representing the strength of EEG signals in the alpha rhythm. Although alpha power is associated with attentional engagement and correlates with the P300 [46], it is not directly targeted by the P300 speller, and thus changes in alpha power may not reliably reflect training progress. For this reason, alpha power is less suitable as a feedback metric in the context of the P300 speller.

Given the lack of clear standards for calculating P300 amplitude and latency, the sensitivity of all brain signal metrics to specific signal processing steps, and the established efficacy of spelling accuracy as feedback in ERP-based NFB studies [96, 98], it was decided to use spelling accuracy. The objectivity and simplicity of spelling accuracy will facilitate the design of an adaptation algorithm that is computationally efficient and interpretable.

Nevertheless, all brain signal metrics discussed here will be employed in subsequent sections to assess underlying brain processes and cognitive engagement during training. These metrics provide complementary insights into the brain's response, allowing a more comprehensive evaluation of training effects.

4.2.1.2 *Possible Task Conditions*

There are two possible task conditions that potentially modulate difficulty: the timing of flashes and the number of flashes. Both conditions theoretically affect P300 strength, meaning that when these are adjusted, the user must improve their focus to maintain performance, thereby enhancing attention. The conditions also make the training more fast-paced, which may help keep users engaged.

By adjusting the inter-stimulus interval (ISI), which is the time between flashes, or by reducing the number of flashes, the target-to-target interval (TTI), i.e. the time between target stimuli, is modified. It is known that increasing the TTI increases P300 strength [46]. Conversely, decreasing the TTI may reduce P300 strength. The ISI is usually a fixed parameter of the P300 speller, so it is unclear how the ISI should be adapted based on the user's performance. Additionally, the ISI is limited to a few hundred milliseconds, making it uncertain whether changing the timing of the flashes would sufficiently alter P300 strength.

Reducing the number of flashes decreases the SNR, since there are fewer trials to average, in addition to decreasing the TTI. Unlike timing adjustments, changing the number of flashes is more straightforward, with a clear range and minimum step size. Evidence suggests

that adapting the number of flashes influences the efficacy of NFB training using a P300 speller [98].

Changing the number of flashes to adapt the speller difficulty is an obvious choice, as it potentially modulates the P300 strength in two ways: by decreasing the SNR, and the TTI. It is also a preferable choice, as it is more clearly defined, and there is supporting evidence for its use in modulating task difficulty [96].

4.2.2 *Implementation*

An open-source software platform for BCI applications, OpenViBE [136], is used to implement the P300 speller. The software comes with two versions of a P300 speller. The only difference between the two versions is the use, or absence, of a spatial filter. The standard version of the OpenViBE P300 speller does not use a spatial filter; instead, it uses all EEG channels directly for target classification. In contrast, the xDAWN P300 speller uses the xDAWN spatial filter to reduce the incoming EEG channels to a predefined number of components [137]. The spatial filter is designed to optimise the SNR in ERPs. It uses the covariance matrices of the signal (in this case, target trials) and noise (nontarget trials) to calculate spatial filters that are linear combinations of the EEG channels [137]. Modified versions of both of these spellers are used in this thesis. How the spellers work, and the modifications, are briefly outlined in this section.

Both versions of the speller use a linear discriminant analysis (LDA) classifier to distinguish between target and nontarget trials [138]. In this implementation, the trials are 600 ms long, starting with stimulus onset. This classifier is trained for each user at the start of every experimental session, using EEG signals from two calibration runs.

In OpenViBE default P300 spellers, each individual trial is classified as either target or nontarget, before voting is conducted to select the row and column with the most target classifications. However, in other speller implementations, and more commonly, trials are first averaged, and only then classified into target and nontarget rows and columns. Since this is the more common classification approach in a P300 speller, the code is modified to reflect this. In this classification approach, the average of all trials is calculated for each row and column, before each row and column is classified as target or nontarget. The row and column with the highest probability of belonging to the target class are then selected, even if the probability is below 50%.

Additionally, OpenViBE default P300 spellers do not calculate the spelling accuracy at the end of a run. Spelling accuracy calculation is implemented, as spelling accuracy is needed for feedback and the task difficulty adaptation module. The spelling accuracy is calculated by checking whether the selected row and column match the actual target row and column, and increasing a counter if they do. At the

end of a run, the count of correct letters is divided by the total letter count to get the spelling accuracy for that run.

Finally, to facilitate comparison of the task difficulty adaptation module developed in this thesis with an existing approach [96], spelling accuracy must be calculated for each number of flashes up to the one actually used. This is referred to as *cumulative spelling accuracy* in this thesis. This means that classification and subsequent spelling accuracy calculation need to be performed using only the first 12 trials (2 target and 10 nontarget trials), then the average of the first 24 trials, and so on, until the average of all trials is used.

4.3 DEVELOPMENT OF TASK DIFFICULTY ADAPTATION MODULE

4.3.1 *Model and Simulation of Performance in a P300 Speller*

This section details the development of a phenomenological model of performance in a P300 speller. This model is created to facilitate easy and frequent testing of the controller, described in Section 4.3.2. Testing the controller through human experiments would be impractical and could lead to biased comparisons due to unavoidable confounders. Therefore, developing a model and using simulation is the preferred approach.

Another reason for the model development is that ILC can be model-based, as discussed in Chapter 3. Having a system model therefore opens up the possibility of using it in the controller.

To understand the behaviour that should be modelled, the dataset described in Section 4.3.1.1 is analysed. The model structure is then explained in Section 4.3.1.2, followed by an overview of the model training process in Section 4.3.1.3. The simulation developed using this model is outlined in Section 4.3.1.4.

4.3.1.1 *Overview of the Training Dataset*

The Akimpech dataset [139] is a publicly available dataset containing raw EEG data from 30 healthy individuals who completed several runs of a P300 speller task, along with classifier weights.

The experiments consisted of four sessions. In the first session, three runs were completed without feedback to collect calibration data. In the second session, all participants copy-spelled the word "SUSHI" with 15 flashes per row and column. In the third session, all participants completed three free-spelling runs with 15 flashes, allowing them to choose which words to spell. In the fourth session, both the number of runs and the number of flashes varied for each participant. The number of letters spelled in each of these free-spelling runs ranged between 2 and 15. Some participants did not participate in the fourth session, resulting in a range of 4 to 11 runs across participants

in sessions two to four. Only data from sessions two to four are used in this analysis, as the first session did not include feedback.

9 of the 30 participants in the dataset are excluded due to poor or incomplete data. The remaining 21 participants are split into 16 training and 5 validation participants.

The cumulative spelling accuracy is calculated for each participant. Figure 4.4 shows the mean spelling accuracy, termed J_1, for all participants in the first 4 runs.

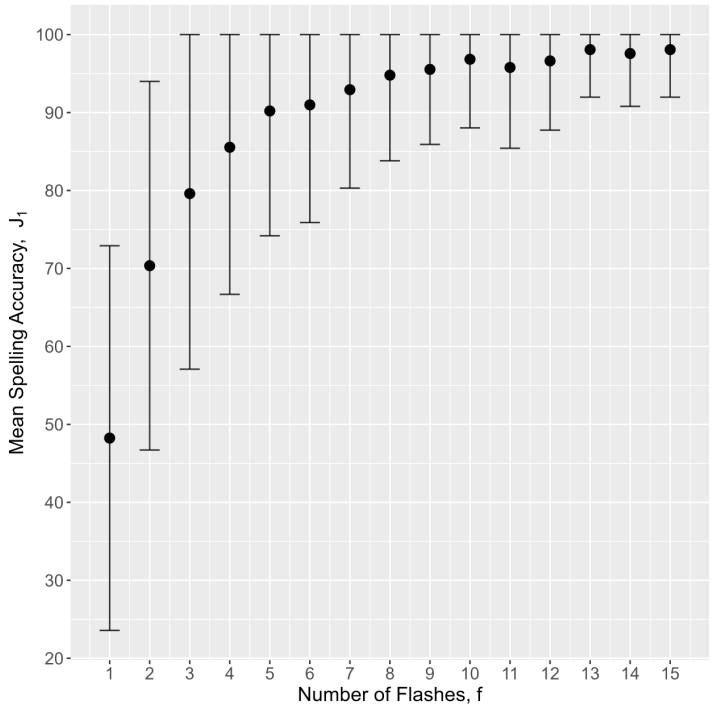


Figure 4.4: Mean spelling accuracy, J_1, for all participants, over the first 4 runs.

J_1 represents the absolute performance of a participant. However, in NFB training, absolute performance is not the primary focus. For this reason, spelling accuracy is divided by the number of flashes, denoted as $J_2 = J_1/f$, where f is the number of flashes. J_2 indicates performance relative to task difficulty and is the metric targeted for maximisation during training. The mean J_2 for all participants, over the first 4 runs, is shown in Figure 4.5.

Figure 4.6 illustrates the progression of J_2 for participant 20 over all runs, depicting the participant's learning curve. This learning evolution is what the phenomenological model aims to capture.

Based on the observations from this dataset, the following behaviours are identified:

- *Behaviour 1:* If performance is perfect and task difficulty remains constant, then performance does not change. While long-term

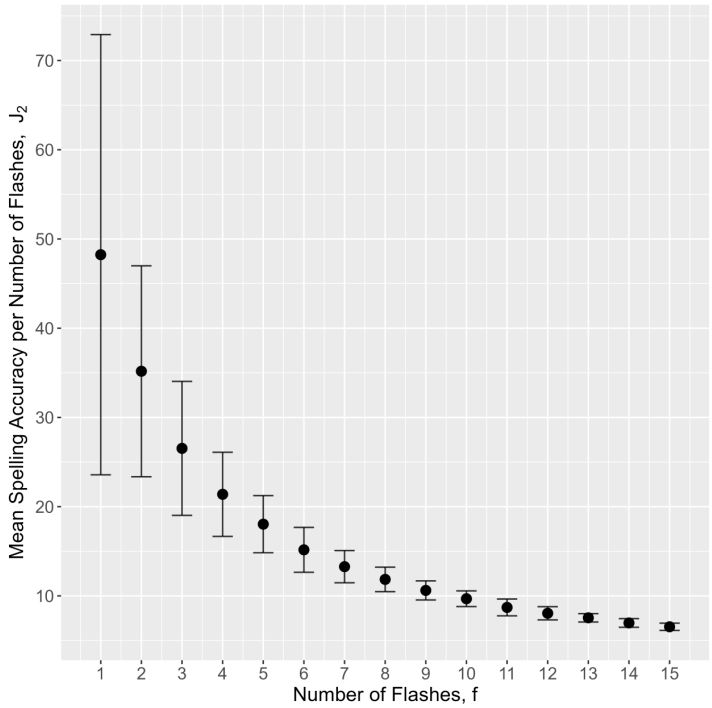


Figure 4.5: Mean spelling accuracy per number of flashes, J_2, for all participants, over the first 4 runs.

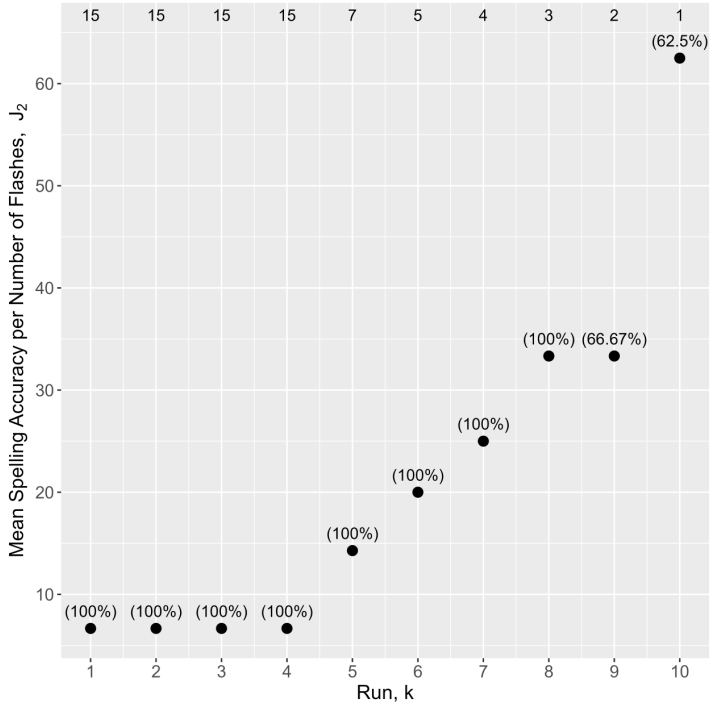


Figure 4.6: Example progression of J_2 over all runs. The sequence of flashes used is specified at the top of the figure. The absolute spelling accuracy, J_1, for each run is given in brackets.

factors like fatigue or loss of motivation might eventually alter performance, this behaviour holds for the few runs simulated.

- *Behaviour 2:* If performance is not perfect with a constant difficulty level, performance typically improves.
- *Behaviour 3:* Changes in task difficulty impact performance.

Additionally, it is assumed that there exists an optimal task difficulty level that challenges the individual without causing excessive frustration [140]. Therefore, *Assumption 1* posits that performance improvement is maximized at this optimal task difficulty and decreases as the actual task difficulty deviates further from this optimal level.

4.3.1.2 *Model Structure*

The proposed model tracks the progression of participant performance (in terms of spelling accuracy per number of flashes) over P300 speller runs, representing the participant learning curve during training. Spelling accuracy per number of flashes can be viewed as an indirect measure of participant attention level, or overall cognitive ability.

The model is defined by the following equation:

$$J_2(k) = J_2(k-1) + \beta e(k-1) + \gamma \Delta t(k), \tag{4.1}$$

where $J_2(k)$ is the spelling accuracy achieved in run k divided by the number of flashes $f(k)$ used in run k. β and γ are parameters determined in Section 4.3.1.3. $e(k)$ represents the error in a run, defined as

$$e(k) = \frac{1 - J_1(k)}{f(k)}, \tag{4.2}$$

and $\Delta t(k)$ is the change in task difficulty in run k relative to the optimal task difficulty, defined as

$$\Delta t(k) = \frac{f(k-1) - f(k)}{1 + |f^0(J_2(k-1)) - f(k)|}. \tag{4.3}$$

f^0 is the optimal task difficulty, which is defined as the number of flashes that maximises $J_2(k-1)$.

This model satisfies *Behaviour 1* since $e(k-1)$ and $\Delta t(k)$ are zero if the performance in run $(k-1)$ is perfect and the task difficulty does not change. *Behaviours 2 and 3* are satisfied due to the inclusion of the second term, $e(k-1)$, and the third term, $\Delta t(k)$, respectively. Finally, *Assumption 1* is addressed by incorporating f^0 in $\Delta t(k)$. The model thus fulfills all the behavioural requirements identified in Section 4.3.1.1.

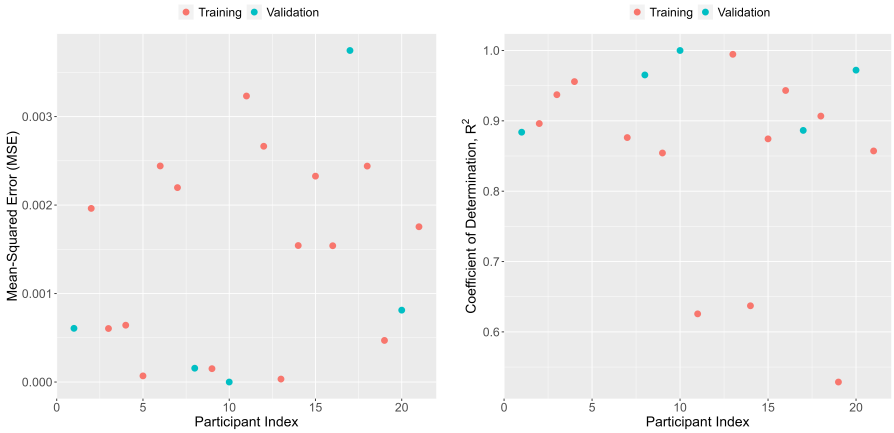
4.3.1.3 *Parameter Estimation*

The parameters β and γ in Equation (4.1) are estimated using least-squares, subject to the constraint that the maximum possible spelling accuracy is 100%. The parameter values obtained using least-squares are listed in Table 4.1.

Table 4.1: Estimated model parameter values.

Parameter	Value
β	1
γ	0.068

Figure 4.7 shows the model fit for all participants in terms of mean-squared error (MSE) and the coefficient of determination (R^2), which assesses how well the model explains variance in the data. Figure 4.7b excludes the R^2 scores for participants 5, 6 and 12 to allow for a better scale. The R^2 values for these participants are -0.3333, -8.6550 and -1.0313, respectively.



(a) (b)

Figure 4.7: Model fit for all participants. (a) Mean-squared error (MSE). (b) Coefficient of determination, R^2. Participants 5, 6 and 12 are excluded to allow for better scale.

An average MSE of 0.0015 on the training data, and 0.0011 on the validation data, is achieved. The average R^2 on the training data is 5.42% (83.75% when the same participants as above are excluded) and 94.15% on the validation data.

This discrepancy between the R^2 score on training and validation data can be explained by the fact that the randomly selected validation participants performed reasonably well in the P300 speller task, whereas some training participants, particularly participants 5, 6, and

12, performed poorly. This suggests that the model is good at capturing strong performance, but struggles to simulate poor performance.

Figure 4.8 shows the actual data and the simulated learning curve of validation participants 8 (second-best model fit in terms of MSE) and 17 (worst model fit in terms of MSE). The best validation participant (index 10) is not shown because they only completed the first four runs with a constant number of flashes and 100% spelling accuracy, leading to perfect model fit by default.

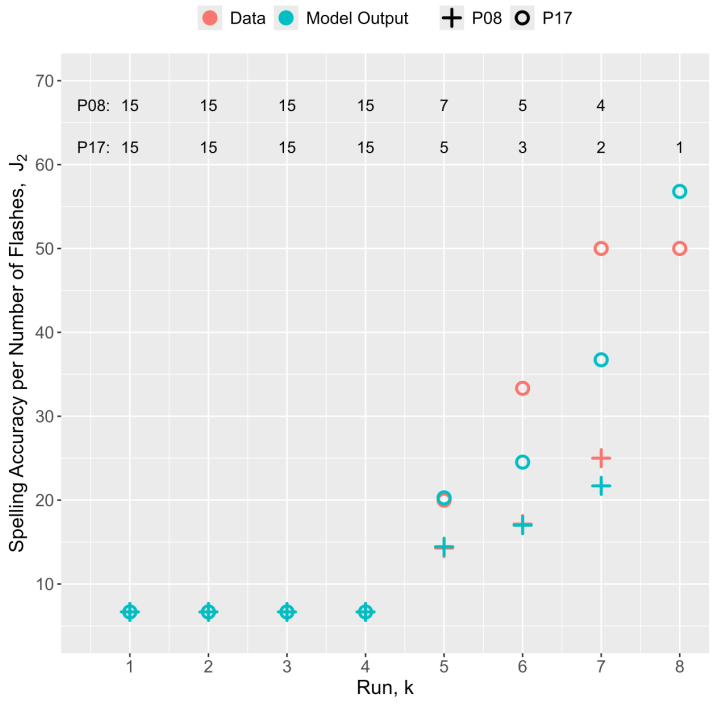


Figure 4.8: Model output and measured data of validation participants 8 and 17. The sequence of flashes used is specified at the top of the figure.

While the model can accurately simulate the learning curve of participant 8, there is a larger error in the last three runs for participant 17. However, the model still captures the overall trend of the learning curve.

4.3.1.4 *Simulation*

This section outlines the development of a performance simulator in a P300 speller using the model described in Section 4.3.1.3. The simulation is developed to enable quick and easy testing, and initial tuning, of the controller, described in Section 4.3.2.

For the simulation, noise, characterised by the residuals in J_1 from the model described in Section 4.3.1.3, is added to the model to ensure that different runs of the simulation are unique and to represent the inter- and intra-participant variations in performance observed

in reality. Figure 4.9 shows a histogram of the residuals for both the training and validation data. As can be seen, the distributions for both the training and validation data are similar, with most residuals centered around zero. Several different distributions are fitted to the histogram of the residuals to model the noise. It can be seen that a Cauchy distribution models the residuals most accurately.

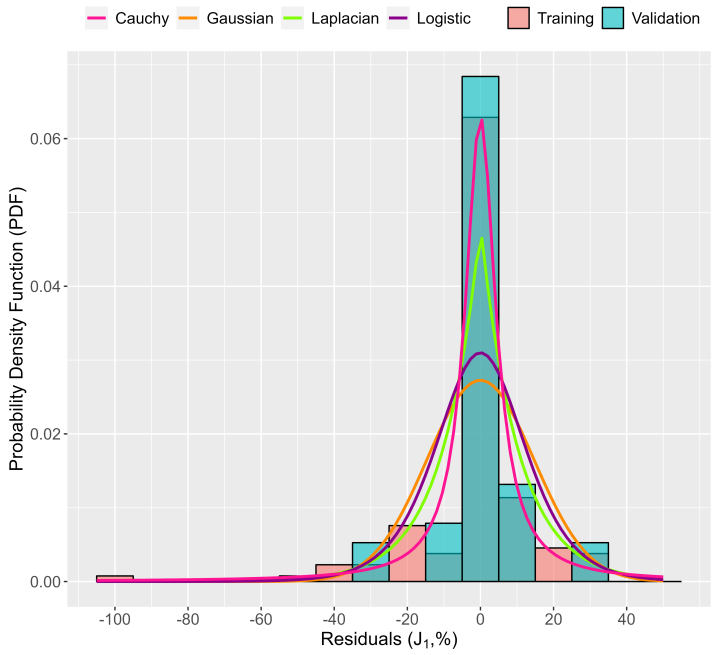


Figure 4.9: Histogram of model residuals (J_1, %) with possible distributions.

The Cauchy distribution is therefore chosen for the simulation, with the following specification:

$$f(x) = \frac{1}{\pi\gamma(1 + (\frac{x-x_0}{\gamma})^2)}, \tag{4.4}$$

where x_0 is the location parameter, defined as the peak of the distribution, which in this case is 0, and γ is the scale parameter, defined as the half-width at half-maximum, which is 5.06 for the distribution of the residuals. The equation for the Cauchy distribution therefore becomes:

$$f(x) = \frac{1}{5.06\pi(1 + (\frac{x}{5.06})^2)}. \tag{4.5}$$

To implement a simulation, an arbitrary starting run with 10 flashes and a J_1 performance per flash, as shown in Table 4.2, is defined. This performance is chosen to resemble the mean curve observed in the Akimpech dataset. 10 runs are then simulated using the model described in Section 4.3.1.3 with additive noise drawn from the Cauchy distribution (Equation (4.5)) to determine performance in a given run. This simulation is then used for the design of an ILC controller

and subsequent evaluation of different task difficulty adaptation approaches, discussed in Section 4.3.2.

Table 4.2: Starting performance in simulation.

Number of flashes	Spelling accuracy
1	0.4
2	0.5
3	0.6
4	0.7
5	0.8
6	0.9
7	1
8	1
9	1
10	1

4.3.2 *Iterative Learning Controller Development*

As discussed in Section 3.4, using ILC to adapt the task difficulty in NFB training is a logical approach. As mentioned in Section 4.2.1, the number of flashes per row and column is chosen as the task difficulty parameter in the P300 speller, which is adapted based on spelling accuracy. Recalling the typical framework of ILC:

$$u_{k+1} = u_k + \epsilon g(e_k), \tag{4.6}$$

this means that the input u represents the number of flashes, and the error e represents the percentage of incorrect letters in a run, i.e. $e_k = 1 - J_{1,k}$. $\epsilon g(e_k)$ therefore is the change in number of flashes based on the error. Unlike traditional ILC, both u and e are scalars.

This section describes the development of the ILC controller for task difficulty adaptation, as well as the results of a comparison to an existing approach and random difficulty levels.

4.3.2.1 *Control Design*

The first step in the control design is defining the requirements. If a person performs very well during training, the training should become more difficult, as strong performance could indicate that they are not sufficiently challenged. In the context of the P300 speller, this means that if the spelling accuracy is high, the number of flashes

should decrease, reducing the SNR and making the task more difficult. Conversely, if a person performs poorly, i.e. the spelling accuracy is low, the task should either remain the same or become easier. This means that the number of flashes should either stay the same or increase. The change in the number of flashes should be proportional to the spelling accuracy, i.e. the higher or lower the spelling accuracy, the greater the increase or decrease, respectively.

Another requirement for the controller is that it should be simple enough to be easily interpreted by end-users and clinicians, aiding in its acceptability and allowing for tuning without the need for an engineer or technician.

Different functions for $g(e_k)$ are tested in simulation based on the main requirement described above. The first function considered is

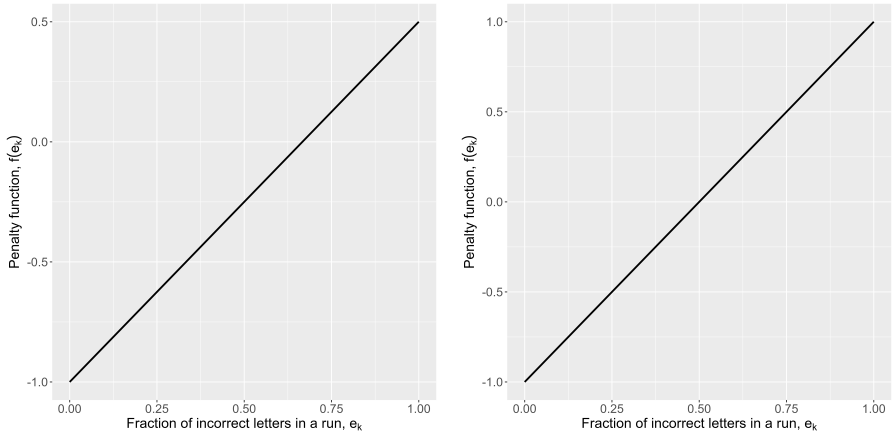
$$g(e_k) = 1.5e_k - 1, \tag{4.7}$$

shown in Figure 4.10a. This function results in a more aggressive controller, which increases task difficulty as soon as about one-third of a run is correct. The maximum step size for a difficulty increase is twice as large as the maximum step size for a difficulty decrease. This function could lead to user frustration due to low performance.

A similar, more balanced, function is also considered;

$$g(e_k) = 2e_k - 1, \tag{4.8}$$

shown in Figure 4.10b. With this function, the threshold for a difficulty increase or decrease is 50%, and the maximum step size is the same for both. This function could potentially be more acceptable to users, though it may result in slower training than the first function (Equation (4.7)).



(a) (b)

Figure 4.10: Penalty functions, $g(e_k)$, considered for the controller in Equation (4.6). (a) Equation (4.7). (b) Equation (4.8).

It is also considered to use a function that is flatter around the threshold, so that only small changes occur near the threshold while

larger step sizes happen with more extreme errors. This behaviour could be achieved with a piecewise linear function or a polynomial function. However, these more complex functions may complicate the algorithm without significantly improving performance, so they have not been pursued.

Instead, Equation (4.8) is chosen due to its more balanced behaviour and lower likelihood of causing frustration. This function also results in better performance in simulation compared to Equation (4.7) and a polynomial function.

In addition to the function $g(e_k)$, the parameter ϵ in Equation (4.6) needs to be determined. Since $g(e_k)$ is unity at 100% and 0% error, ϵ can be interpreted as the maximum difficulty increase or decrease. Different values for ϵ are tested in simulation, but it is ultimately decided that the maximum step size should depend on the current difficulty level. There should not be drastic changes when the difficulty is very high (i.e. the number of flashes is low), but the step size can be large when the training is easy, particularly at the beginning of a session when the optimal task difficulty is first being approached. It is found in simulation that $\epsilon = \frac{u_k}{2}$ leads to the best results.

This results in the following algorithm:

$$u_{k+1} = u_k + \frac{u_k}{2}(2e_k - 1), \tag{4.9}$$

which can be simplified to

$$u_{k+1} = u_k(e_k + 0.5). \tag{4.10}$$

4.3.2.2 *Comparison to Other Approaches in Simulation*

To evaluate the developed ILC controller, and test the hypothesis that using ILC to adapt task difficulty accelerates NFB training, a comparative simulation study is conducted using the simulation described in Section 4.3.1.4.

The ILC controller is compared to the task difficulty adaptation approach used in [96], referred to as the Benchmark algorithm, and to a random difficulty adjustment, referred to as the Random approach. The different approaches are explained in more detail in Section 5.3.

Early stopping, which personalises the number of flashes in the P300 speller (discussed in Section 2.5), is not simulated here. The primary reason is that the simulation used in this study focuses on performance across multiple runs, while early stopping typically requires EEG data to determine when to stop the task. Since EEG data are not simulated, applying early stopping is not feasible in this context.

Additionally, the goal of both the ILC and the Benchmark algorithm is to challenge the user and drive learning by continuously increasing task difficulty in response to the user's performance. In contrast, early stopping aims to optimise the number of flashes based

on the user's current ability, without necessarily pushing for continuous improvement. However, early stopping could complement ILC in cases where the task difficulty set by the ILC is too easy. In such cases, early stopping could help identify a more appropriate difficulty level, and this information could then be used by the ILC to adjust the difficulty for subsequent runs, ensuring an even more effective training experience.

The simulation is repeated for 50 iterations. The three metrics, J_1, J_2, and f, are then averaged over all runs to simplify the statistical analysis, as illustrated in Figure 4.11. Due to non-normality in the data, between-group differences are analysed using Kruskal-Wallis tests, followed by Tukey-Kramer Nemenyi tests, to investigate where significant differences lie. A detailed explanation of all statistical methods used in this thesis can be found in Appendix A.

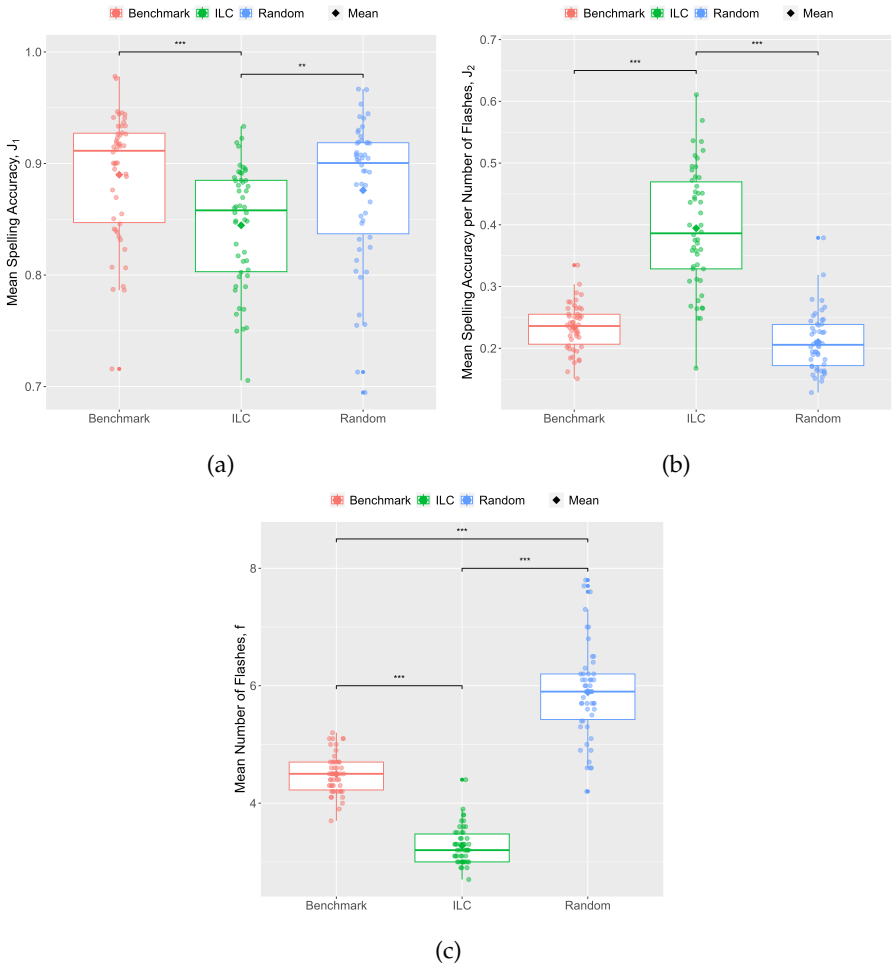


Figure 4.11: Simulation metrics calculated across runs for the different task difficulty adaptation approaches. (a) Mean J_1. (b) Mean J_2. (c) Mean f. Statistical analysis by Kruskal-Wallis tests, $p < 0.001$ ***, $p < 0.01$ **.

A Kruskal-Wallis test shows a significant difference in J_1 between the different task difficulty adaptation approaches ($\chi^2_{(2)} = 19.99$, $p < 0.001$). A Tukey-Kramer Nemenyi test reveals that the ILC approach is significantly different from the two other approaches (Benchmark: $p < 0.001$, Random: $p = 0.004$).

Similarly, J_2 is significantly different between all groups ($\chi^2_{(2)} = 86.07$, $p < 0.001$), with the ILC approach resulting in higher J_2 compared to the other approaches (Tukey-Kramer Nemenyi, Benchmark: $p < 0.001$, Random: $p < 0.001$).

Lastly, a Kruskal-Wallis test reveals significant differences in f ($\chi^2_{(2)} = 124.73$, $p < 0.001$), with a Tukey-Kramer Nemenyi test showing that all approaches result in significantly different number of flashes with a p-value below 0.001.

Although J_1 is statistically significantly different with the ILC approach compared to the other approaches, the difference is small, and all approaches result in an average spelling accuracy higher than 80% (ILC: 84.5%, Benchmark: 89.0%, Random: 87.6%). However, f is significantly lower in the ILC group, resulting in a higher J_2 in that group. This suggests that the ILC approach leads to faster training without significantly affecting performance.

4.4 SUMMARY

This chapter describes the development of an ERP-based NFB training system using a P300 speller and ILC to dynamically adapt task difficulty.

The design choices and implementation of the P300 speller for NFB training are discussed, followed by an explanation of a phenomenological performance model. This model facilitates efficient testing of ILC controllers in simulation.

The design of the ILC controller is then presented, alongside results from a comprehensive simulation study that compares the ILC controller to a benchmark adaptation approach and random adaptation. The findings from this comparative study demonstrate that the ILC controller significantly accelerates the training process while maintaining high performance. The controller therefore addresses one of the key challenges in NFB training: enhancing training efficiency.

The ILC controller has two main hyperparameters: the penalty function $g(e_k)$ and the learning rate ϵ. These parameters were tuned using the simulation developed in this chapter, capturing a high level of performance in the P300 speller (as discussed in Section 4.3.1.3). The tuning is effective within the controlled simulation environment, yielding satisfactory performance.

However, performance in the P300 speller can vary significantly between individuals and even across sessions for the same individual. This variability is due to the dynamic nature of the "system",

which changes with factors like signal processing, experimental conditions, and individual cognitive abilities. Consequently, fixed hyperparameters may not consistently achieve optimal performance across all users or conditions. Tuning the controller for each individual and training session would likely yield the best results, achieving faster convergence to an optimal level of training difficulty. Advanced ILC algorithms like NOILC might offer enhanced performance if they incorporate an individualised model of the system, potentially achieving convergence within a few runs of the speller.

It is important to note, however, that true convergence is not fully attainable in this dynamic system because of the constant changes in user state, environmental conditions, and signal variability. Therefore, the controller must continuously adapt to maintain effective performance rather than reaching a stable, unchanging state.

Since the training in this thesis is limited to a single session, it is not feasible to tune the controller for each individual. Therefore, the controller with the hyperparameters described in Section 4.3.2.1 is used throughout the remainder of the thesis.

The effectiveness of the ILC controller in optimising training speed and maintaining user engagement is further validated through a clinical trial with human participants, which is detailed in the following chapter. This next step is crucial to confirm that the ILC can deliver the same benefits observed in simulation, thus solidifying its potential for real-world NFB applications.

Part III

EXPERIMENTAL WORK

5 ATTENTION TRAINING IN HEALTHY ADULTS

5.1 MOTIVATION

To test the efficacy of the ERP-based neurofeedback training system with adaptive task difficulty using ILC, as described in Chapter 4 of this thesis, a clinical trial aiming to train attention in healthy adults within a single neurofeedback training session is conducted. Based on the simulation results presented in Section 4.3.2, the hypothesis is that using ILC leads to a quicker training session than the benchmark algorithm by Arvaneh et al. [96], without compromising training efficacy. ILC is therefore compared to the benchmark algorithm, as well as to random task difficulty, to compare the efficacy between personalised and non-personalised training.

Arvaneh et al. [96] already demonstrated that receiving feedback in the P300-based neurofeedback training is crucial in enhancing attention, i.e. that there is no training effect without feedback, so the focus of this study is the adaptation of task difficulty.

By establishing a pre-registered clinical trial with a peer-reviewed protocol, this research also aims to address the concerns raised in Chapter 2 about the lack of well-controlled NFB studies. This approach ensures that the study meets high standards of scientific rigor and reliability.

This chapter outlines the study design in Section 5.2 and gives an overview of the different task difficulty adaptation approaches included in the study in Section 5.3. The analysis of study outcomes is described in Section 5.4, with the results presented in Section 5.5. These results are interpreted and discussed in Section 5.6 and a brief summary of this study is given in Section 5.7.

5.2 STUDY DESIGN

This is a single-blind, 3-arm randomised controlled trial. It was approved by the Maynooth University Ethics Committee (BSRESC-2022-2474456) and pre-registered on ClinicalTrials.gov (NCT05576649). The study protocol underwent peer-review prior to the commencement of the clinical trial [141].

51 healthy adults, with no self-reported history of neurological diseases or conditions, and normal or corrected-to-normal vision, were recruited. Of the 51 recruited participants, 4 did not complete the experiment, and 2 were excluded from the analysis due to a minor adjustment in the ILC controller. The sample size was selected based

on a previous study [96] that demonstrated significant results with a similar number of participants.

Participants were randomly assigned to one of 3 groups, which are described in Section 5.3. The groups consisted of 15 participants each, with mean ages of 27.2 ($\pm$10.3), 26.1 ($\pm$9.4), and 27.4 ($\pm$10.6) for the ILC, benchmark, and random difficulty groups, respectively. Each group included 6 female participants, and one participant in the random difficulty group preferred not to disclose their gender.

Each participant completed a single experimental session that lasted no more than 2 hours. The experiments were conducted in an electrically shielded, sound attenuated and dimly lit room on the Maynooth University campus.

All participants gave informed consent prior to the experiment and an allergy patch test for the electroconductive gel was performed.

The following subsections describe each task the participants completed in detail.

5.2.1 *Questionnaire*

Participants were instructed to complete two questionnaires. The first questionnaire, referred to as the fatigue-boredom questionnaire and inspired by Arvaeh et al.'s study [96], asked participants to rate the following four questions on a 10-point Likert scale:

Q1: How tired are you now?

Q2: How alert do you feel?

Q3: How bored do you feel?

Q4: Do your eyes feel tired?

This questionnaire was completed before and after the training to assess the training impact in these aspects.

The second questionnaire was the NASA task load index (TLX) [142], to be completed only at the end of the training. It was used to assess the subjective workload of the training. The NASA TLX contains the following questions, to be scored on a scale between 1 and 20:

Q1 – Mental Demand: How mentally demanding was the task?

Q2 – Physical Demand: How physically demanding was the task?

Q3 – Temporal Demand: How hurried or rushed was the task?

Q4 – Performance: How successful were you in accomplishing what you were asked to do?

Q5 – Effort: How hard did you have to work to accomplish your level of performance?

Q6 – Frustration: How insecure, discouraged, irritated, stressed and annoyed were you?

5.2.2 *Random Dot Motion Task*

Before and after the NFB training, participants completed the random dot motion (RDM) task to assess changes in cognitive abilities.

In the RDM task, moving dots on the screen are observed, where a fraction of the dots will move coherently in one direction, while the rest of the dots are moving in random directions. The participant then has to indicate the direction the coherent dots are moving in, often with a button press or joystick. The task usually involves discrete trials, where a fraction of dots are always moving coherently, with breaks in-between. However, in this study, a continuous version of the RDM task is used. In the continuous version of the task, the dots transition from incoherent motion, where all dots are moving in random direction, to coherent motion, where a fraction of dots moves in the same direction, in a continuous manner. This version of the task was used in [96] and is used here for better comparability. Figure 5.1 shows a schematic of the RDM task, where the coherence level is the percentage of dots that move in the same direction. In this specific task, the dots move either left or right during coherent motion phases.

Participants were asked to indicate the direction of the coherent motion by pressing the left or right arrow key once they are sure of the motion direction.

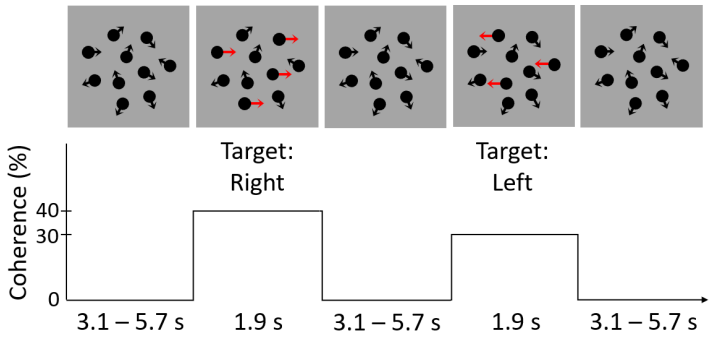


Figure 5.1: Schematic of Random Dot Motion (RDM) task. The task is to indicate the direction a fraction of dots are moving in. The dots switch between incoherent and coherent motion in certain intervals. For illustrative purposes, two target trials with coherence levels (i.e. percentage of coherently moving dots) of 40% and 30%, respectively, are shown.

An average of 118 dots, with a size of 6 by 6 pixels each, are displayed in a circular aperture of 5 ° at a viewing distance of 70 cm, resulting in a dot density of approximately 10.8%. Each dot is black

against a grey background. The dots are moving at a speed of 3.33 °/s. The incoherent motion lasts for either 3.1, 4.2, or 5.7 seconds, and the coherent motion always lasts for 1.9 seconds.

The code for this task was developed using PsychoPy [143]. The stimuli are presented on an LED monitor with a refresh rate of 60 Hz and a resolution of 1920 x 1080 pixels.

To familiarise participants with the task and reduce learning effects, they first completed 3 practice runs of 6 trials each, where verbal feedback of hits, misses and false alarms was given. In the first practice run, the coherence levels alternated between 80% and 60%, in the second run between 40% and 30%, and in the final run, between 25% and 20%.

Following these practice runs, participants completed the pre-training run of 40 trials without feedback. Here, the coherence level was either 25% or 19%, chosen randomly. The same was repeated after the training to evaluate any changes in attention.

5.2.3 *P300 Speller Task*

For the NFB training, participants copy-spelled 9 words, i.e. completed 9 runs, of the P300 speller described in Chapter 4. The runs are detailed in Table 5.1.

Table 5.1: Runs in the P300 speller.

Stage	Run	Word	Number of flashes	Feedback
Calibration	1	THE	12	No
	2	QUICK		No
	3	DOG		Yes
Training	4	BEAUTIFUL	10	Yes
	5 – 8	BEAUTIFUL	varying	Yes
Post-Training	9	DANCE	12	Yes

The first 2 runs, 'THE' and 'QUICK', are used to collect training data for the LDA classifier and xDAWN spatial filter. The third run is used to evaluate the BCI: If less than 2 letters are selected correctly, the classifier and spatial filter are retrained with the data from that run. In that case, participants are asked to copy-spell the word 'FOX', using 12 flashes per row and column, to reassess the speller. If less than 2 letters are correct, the training is stopped at that point. However, this did not happen in this study, with all participants achieving at least 2 correct letters in the third run. These first 3 runs make up the calibration stage.

The next stage is the training stage. First, participants spell the word 'BEAUTIFUL' with 10 flashes per row and column. They then repeat this run 4 times with a varying number of flashes according to their assigned group; how the number of flashes is adapted is explained in more detail in Section 5.3.

To compare post-training performance in the speller between groups, all participants copy-spelled the word 'DANCE' with 12 flashes per row and column, which is the post-training stage.

5.2.3.1 *EEG Acquisition*

During the P300 speller task, EEG signals for each participant are recorded using the Ant Neuro eego rt amplifier [144] with a 32-channel waveguard cap [145], with electrodes positioned according to the standard 10-10 system [23]. AFz and CPz serve as the ground and reference electrodes, respectively. During the online use of the P300 speller, all EEG signals are filtered between 1 and 20 Hz using a 4th-order Butterworth filter and then downsampled by a factor of 4. This EEG recording system and these settings for online use of the P300 speller are maintained across all studies described in this thesis.

The xDAWN spatial filter [137], previously explained in Section 4.2.2, is applied to reduce the 32 EEG channels to 3 xDAWN components, which maximise the difference between target and nontarget trials.

5.2.4 *Automation of Experimental Sessions*

All studies described in this thesis are automated using batch files to mitigate the possibility of human error, such as starting the wrong run or failing to save a recording properly. Automation also standardises the experiments, reducing inter-session variability. Additionally, automating the experimental sessions increases efficiency by eliminating time spent searching for files and inputting session information, which helps participants remain engaged and focused throughout the session. Lastly, the use of batch files streamlines log file creation and reduces errors, simplifying subsequent data analysis.

The batch files automatically launch all required programs, including the EEG acquisition software, the RDM task, the P300 speller with varying inputs for each run, and MATLAB for task difficulty adaptation. A button press is required between different tasks and runs of the P300 speller to allow participants to take breaks as needed.

To anticipate technical and non-technical issues, a secondary batch file was created for each study, enabling the experiment to be resumed from any point.

5.3 TASK DIFFICULTY ADAPTATION APPROACHES

5.3.1 *Iterative Learning Control Group*

In the ILC group, the number of flashes, i.e. task difficulty, in the training runs is adapted by the iterative learning controller. The development of the controller is described in Section 4.3.

As a reminder, the update law of the controller is

$$u_{k+1} = u_k(e_k + 0.5), \tag{5.1}$$

where u_k is the number of flashes in run k, and e_k is the percentage of incorrectly identified letters. This means that the number of flashes is increased if more than half the letters in a word are incorrect, and decreased otherwise.

5.3.2 *Benchmark Group*

For participants in the benchmark group, the number of flashes is adapted by the algorithm of Arvaneh et al. [96],

$$u_{k+1} = \frac{u_k + u_{k(66)}}{2}, \tag{5.2}$$

where u_k is the number of flashes in run k, and $u_{k(66)}$ is the lowest number of flashes that would have resulted in at least 66% spelling accuracy in the previous run. The algorithm therefore sets the number of flashes in the next run to the average of the number of flashes in the previous run that was actually used, and the number of flashes that would have resulted in 66% accuracy.

$u_{k(66)}$ is determined by assessing the spelling accuracy for each number of flashes from 1 to the number of flashes actually used. This is done by only considering the EEG signals associated with the first flash for classification of target trials, then the average EEG signals associated with the first two flashes for each row and column, and so on until the average EEG signals in response to all flashes are considered.

If the spelling accuracy in a run is lower than 66%, the number of flashes is increased by 1 in the next run:

$$u_{k+1} = u_k + 1. \tag{5.3}$$

5.3.3 *Random Difficulty Group*

As the name suggests, in the random difficulty group, the number of flashes is set to a random number between 1 and 10. This is done to include a wide range of difficulty levels, that do not go below the initial difficulty level (10 flashes per row and column), are unpredictable and non-personalised.

This group is included to compare the two personalised adaptation approaches to a non-personalised approach. To allow for a fair comparison, the task difficulty in the non-personalised approach should still vary and not remain the same throughout the training so there is no bias due to an easier task difficulty [146], and the task difficulty should be unpredictable, similar to the other approaches, and not increase systematically.

While the task difficulty for participants in this group does not adapt based on their performance, they do indeed undergo NFB training as the feedback they receive in the P300 speller is genuinely based on their brain signals.

5.4 DATA ANALYSIS

5.4.1 *Offline EEG Processing*

The recorded EEG signals are processed in a similar fashion to Arvaneh et al. [96]. This includes bandpass filtering the signals between 0.5 and 35 Hz with a Hamming-windowed sinc filter, and re-referencing to Fz. Only electrodes C3, Cz, C4, P3, Pz, and P4 are included in this analysis, following the protocol of Arvaneh et al. [96], due to the prominence of the P300 component in these centro-parietal regions of the brain [46].

EEG signals are segmented into baseline-corrected epochs of 150 ms to 550 ms post-stimulus, where the 150 ms period preceding the stimulus is used as the baseline. Any epochs with an amplitude greater than 75 μV, or with a voltage step of more than 150 μV within a 200 ms window, are excluded from analysis as these are likely caused by artifacts.

5.4.2 *Questionnaire*

The scores from the fatigue-boredom questionnaire are analysed using repeated measures analysis of variance (ANOVA) with stage (pre- and post-training), questions (fatigue, alertness, boredom, eye fatigue) and group (ILC, benchmark, random difficulty) as factors. This is done on ranked scores due to non-normality of the data.

This is followed by one-way ANOVA for normally distributed data and a Kruskal-Wallis test for non-normally distributed data, to investigate between-group differences, as well as paired t-tests and Wilcoxon signed-rank tests for within-group differences.

The scores for each question, as well as the total score, from the NASA TLX are analysed using one-way ANOVA and Kruskal-Wallis tests for between-group differences.

5.4.3 *Random Dot Motion Task*

Performance in the RDM task is assessed using three key metrics: response time (RT), accuracy, and a combined score of accuracy divided by RT. The inclusion of a combined score allows for a holistic measure of performance, considering that improvements in either accuracy or RT may not fully represent a participant's progress.

Repeated measures ANOVA with stage (pre- and post-training) and group (ILC, benchmark and random difficulty) are conducted to analyse changes in all three metrics. These tests are done on ranked data for accuracy to account for the non-normal distribution.

For all metrics, one-way ANOVA and Kruskal-Wallis tests are performed for between-group differences, and paired t-tests and Wilcoxon signed-rank tests for within-group differences.

5.4.4 *P300 Speller Task*

The first outcome from the P300 speller task that is analysed is the spelling accuracy, which is defined as the percentage of correctly identified letters in a run. For a fair comparison, spelling accuracy analysis includes only runs where feedback was provided and the number of flashes remained constant (runs 3, 4, and 9), ensuring uniformity in the conditions across all participants and groups. Between-group differences are analysed by Kruskal-Wallis tests due to non-normality of the data.

Since the hypothesis of the study is that ILC accelerates the training, another important outcome is training length. Here, training length is defined as the total number of flashes in the variable runs (i.e. runs 5 to 8), to ensure a fair comparison. This approach accounts for variations in breaks, as well as loading and setup times, which would affect the total time comparison. Between-group differences in training length are assessed using Kruskal-Wallis tests.

The EEG signals that were recorded during the P300 speller task are analysed in terms of P300 amplitude, latency, total power and power in the alpha band. The P300 amplitude is defined as the voltage difference between the minimum and maximum in the target epoch average (peak-to-peak voltage). This is chosen to account for epoch drift and to include the N200, which is closely related to the P300 [26]. The P300 latency is defined as the time between stimulus onset and the positive peak (maximum voltage) in the target epoch average.

The total power is calculated by averaging the squared samples in the epoch average. It is included since the P300 amplitude alone may not accurately reflect increases in P300 strength; such increases could be a general rise in amplitude rather than a single distinct peak. The total power is calculated for both target and nontarget trials, separately.

The alpha power is defined as the power between 7 and 12 Hz, calculated in the same way as the total power. The calculation is performed exclusively for the 150 ms period following stimulus onset in nontarget trials, provided these did not immediately succeed target trials. It is included to evaluate participants' attentional state, since alpha power desynchronisation is known to be correlated with selective attention [147].

All EEG metrics are analysed by first calculating the average of the metric for each stage (calibration, training and post-training), and then calculating the ratio of training to calibration, and post-training to calibration, for each metric. If a ratio is greater than unity, it means that the metric increased compared to calibration (baseline), and if it is less than unity, the metric decreased.

Repeated measures ANOVA with stage and group indices as factors (on ranked data if necessary) is applied to all ratios. For total power ratios, repeated measures ANOVA is conducted for each stage individually, with trial (target and nontarget) and group as factors. For all EEG metrics, one-way ANOVA and Kruskal-Wallis tests are used for between-group differences, and paired t-tests and Wilcoxon signed-rank tests for within-group differences.

5.4.5 *Correlation between P300 Speller Task and Random Dot Motion Task*

To investigate how the training (i.e. the P300 speller task) influences attention improvement, as measured by the RDM task, correlation analysis is conducted between performance and EEG changes in the speller task, and performance in the RDM task. The hypothesis underlying the P300-based training is that, by increasing the task difficulty, participants are encouraged to improve their focus to maintain performance. It is therefore expected that participants with good speller performance have rather small improvements in the RDM task, since they are not sufficiently challenged to improve their attention. Similarly, participants with large (positive) EEG changes should have larger increases in the RDM task.

For the P300 speller, minimum and average spelling accuracy in feedback runs are included in the correlation analysis, as well as all EEG metrics. The minimum spelling accuracy is also included, since the average might not capture the level of challenge accurately, due to easier runs at the beginning of the training.

For the RDM task, the ratio of post-training to pre-training RT and accuracy are used.

Pearson's correlation coefficient is used for normally distributed data, and Spearman's correlation coefficient otherwise.

5.4.6 *Post-Hoc Sensitivity Analysis*

Despite randomised group allocation of participants, a baseline imbalance in boredom and eye fatigue is observed according to the fatigue-boredom questionnaire, as presented in Section 5.5.1. This is not taken into account in the primary statistical analysis described in this section, which is why a sensitivity analysis, where the baseline for boredom and eye fatigue are factored in, is conducted.

Due to imbalanced distribution of scores, as shown in Table 5.2 and Table 5.3, the baseline scores cannot be treated like continuous variables and therefore covariates. That is why the scores are aggregated and then treated as categorical factors, with low, medium and high levels. How the scores are aggregated is detailed in Table 5.2 and Table 5.3.

To investigate between-group differences of all outcomes, a two-way ANOVA test for normally distributed data and an aligned rank transformation (ART) ANOVA for non-normal data are conducted, both with baseline level and group as factors.

Table 5.2: Distribution and aggregation of baseline boredom scores.

Level	Low			Med			High				
Likert Score	0	1	2	3	4	5	6	7	8	9	10
ILC	5	3	1	3	0	3	0	0	0	0	0
Benchmark	4	7	1	1	0	1	1	0	0	0	0
Random Difficulty	0	2	5	1	5	1	1	0	0	0	0

Table 5.3: Distribution and aggregation of baseline eye fatigue scores.

Level	Low			Med			High				
Likert Score	0	1	2	3	4	5	6	7	8	9	10
ILC	3	4	2	1	2	2	0	1	0	0	0
Benchmark	4	2	2	1	1	1	4	0	0	0	0
Random Difficulty	0	0	1	3	0	3	4	3	1	0	0

5.5 RESULTS

5.5.1 *Questionnaire*

Figure 5.2 shows the scores of the fatigue-boredom questionnaire.

The repeated measures ANOVA on the ranked scores of the fatigue-boredom questionnaire reveals a significant main effect of group ($F_{(2,42)} = 7.94$, $p < 0.001$), stage ($F_{(1,44)} = 28.03$, $p < 0.001$), and question ($F_{(3,132)} = 28.44$, $p < 0.001$). Significant interactions between stage and question ($F_{(3,132)} = 3.31$, $p = 0.020$), and question and group ($F_{(6,264)} = 2.16$, $p = 0.046$) are also revealed.

A Kruskal-Wallis test shows that there is a significant difference in pre-training boredom scores ($\chi^2_{(2)} = 7.30$, $p = 0.026$) specifically between the benchmark and random difficulty groups (Tukey-Kramer Nemenyi: $p = 0.03$). Similarly, a significant difference in pre-training eye fatigue scores is observed between groups ($\chi^2_{(2)} = 11.93$, $p = 0.003$), with a Tukey-Kramer Nemenyi test showing that the random difficulty group is different from the others (ILC: $p = 0.004$, benchmark: $p = 0.018$).

A significant increase in tiredness is observed in all groups (ILC: $t_{(14)} = -2.98$, $p = 0.010$; benchmark: $t_{(14)} = -3.06$, $p = 0.009$; random difficulty: $W = 7$, $p = 0.021$). Additionally, participants in the ILC and benchmark groups experienced a significant increase in eye fatigue (ILC: $t_{(14)} = -4.58$, $p < 0.001$; benchmark: $W = 0$, $p = 0.001$). Participants in the ILC group reported significantly higher levels of boredom post-training ($W = 2$, $p = 0.029$).

Table 5.4 shows the scores of the NASA TLX questionnaire.

A Kruskal-Wallis test reveals a significant difference in physical demand ($\chi^2_{(2)} = 8.35$, $p = 0.015$) between the benchmark and random difficulty groups (Tukey-Kramer Nemenyi: $p = 0.024$). There are no significant differences in the other questions and the total score (i.e. the sum of all question scores).

Table 5.4: Mean NASA Task Load Index (TLX) scores for all 3 groups. Total score is the sum of all questions. Standard deviation is shown in brackets.

Question	ILC	Benchmark	Random Difficulty
Mental demand	14 (2.95)	10.50 (5.55)	13.80 (3.82)
Physical demand	3.27 (2.96)	3.53 (4.93)	8.00 (5.89)
Temporal demand	10.50 (4.47)	7.47 (4.21)	9.93 (5.11)
Performance	8.87 (2.61)	7.20 (4.21)	6.80 (4.33)
Effort	13.9 (2.77)	13.00 (3.80)	12.70 (3.51)
Frustration	7.00 (4.57)	5.87 (4.24)	7.60 (5.67)
Total score	57.5 (11.10)	47.6 (13.50)	58.90 (17.00)

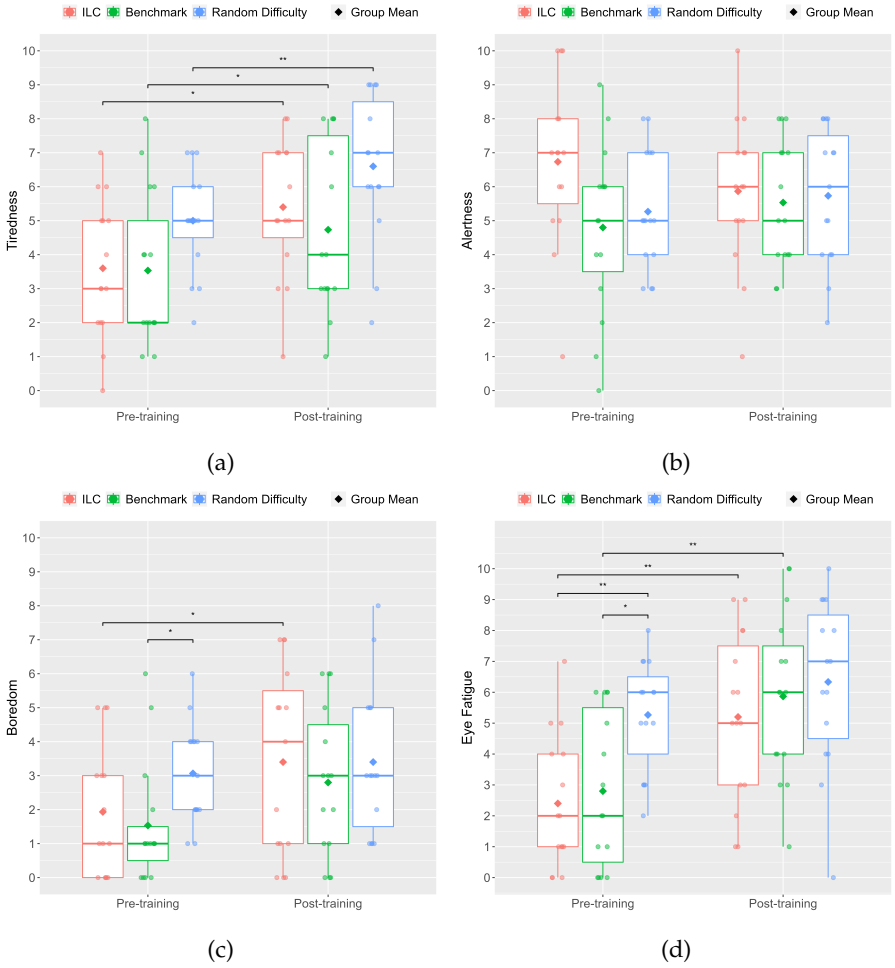


Figure 5.2: Scores of the fatigue-boredom questionnaire. (a) Q1 - Fatigue. (b) Q2 - Alertness. (c) Q3 - Boredom. (d) Q4 - Eye Fatigue. Statistical analysis by paired t-test and Wilcoxon signed-rank test, $0.001 \leqslant p < 0.01$ **, $p \leqslant 0.05$ *.

5.5.2 *Random Dot Motion Task*

Figure 5.3 illustrates the performance in the RDM task for all groups and the three performance metrics analysed in this study.

Although an average decrease in RT is observed across all groups, it is not statistically significant according to a repeated measures ANOVA. There are also no significant between-group differences in either stage.

In contrast, a significant main effect of stage is revealed by a repeated measures ANOVA on ranked accuracy ($F_{(1,44)} = 10.91$, $p = 0.001$). One-way ANOVA and Kruskal-Wallis tests confirm that there are no significant between-group differences. Paired t-tests and Wilcoxon signed-rank tests show that all groups experienced a significant increase in accuracy (ILC: $t_{(14)} = -3.83$, $p = 0.002$; benchmark: $t_{(14)} = -2.54$, $p = 0.024$; random difficulty: $W = 15$, $p = 0.036$).

Despite this significant increase in accuracy in all groups, only the ILC group achieved a significant increase in score, defined as accuracy divided by RT ($t_{(14)} = -2.79$, $p = 0.015$).

5.5.3 *P300 Speller Task*

5.5.3.1 *Spelling Accuracy*

All participants demonstrated high proficiency with the P300 speller, as evidenced by the high average spelling accuracy across common runs that provided feedback, shown in Table 5.5. There are no statistically significant differences between groups according to a Kruskal-Wallis test.

Table 5.5: Mean spelling accuracy (%) in the P300 speller runs that provided feedback and that were the same for all groups. Standard deviation is shown in brackets.

Group	Run 3	Run 4	Run 9
ILC	97.80 (8.61)	100.00 (0.00)	100.00 (0.00)
Benchmark	97.80 (8.61)	98.50 (3.91)	98.70 (5.16)
Random Difficulty	97.80 (8.61)	94.10 (10.20)	96.00 (8.28)

5.5.3.2 *Training Length*

Figure 5.4 shows the training length in terms of total number of flashes in the variable runs for all groups.

A Kruskal-Wallis test reveals a significant between-group difference ($\chi^2_{(2)} = 22.30$, $p < 0.001$). This difference lies between the ILC group and the others according to a Tukey-Kramer Nemenyi test

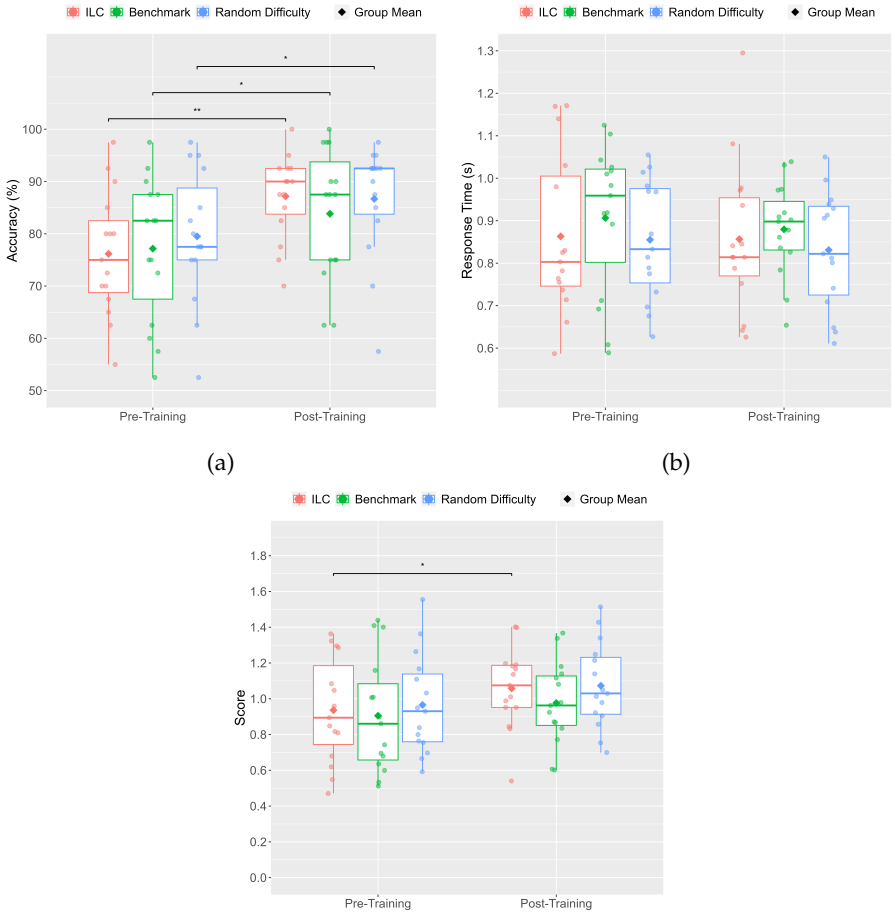


Figure 5.3: Performance in the Random Dot Motion (RDM) task. (a) Accuracy (%). (b) Response time (RT, s). (c) Score, calculated as accuracy divided by response time. Statistical analysis by paired t-test and Wilcoxon signed-rank test, $0.001 \leqslant p < 0.01$ **, $p \leqslant 0.05$ *.

(benchmark: $p = 0.007$, random difficulty: $p < 0.001$). The difference in training length between the benchmark and random difficulty groups is not significant ($p = 0.255$).

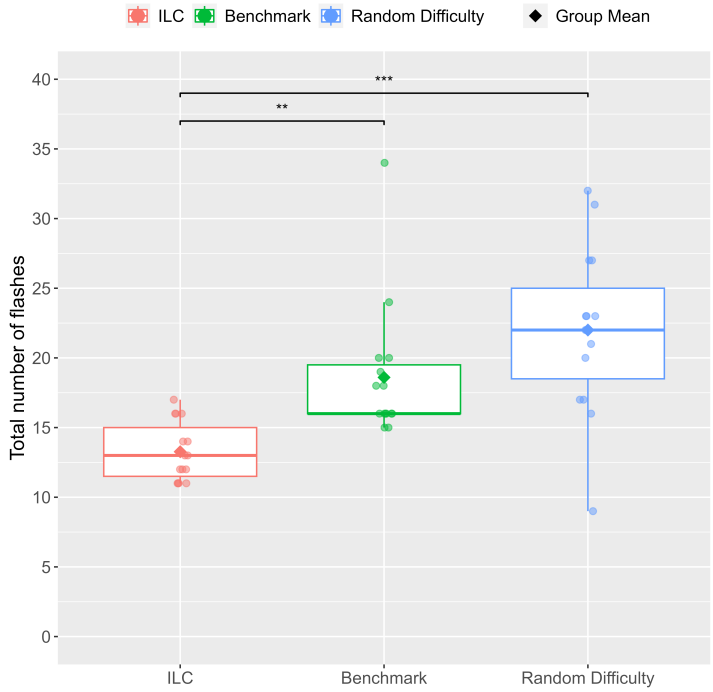


Figure 5.4: Training length in terms of total number of flashes in runs 5 to 8. Statistical analysis by Kruskal-Wallis tests, $p < 0.001$ ***, $p < 0.01$ **.

5.5.3.3 *P300 Event-Related Potential*

A repeated measures ANOVA on P300 amplitude ratios, as pictured in Figure 5.5, reveals a significant main effect of stage ($F_{(1,44)} = 20.13$, $p < 0.001$), and group ($F_{(2,42)} = 4.74$, $p = 0.011$). A one-way ANOVA shows that there is a significant difference in the training-to-calibration amplitude ratio between the ILC and random difficulty groups ($F_{(2,42)} = 3.25$, $p = 0.049$, Tukey test: $p = 0.044$).

There are significant differences between training-to-calibration and post-training-to-calibration amplitude ratios in the ILC group ($t_{(14)} = 3.4$, $p = 0.002$) and the random difficulty group ($t_{(14)} = 4.42$, $p < 0.001$).

There are no significant between- or within-group differences in P300 latency, which is shown in Figure 5.6, according to a repeated measures ANOVA test on ranked data.

5.5.3.4 *Total Power*

As can be seen in Figure 5.7, there is an average increase in post-training total power of 4% in the ILC group and 18% in the bench-

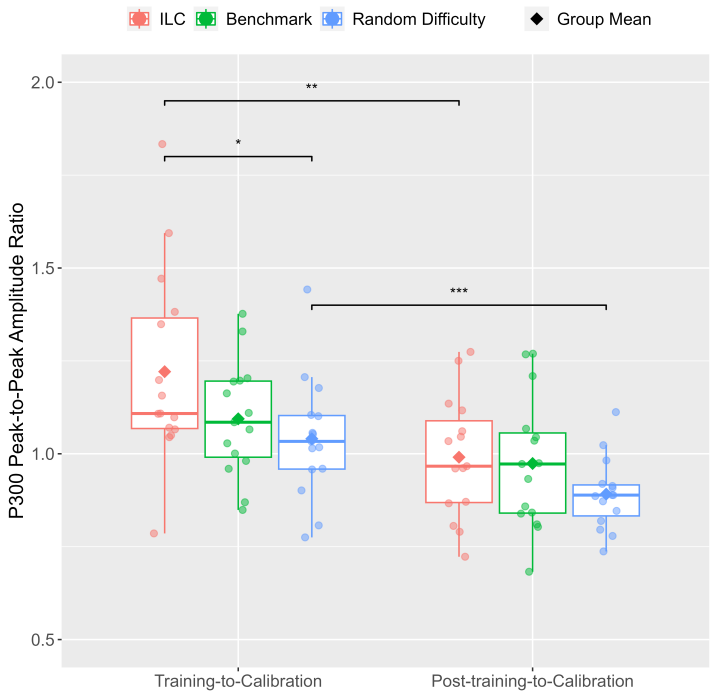


Figure 5.5: P300 amplitude ratios. Statistical analysis by one-way ANOVA/ Kruskal-Wallis tests and paired t-tests/ Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.

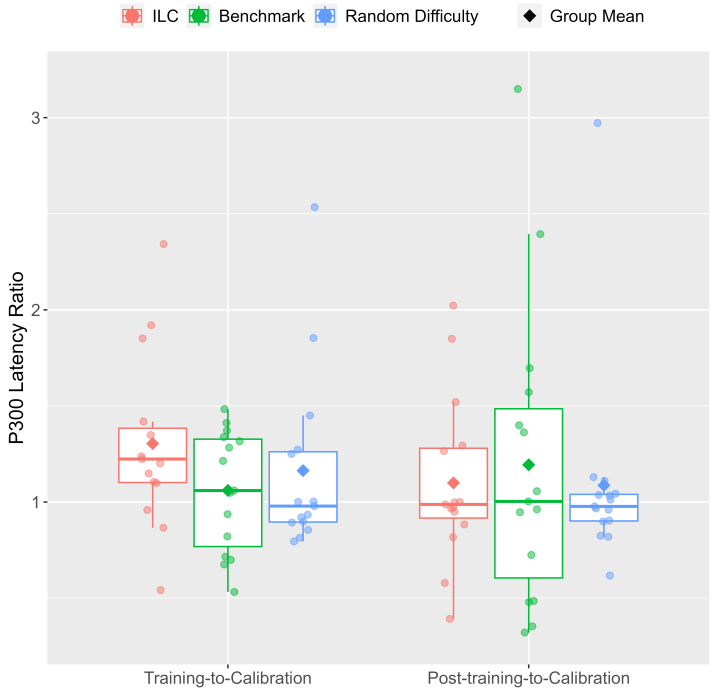


Figure 5.6: P300 latency ratios. Statistical analysis by one-way ANOVA/ Kruskal-Wallis tests and paired t-tests/ Wilcoxon signed-rank tests.

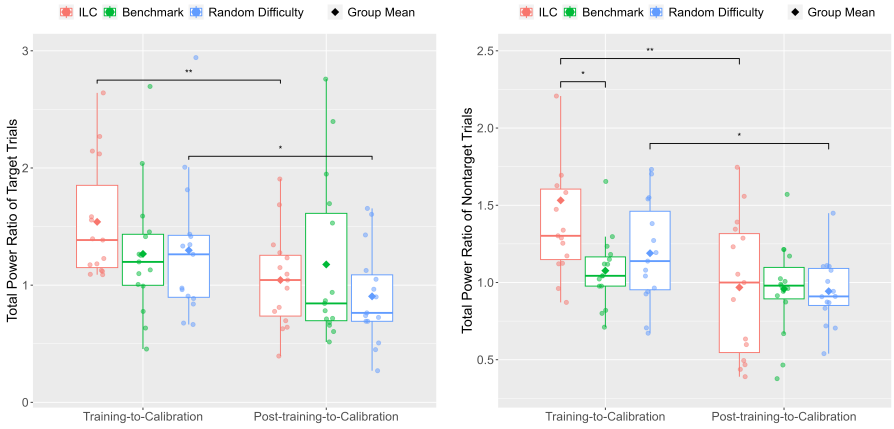
mark group in target trials, and a decrease of 3% and 4% in the ILC and benchmark groups, respectively, in nontarget trials. In the random difficulty group, post-training total power decreased both in target and nontarget trials by 10% and 6%, respectively.

According to a repeated measures ANOVA on the ranked total power ratios, there is a significant main effect of stage ($F_{(1,44)} = 26.19$, $p < 0.001$), and group ($F_{(2,42)} = 3.25$, $p = 0.041$).

Conducting a repeated measures ANOVA on the ranked training-to-calibration ratios only revealed a significant main effect of group ($F_{(2,42)} = 4.79$, $p = 0.011$). There are no significant between-group or between-trial differences in the post-training-to-calibration ratios.

A significant difference in nontarget training-to-calibration ratios is found between the ILC and benchmark groups ($\chi^2_{(2)} = 6.64$, $p = 0.036$, Tukey-Kramer Nemenyi: $p = 0.030$).

Significant within-group differences between both target and nontarget training-to-calibration and post-training-to-calibration ratios is found in the ILC group (target: $W = 108$, $p = 0.004$, nontarget: $W = 107$, $p = 0.005$) and random difficulty group (target: $W = 101$, $p = 0.018$, nontarget: $t_{(14)} = 2.91$, $p = 0.011$).



(a) (b)

Figure 5.7: Total power ratios. (a) Target trials. (b) Nontarget trials. Statistical analysis by one-way ANOVA/ Kruskal-Wallis tests and paired t-tests/ Wilcoxon signed-rank tests, $0.001 \leqslant p < 0.01$ **, $p \leqslant 0.05$ *.

5.5.3.5 *Alpha Power*

Figure 5.8 shows the training-to-calibration and post-training-to-calibration alpha power ratios.

A repeated measures ANOVA test on ranked data shows a significant main effect of stage ($F_{(1,44)} = 9.64$, $p = 0.003$), and a significant interaction of stage and group ($F_{(2,42)} = 3.70$, $p = 0.029$).

There is a significant between-group difference in training-to-calibration ratios ($F_{(2,42)} = 5.93$, $p = 0.005$), specifically between the ILC group and the others according to a Tukey test (benchmark: $p = 0.006$, random difficulty: $p = 0.034$).

In the ILC group, the post-training-to-calibration ratios are significantly lower than the training-to-calibration ratios ($t_{(14)} = 5.46$, $p < 0.001$).

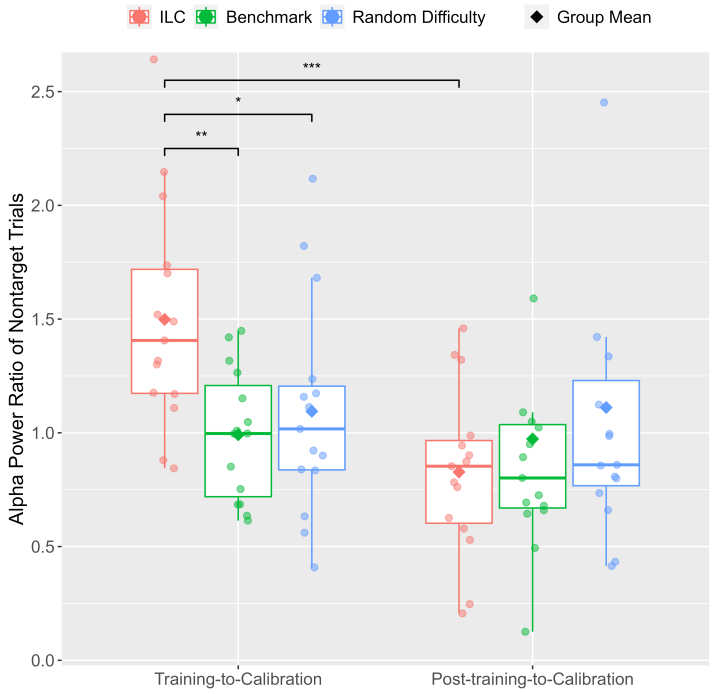


Figure 5.8: Power in the alpha band (7 to 12 Hz) in nontarget trials. Statistical analysis by one-way ANOVA/ Kruskal-Wallis tests and paired t-tests/ Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.

5.5.4 *Correlation between P300 Speller Task and Random Dot Motion Task*

No correlation is found between spelling accuracy in the P300 speller task and RDM task performance within the ILC and random difficulty groups. However, a significant negative correlation emerges between mean spelling accuracy and RDM accuracy ratio ($\rho_{(13)} = -0.64$, $p = 0.010$), as well as between minimum spelling accuracy and RDM accuracy ratio ($r_{(13)} = -0.58$, $p = 0.023$), in the benchmark group.

Furthermore, a significant negative correlation between training-to-calibration latency ratio and RDM RT ratio is found both in the benchmark group ($r_{(13)} = -0.53$, $p = 0.045$) and the ILC group ($r_{(13)} = -0.65$, $p = 0.008$).

Additionally, a significant positive correlation between training-to-calibration alpha ratio and RDM RT ratio is observed in the ILC group ($r_{(13)} = 0.56$, $p = 0.032$).

5.5.5 *Post-Hoc Sensitivity Analysis*

The post-hoc sensitivity analysis, which considers baseline boredom and eye fatigue levels, identifies three outcomes with results differing from the primary analysis. This indicates that these outcomes are influenced by baseline levels of boredom and/or eye fatigue.

The first outcome is the NASA TLX scores, where significant differences are found in mental demand between the benchmark and random difficulty groups when eye fatigue is taken into account (two-way ANOVA: $F_{(2,42)} = 4.08$, $p = 0.025$; Tukey: $p = 0.042$). In contrast to the primary analysis, no significant between-group differences are found in physical demand when eye fatigue is included as a factor in ART ANOVA.

The second outcome that is different is the P300 spelling accuracy. In run 4, significant differences in spelling accuracy are revealed when baseline boredom levels are considered (ART ANOVA: $F_{(2,42)} = 4.81$, $p = 0.014$). Contrast tests showed that the difference between the ILC and benchmark groups is tending to significance ($p = 0.059$), while the difference between the ILC group and random difficulty group is significant ($p = 0.023$). The same goes for run 9, where a significant between-group difference is found ($F_{(2,42)} = 5.60$, $p = 0.008$), again between the ILC group and the others (benchmark: $p = 0.048$, random difficulty: $p = 0.011$).

The last outcome that is affected by baseline boredom and eye fatigue levels is the P300 amplitude. A two-way ANOVA test indicates that there are no significant between-group differences in training-to-calibration amplitude ratios when eye fatigue baseline levels are taken into account.

All other study outcomes seem to be robust against the observed baseline imbalances.

5.6 DISCUSSION

In this study, three different methods to adapt the task difficulty in P300-based NFB training are compared. A summary of the main findings, along with the statistical tests used, results, and their implications, can be found in Appendix B.

All groups show significant improvements in the RDM task. This finding indicates that P300 speller-based attention training is effective across various task difficulty adaptation methods. However, it is found that the ILC controller developed in this thesis (Section 4.3) ac-

celerates the training significantly compared to the other approaches, without compromising the training efficacy.

Participant feedback through questionnaires shows that the training was both tiring and mentally demanding, which is to be expected as the training is meant to be challenging. Additionally, participants reported an increase in eye fatigue and boredom. Eye strain might be mitigated in future studies by encouraging participants to take frequent breaks away from the screen. While frequent breaks between each run were offered in this study, most participants preferred to continue the training without breaks. Although boredom scores slightly increased after the training, they are still low.

Interestingly, some participants rated the training as being physically demanding despite no physical aspect being present in the training. It is believed that for these participants, the high mental demand of the training and the increase in tiredness was perceived as physical demand.

Overall, the NASA TLX scores are mostly neutral, indicating an acceptable workload of the training and that the training is challenging but not overly frustrating.

The peak-to-peak P300 amplitude increased for all groups during the training, with it being the highest in the ILC group. The ILC group is the group with the lowest number of flashes, which supports the hypothesis that decreasing the number of flashes drives the attention improvement and therefore P300 amplitude increase.

Further positive EEG changes are observed in the ILC and benchmark groups, with a power increase in target trials and power decrease in nontarget trials, as well as a decrease in alpha power after the training. This indicates that participants in these groups were more focused on target trials and less distracted by nontarget trials. The decreased alpha power also indicates that they were in a more focused state post-training.

However, these positive changes are not observed in the random difficulty group, where total power decreased in both target and nontarget trials, and alpha power increased post-training. This means that non-personalised training might not be as effective as personalised training.

A significant negative correlation between P300 spelling accuracy and accuracy ratios in the RDM task is found, indicating that higher spelling accuracy during training corresponded to lesser improvement in RDM task accuracy. This makes sense, since it might mean that the training was not sufficiently challenging to improve attention. Similarly, a positive correlation between alpha power and RT ratios in the RDM task is observed, which means participants with lower alpha power, i.e. participants in a more focused state, improved their RT post-training. An unexpected negative correlation is found between

P300 latency and RT ratios. A positive correlation would be expected as lower latency is associated with faster reaction times [46].

The post-hoc sensitivity analysis due to baseline imbalances shows that perceived mental and physical demand, P300 spelling accuracy, as well as P300 amplitude, are all affected by baseline boredom and/or eye fatigue levels. However, the main outcomes of this study, i.e. RDM performance and training length, are unaffected by the baseline imbalances.

The study shows that the ILC and benchmark adaptation approaches are similar in terms of efficacy, however, the ILC algorithm is faster and more computationally efficient than the benchmark algorithm, since only the spelling accuracy for the number of flashes that was actually used needs to be calculated for the ILC algorithm.

While the ILC controller used in this study significantly accelerates training without compromising efficacy, further enhancements may be possible. For example, fine-tuning ILC parameters specifically for each individual could allow for more precise adaptation to each user's unique cognitive responses and engagement levels, potentially improving convergence speed and overall efficiency. However, such an approach would require sufficient data from each individual prior to tuning the controller, which is not feasible within a single training session.

In contrast, commonly used early stopping methods, as discussed in Section 2.5, could assist the ILC in converging faster by filtering out unnecessary repetitions, especially during early stages of training. However, as these approaches primarily adapt to current performance without challenging the user's cognitive engagement, they may not, in isolation, produce the same attention-enhancing effects as the ILC, which intentionally pushes users by progressively adjusting task difficulty. Future research could compare the ILC adaptation with early stopping methods to determine if they can achieve similar training effects or if the ILC adaption approach yields more sustained attention improvements.

5.7 SUMMARY

This chapter describes an attention training study with healthy adults to compare different, personalised and non-personalised, task difficulty adaptation approaches.

The results of the study are promising with statistically significant improvements in the RDM task and positive EEG changes with the personalised approaches, despite only a single training session.

The ILC controller developed in this thesis is significantly faster than the other approaches, while being computationally more efficient than the benchmark algorithm.

6 OPTIMISING THE EXPERIMENTAL PROTOCOL

6.1 MOTIVATION

The results from the study described in Chapter 5 show that the proposed P300-based neurofeedback training can effectively train attention in healthy adults. The results also demonstrate that personalising the task difficulty in the training session leads to better outcomes, and that using ILC to adapt the task difficulty accelerates the training.

However, the experimental procedure was not optimal. While using all available 32 EEG channels means that the signal quality was very good, setting up the EEG cap took a long time. For some participants, setting up the EEG cap even took longer than the training itself. Other disadvantages include increased cost of the EEG system, higher processing power requirements, greater discomfort for participants and the need for a large amount of electroconductive gel. The aim was therefore to reduce the number of electrodes used for the training.

Another suboptimal aspect of the training in the previous study was that there were only 4 runs with adaptive task difficulty in the training. This meant that the optimal task difficulty was not achieved for some participants. To mitigate this issue, it was desired to increase the number of adaptive runs, so that there are more opportunities for finding a participant's optimal task difficulty.

This chapter outlines the development and testing of a refined training protocol featuring fewer electrodes and more adaptive runs, thus making the NFB system a more practical intervention. Initially, the most crucial electrodes for target trial classification were identified through offline data analysis, detailed in Section 6.2. Subsequent steps included a comparative study of electrode sets (Section 6.3) and an evaluation of the new protocol's effectiveness and participant satisfaction (Section 6.4).

6.2 ELECTRODE SELECTION: OFFLINE DATA ANALYSIS

6.2.1 *Determining Electrode Sets*

Many different methods have been proposed for electrode selection in P300 speller applications. Often, these methods use an iterative search, where electrodes are either removed from (sequential reverse selection) or added to (sequential forward selection) the set, with classification results assessed after each change. However, the iterative na-

ture of these methods makes them relatively slow and computationally intensive [148]. Other approaches determine relevant electrodes based on factors such as correlation with the target signal, neuroscientific insights into the source of the signal of interest, or the SNR of each electrode [149].

In this study, xDAWN spatial filtering is chosen to guide electrode selection for two main reasons. First, xDAWN has demonstrated effectiveness in optimising P300 signals by weighting channels based on their contribution to signal detection, thus providing a reliable indication of SNR for each electrode [137]. Second, leveraging xDAWN weights for electrode ranking is both efficient and practical, as this information is already available in the dataset from Chapter 5, which serves as the basis for electrode selection in the current study.

In the previous study, described in Chapter 5, the xDAWN spatial filter was used to reduce the 32 EEG electrodes to 3 xDAWN components. At the start of each session, this filter is customised for the participant, assigning unique weights to each electrode. Figure 6.1 shows an example of these weights. For each of the 3 xDAWN components, there is a weight for each electrode, resulting in 96 weights.

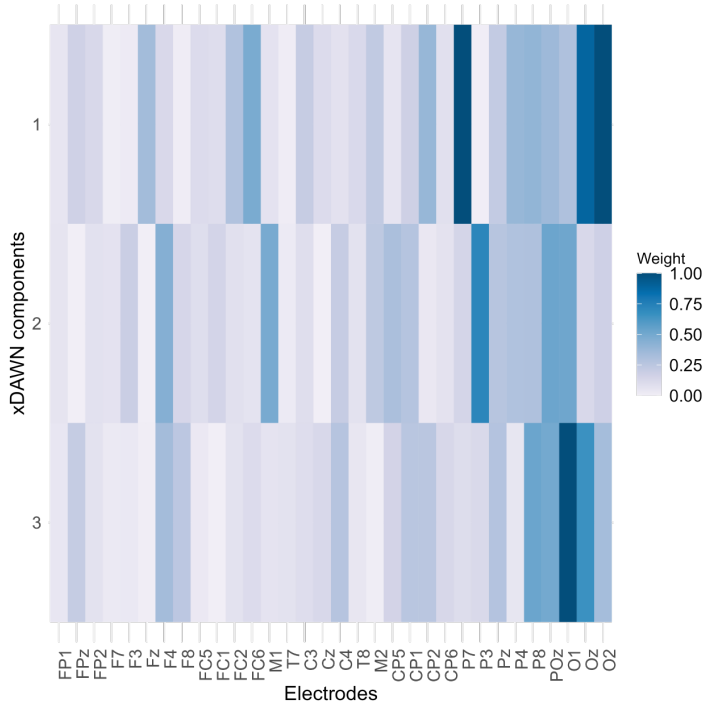


Figure 6.1: Heatmap showing scaled xDAWN weights for each electrode.

For this analysis, 5 electrode sets, comprising 32, 16, 8, 6, and 4 electrodes, respectively, are selected to align with the number of electrodes commonly used in EEG research. To identify which electrodes to include in each set, the weights assigned by the xDAWN spatial filter are leveraged. Specifically, the 32 electrodes are ranked according to the frequency with which they appear among the top

16 highest-weighted electrodes across the 47 participants from the study of Chapter 5. Consequently, the 4 electrodes with the highest frequency of appearance are allocated to the 4-electrode set, the top 6 to the 6-electrode set, and so forth, ensuring that each set is composed of electrodes most indicative of significant signal contributions as determined by their xDAWN weights. To maintain symmetry in the electrode sets, adjustments are made based on the electrode positions. For example, P3, P4, O1, and O2 are selected instead of the top-ranked P3, P4, O1, and Oz for the 4-electrode set.

In addition to the sets determined using the ranking, a 4-electrode set inspired by literature is included for benchmarking, based on the findings of Speier et al. [150]. They report comparable performance in a P300 speller between 4 electrodes (PO7, PO8, Pz and POz) and 32 electrodes. As PO7 and PO8 were not included in the original setup, P7, P8, Pz, and POz are selected as the closest alternatives.

To simplify reference to each set, a naming convention is adopted based on the number of electrodes in each set: Set 32 (32 electrodes), Set 16 (16 electrodes), Set 8 (8 electrodes), Set 6 (6 electrodes), Set 4 (4 electrodes based on ranking), and Set 4L (4-electrode set based on the literature findings of Speier et al. [150]). Figure 6.2 shows which electrodes are included in the six sets that are analysed.

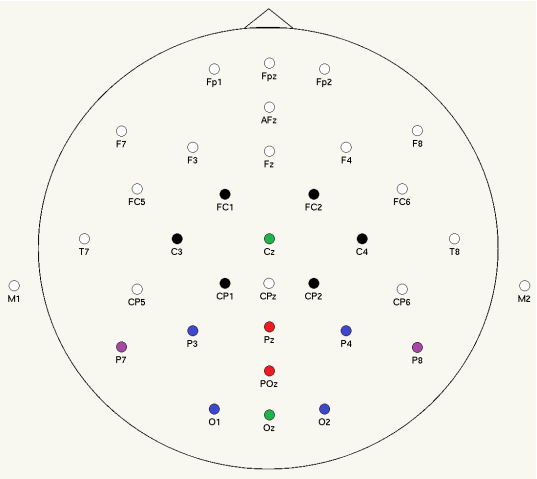


Figure 6.2: Electrode sets used in the analysis. AFz and CPz are used as ground and reference, respectively. Set 32: all electrodes shown on the image. Set 16: electrodes of any colour. Set 8: all electrodes coloured in red, green and blue. Set 6: all electrodes coloured in red and blue. Set 4 (based on ranking): all electrodes coloured in blue. Set 4L (based on [150]): all electrodes coloured in purple and red.

6.2.2 *Evaluating Electrode Sets*

The 6 electrode sets described in Section 6.2.1 are analysed by replaying the experiment outlined in Chapter 5 for all participants across all sets, using the recorded EEG signals. Specifically, the spelling accuracy achievable with each electrode set and the xDAWN spatial filter impact on their performance are investigated. Consequently, a total of 12 configurations are examined, encompassing the 6 sets both with and without the application of the xDAWN spatial filter.

To achieve this, the cumulative spelling accuracy is calculated for each participant across a range of 1 to 12 flashes, aligning with the maximum number of flashes used in Chapter 5. This is done by only considering the EEG signals in response to the first flash per row and column for classification, then the EEG signals in response to the first two flashes per row and column, and so on until the EEG signals in response to all 12 flashes per row and column are used for classification. This results in a curve of spelling accuracy against number of flashes for every configuration, as can be seen in Figure 6.3 for the grand mean. Given the similarity of curves across all configurations, namely, the logarithmic relationship between spelling accuracy and number of flashes, the analysis is simplified by computing the mean spelling accuracy across all flash counts, yielding a single value per configuration.

Subsequently, the differences in mean spelling accuracy between configurations are assessed employing a Friedman test, complemented by paired t-tests and Wilcoxon signed-rank tests, with Bonferroni adjustments, for pairwise comparisons.

The resolution of spelling accuracy in the P300 speller is very small, e.g. a 1-letter difference in the word 'BEAUTIFUL' corresponds to an 11% difference in spelling accuracy. This indicates that a statistically significant outcome in the aforementioned pairwise comparisons may not translate to practical significance, as the observed difference between any two configurations could be minimal, potentially less than the difference of a single letter, despite statistical significance. To address this, pairwise equivalence testing is conducted using the two one-sided tests (TOST) procedure [151].

The null hypothesis for a difference test posits that there is no difference between two or more populations, implying that a significant outcome indicates a non-zero difference. Conversely, the null hypothesis for an equivalence test suggests that any difference lies outside predefined bounds, which represent the minimum meaningful difference for the study context. Thus, a significant result from an equivalence test indicates that the observed difference is within these specified bounds [151].

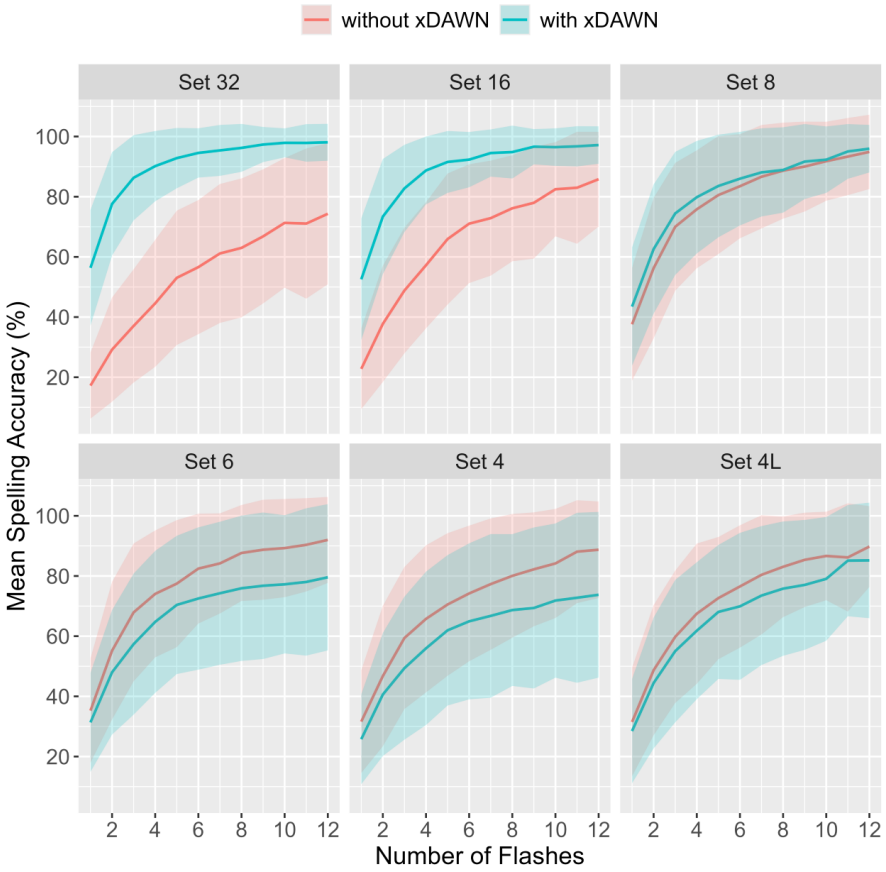


Figure 6.3: Grand mean spelling accuracy achieved with different electrode sets, with and without using the xDAWN spatial filter. Standard deviation illustrated by shading.

Testing data for both differences and equivalences can therefore reveal whether there are any differences (difference test), and whether those differences are practically meaningful (equivalence test).

6.2.3 *Results*

6.2.3.1 *Difference Tests*

Given that only the spelling accuracies for Sets 32 and 16 without the xDAWN spatial filter, and Set 4L with the xDAWN filter, conformed to normal distribution, a Friedman test is conducted, which reveals a significant effect of configuration on mean spelling accuracy ($\chi^2_{(11)} = 352.37$, $p < 0.001$).

The mean spelling accuracy of each configuration is plotted in Figure 6.4, illustrating comparisons within sets regarding the xDAWN spatial filter (Figure 6.4a, and between sets with the filter (Figure 6.4b) and without (Figure 6.4c). Paired t-tests are conducted for sets with normally distributed data, as mentioned, and Wilcoxon signed-rank tests for the remaining sets.

Figures 6.4a–6.4c demonstrate that Sets 32, 16, and 8 perform better with the xDAWN spatial filter, whereas Sets 6, 4, and 4L show improved accuracy without it. Consequently, the optimal configurations for each set are plotted to facilitate direct between-set comparisons, as shown in Figure 6.4d.

6.2.3.2 *Equivalence Tests*

To assess the practical equivalence between configurations, pairwise equivalence tests are performed. The t-TOST procedure is applied for normally distributed configurations (Sets 32 and 16 without the xDAWN spatial filter, and Set 4L with the filter) and the Wilcoxon TOST method for the others. Upper and lower equivalence bounds are established at 11%, reflecting the highest resolution of spelling accuracy observed in Chapter 5. This threshold corresponds to a 1-letter difference in the word 'BEAUTIFUL', serving as the benchmark for meaningful difference. A significant p-value indicates that the differences between configuration pairs fall within the predefined bounds, suggesting practical equivalence.

Mirroring the approach in Section 6.2.3.1, the mean spelling accuracy for each configuration is plotted to visually compare within-set (Figure 6.5a) and between-set equivalences. This includes analyses with the xDAWN spatial filter (Figure 6.5b) and without it (Figure 6.5c), as well as the optimal configurations for each set (Figure 6.5d).

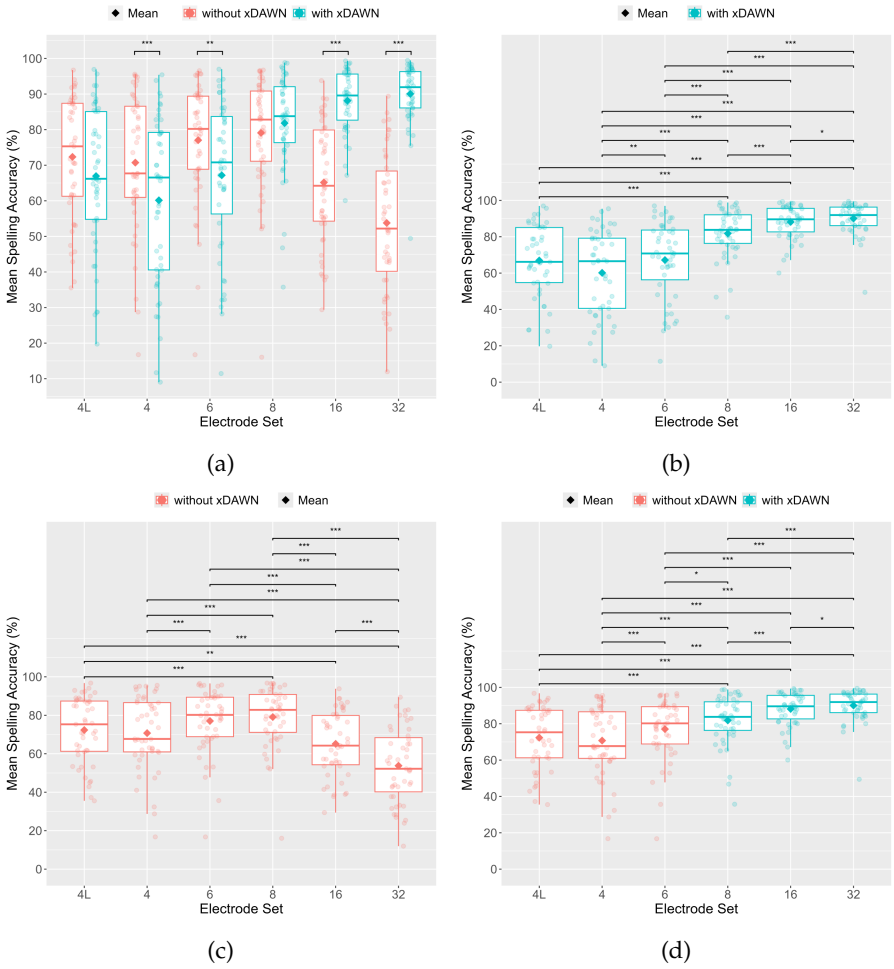


Figure 6.4: Differences in mean spelling accuracy within and between electrode sets. (a) Within-set differences with and without the xDAWN spatial filter. (b) Between-set differences with the xDAWN spatial filter. (c) Between-set differences without the xDAWN spatial filter. (d) Between-set differences with the best configuration for each set. Statistical analysis by paired t-tests and Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.

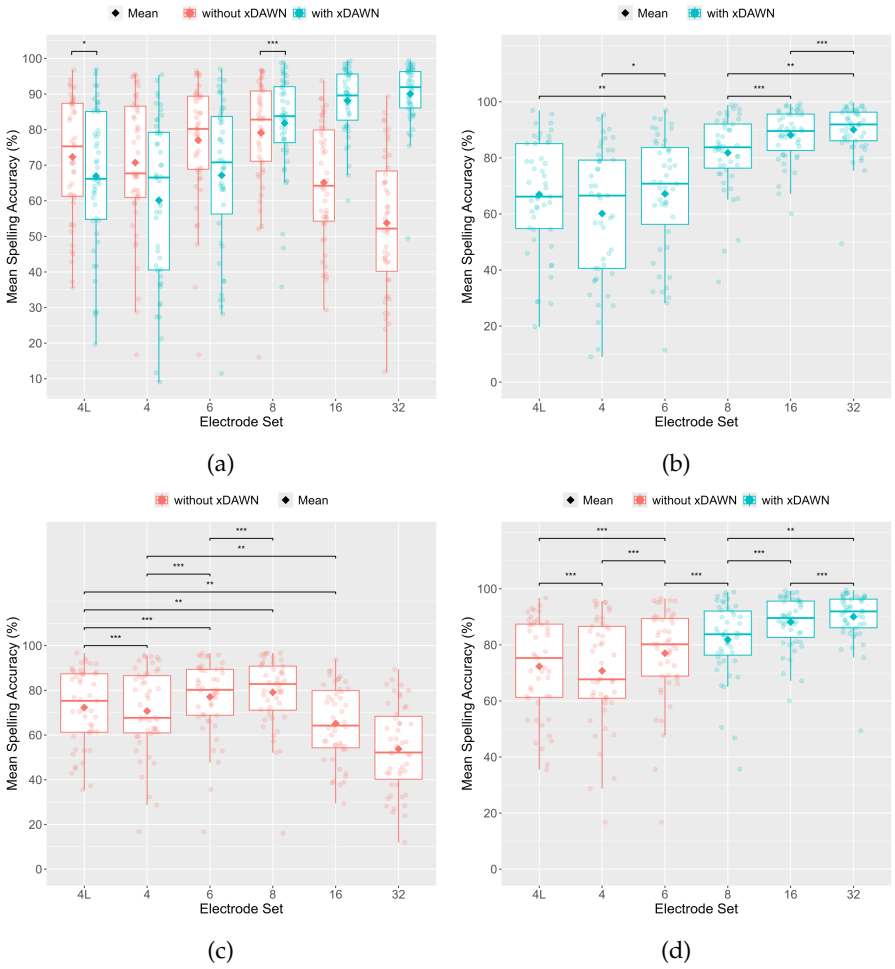


Figure 6.5: Equivalences in mean spelling accuracy within and between electrode sets. (a) Within-set equivalences with and without the xDAWN spatial filter. (b) Between-set equivalences with the xDAWN spatial filter. (c) Between-set equivalences without the xDAWN spatial filter. (d) Between-set equivalences with the best configuration for each set. Statistical analysis by t-TOST and Wilcoxon TOST, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *. TOST = two one-sided tests.

6.2.4 *Discussion*

The findings from the preceding section indicate that the xDAWN spatial filter enhances spelling accuracy for larger electrode sets (Sets 32, 16, and 8), while it appears to diminish performance in smaller sets (Sets 6, 4, and 4L). Application of the xDAWN spatial filter significantly increases spelling accuracy in the 32- and 16-electrode configurations, yet significantly reduces it in the smaller configurations of Sets 6 and 4. Interestingly, whether or not the xDAWN spatial filter is used does not make a difference in the 8-electrode set and the 4-electrode set based on [150] (Set 4L) as the spelling accuracy with these sets is statistically not significantly different, and equivalent, indicating that there is no meaningful difference. Nevertheless, using the xDAWN spatial filter still leads to better spelling accuracy on average in Set 8, and poorer spelling accuracy in Set 4L.

Accordingly, employing the xDAWN spatial filter for Sets 32, 16, and 8 is recommended, while its use is advised against for Sets 6, 4, and 4L.

Focusing on these configurations, the analysis reveals no significant or practical differences between Sets 6 and 4L, or between Sets 4 and 4L. Although Sets 6 and 4 are equivalent, they exhibit a statistically significant difference, suggesting a discernible but not practically meaningful disparity. The same goes for Sets 32 and 16, when the xDAWN spatial filter is used; Sets 32 and 8, with the xDAWN spatial filter; Sets 16 and 8, with the xDAWN spatial filter; and Set 8, with the xDAWN spatial filter, and Set 6, without the xDAWN spatial filter.

Given that these results derive from offline data analysis, experimental validation is crucial to assess how factors such as frustration may impact performance across different sets. Additionally, since the data were recorded in an electrically shielded room, the results may be more favorable than those typically expected in less controlled, real-world, conditions.

Informed by this data analysis study, it is decided to proceed with Sets 16 and 8 using the xDAWN filter, and Sets 6 and 4L without it, for further experimental comparison using real-world data. The choice to proceed with Set 4L over Set 4 is motivated by their statistical and practical equivalence, with a slightly better performance observed in Set 4L.

6.3 ELECTRODE SELECTION: STUDY

6.3.1 *Study Design*

To validate the findings of the data analysis study detailed in Section 6.2, a within-subjects study involving 10 healthy adults with-

out known neurological conditions and with normal or corrected-to-normal vision was undertaken. The study was approved by the Maynooth University Ethics Committee (BSRESC-2023-36713). It involved participants completing several P300 speller runs across different electrode sets. The electrode sets included in this study are Sets 16 and 8 with the xDAWN spatial filter, and Sets 6 and 4L without the xDAWN spatial filter (see Figure 6.2). The P300 speller runs are described in more detail in Section 6.3.1.1.

The study was conducted in two different locations, with 5 participants in each. This approach anticipated the extension of the next study, described in Chapter 7, across several institutions and aimed to assess the impact of location on spelling accuracy and EEG data quality for the selected sets. Both sites were office environments, aligning with the anticipated settings of subsequent studies and facilitating the collection of data reflective of real-world conditions. Photographs of these locations can be seen in Figure 6.6.

(a) (b)

Figure 6.6: Study locations, with 5 participants in each. (a) Location 1. (b) Location 2.

6.3.1.1 *P300 Speller Task*

Given the findings from the study detailed in Chapter 5, which revealed that training could lead to fatigue and eye strain, as well as exhibit learning effects, the sequence of electrode sets for each participant was randomised to mitigate these concerns.

In the attention training study of Chapter 5, the number of flashes progressively decreased due to the strong performance of the participants. The number of flashes in each run was therefore decreased to simulate this behaviour.

The first two runs were used to collect calibration data. The words are 'THE' and 'QUICK' with 12 flashes per row and column, as in Chapter 5.

Once the BCI system was calibrated, participants were instructed to copy-spell the word 'DANCE' 3 times for each of the 4 electrode sets under comparison, totaling 12 runs. For the first iteration of each

set, 10 flashes per row and column were used, reducing to 5 flashes for the second, and further to 3 flashes for the final iteration.

The selection of 'DANCE', a word used in the study of Chapter 5 yet comprising only 5 letters, aimed to maintain brevity. Opting for longer words would have unduly extended the duration of the training, especially considering the thrice-repeated spelling task for each set.

6.3.2 *Data Analysis*

6.3.2.1 *Spelling Accuracy*

Spelling accuracy is a critical outcome measure for the P300 speller, highly influenced by electrode selection, as demonstrated in the data analysis study of Section 6.2. Examination of the spelling accuracy also permits validation of the simulation results.

Similar to the data analysis study, the mean spelling accuracy for each participant is calculated across all runs. Subsequently, a Friedman test is conducted to analyse differences in spelling accuracy between sets. This is followed by paired t-tests and Wilcoxon signed-rank tests for pairwise comparisons. The effect size (Cohen's D [152] and correlation coefficient [153]) is also calculated for all pairwise comparisons due to the small sample size of the study. These tests are conducted without initially taking the different locations into account. The data are then split by location and Friedman/repeated measures ANOVA tests and paired t-tests/Wilcoxon signed-rank tests are conducted again. Finally, within-set differences between the locations are analysed using Wilcoxon rank-sum tests and Student's t-tests.

6.3.2.2 *Predicted Training Length*

In the study described in Chapter 5, the total number of flashes, reflecting training length, was a key outcome, with the aim of accelerating training using ILC over other methods of task difficulty adjustment. Since the number of flashes in the current study was fixed, the predicted number of flashes is analysed instead. The predicted number of flashes for the next run is determined by the ILC controller, using the actual number of flashes (i.e. 10, 5, and 3) and the spelling accuracy from each run. This is done for all 3 runs for each set, and the total predicted number of flashes is then calculated as the sum of the predicted number of flashes of these 3 runs. This yields a single performance measure for each set and each participant; the predicted training length.

The same statistical tests as for spelling accuracy are used for the predicted training length to compare between-set differences without taking location into account, between-set differences within each location, and within-set differences between the locations.

6.3.2.3 *P300 Amplitude*

The P300 amplitude at POz and Pz, common electrodes across the 4 sets included in this study, is also analysed. Target epochs were extracted from the 150 ms to 550 ms post-stimulus period of the EEG signals. These are baseline-corrected, where the 150 ms period preceding a stimulus is used as the baseline. The P300 amplitude is defined as the difference between the negative and positive peak in these target epochs. The amplitude is averaged across trials and runs for each participant and set. The same analysis as for spelling accuracy and predicted training length is then conducted.

6.3.2.4 *Effect of Location on Data Quality*

To explore potential sources of variation between locations, the mains electrical noise is analysed in relation to the EEG signals at POz and Pz. Although 50 Hz was outside the passband of the filter used (1-20 Hz), high power at 50 Hz could still indirectly impact the filtered signal, potentially introducing artifacts or distortions.

To analyse the noise, the power at 50 Hz in the raw EEG signals, which is a measure of the electrical noise in the environment, is calculated first. Then, the average power between 1 and 20 Hz from the same target epochs as described in Section 6.3.2.3, which is a measure of the task-relevant EEG activity, is calculated. Both the noise power and the EEG activity power are averaged over all runs and sets for each participant as they are very similar. Lastly, the average power at 50 Hz is divided by the average power of the EEG activity to get the noise-to-signal ratio (NSR), and the ratios are converted to decibels ($dB = 10log_{10}(\mu V^2/Hz)$).

Student's t-tests are then conducted to analyse the difference in NSR between the locations. The effect size using Cohen's D measure [152] is also calculated due to the small study sample size.

6.3.2.5 *Comparison to Offline Data Analysis Results*

Aiming to validate the data analysis study results (Section 6.2), Wilcoxon rank-sum tests and Wilcoxon TOST are conducted with $\pm$ 11% equivalence bounds to examine differences and equivalences in spelling accuracy between experimental and simulated data.

6.3.2.6 *Comparison to Performance with 32 Electrodes*

To compare the spelling accuracy achieved with the different sets in the current study with that achieved using 32 electrodes in the study described in Chapter 5, Wilcoxon rank-sum tests and Wilcoxon TOST with $\pm$ 11% equivalence bounds are conducted.

6.3.3 *Results*

6.3.3.1 *Spelling Accuracy*

A Friedman test reveals that the choice of electrode set significantly affects spelling accuracy ($\chi^2_{(3)} = 10.14$, $p = 0.017$), when location is not considered. While the pairwise comparisons do not show any significance, the effect sizes between the 16-electrode set and all other sets are large (4-electrode set: $r = 0.68$, 6-electrode set: $r = 0.54$, 8-electrode set: $r = 0.70$).

Exclusively for Location 1, a Friedman test indicates significant differences in spelling accuracy across the sets ($\chi^2_{(3)} = 9.13$, $p = 0.028$). Again, there are non-significant, but large, effect sizes between the 16-electrode set and all other sets (4-electrode set: $r = 0.92$, 6-electrode set: $r = 0.86$, 8-electrode set: $r = 0.91$).

In Location 2, repeated measures ANOVA shows no significant differences in spelling accuracy among electrode sets ($F_{(3,15)} = 0.77$, $p = 0.534$). Nevertheless, there are medium effect sizes between the 16-electrode set and all other sets (4-electrode set: $d = 0.69$, 6-electrode set: $d = 0.40$, 8-electrode set: $d = 0.65$).

Comparing within-set differences across locations reveals a significant difference in the 16-electrode set ($U = 23$, $p = 0.029$). However, large effect sizes are present in all other sets as well (4-electrode set: $r = 0.60$, 6-electrode set: $d = 1.33$, 8-electrode set: $d = 1.21$).

The mean spelling accuracy for all sets across both locations, as illustrated in Figure 6.7, varies notably between the two locations for each set. Although these differences are mostly not statistically significant, the large effect sizes suggest that with a larger sample size, these differences might reach statistical significance.

6.3.3.2 *Predicted Training Length*

Disregarding location, a significant difference in predicted training length is observed, as determined by a Friedman test ($\chi^2_{(3)} = 9.88$, $p = 0.020$). While not significant, there are medium to large effect sizes between the 16-electrode set and all other sets (4-electrode set: $r = 0.68$, 6-electrode set: $r = 0.48$, 8-electrode set: $r = 0.63$).

In Location 1, significant differences in predicted training length between sets are noted ($\chi^2_{(3)} = 10.20$, $p = 0.017$). Again, these differences lie between the 16-electrode set and the others as evidenced by large effect sizes (4-electrode set: $r = 0.92$, 6-electrode set: $r = 0.87$, 8-electrode set: $r = 0.91$).

In Location 2, no significant differences in predicted training length are found between sets ($\chi^2_{(3)} = 2.02$, $p = 0.568$), though medium effect sizes are observed between the 4- and 16-electrode sets ($r = 0.55$), and the 8- and 16-electrode sets ($r = 0.43$).

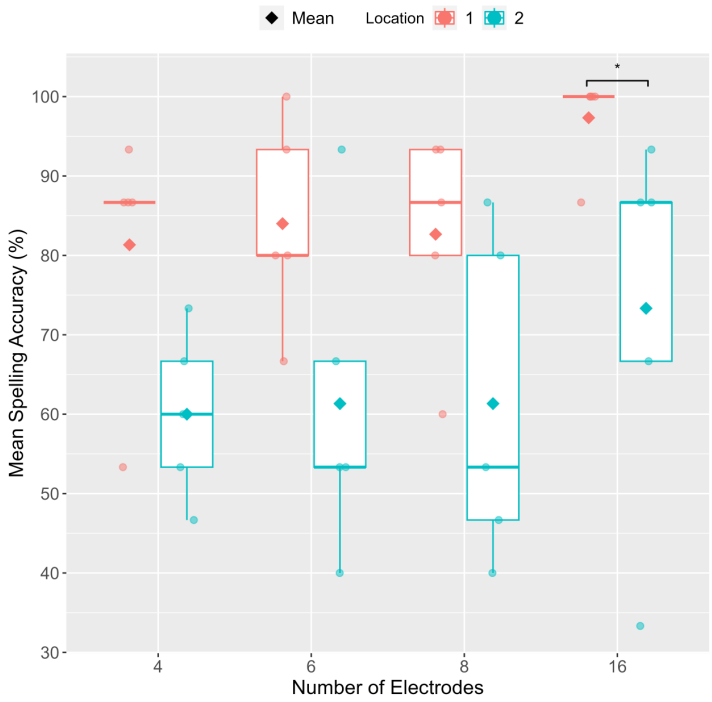


Figure 6.7: Mean spelling accuracy (%) in all sets and both locations. Statistical analysis by Student's t-tests and Wilcoxon rank-sum tests, $0.01 \leqslant p \leqslant 0.05$ *.

Echoing the findings in spelling accuracy, a significant difference in predicted training length emerges between locations for the 16-electrode set, according to a Wilcoxon rank-sum test ($U = 2.5$, $p = 0.041$), but not the others. The effect sizes for the other sets are medium to large (4-electrode set: $r = 0.57$, 6-electrode set: $d = 1.20$, 8-electrode set: $r = 0.46$).

Predicted training length for all sets across both locations is depicted in Figure 6.8. It can be seen that, in both locations, the average predicted training length is similar across the 4-, 6- and 8-electrode sets, with the 16-electrode set resulting in shorter training. The training length is longer for all sets in Location 2, compared to Location 1.

6.3.3.3 *P300 Amplitude*

A Friedman test indicates no significant differences in P300 amplitude among sets, independent of location, both for EEG signals measured at Pz ($\chi^2_{(3)} = 1.44$, $p = 0.696$) and at POz ($\chi^2_{(3)} = 3.36$, $p = 0.339$). Pairwise comparisons confirm this result at Pz; however, at POz, there is a non-significant but large effect size between the 4- and 16-electrode sets ($r = 0.53$).

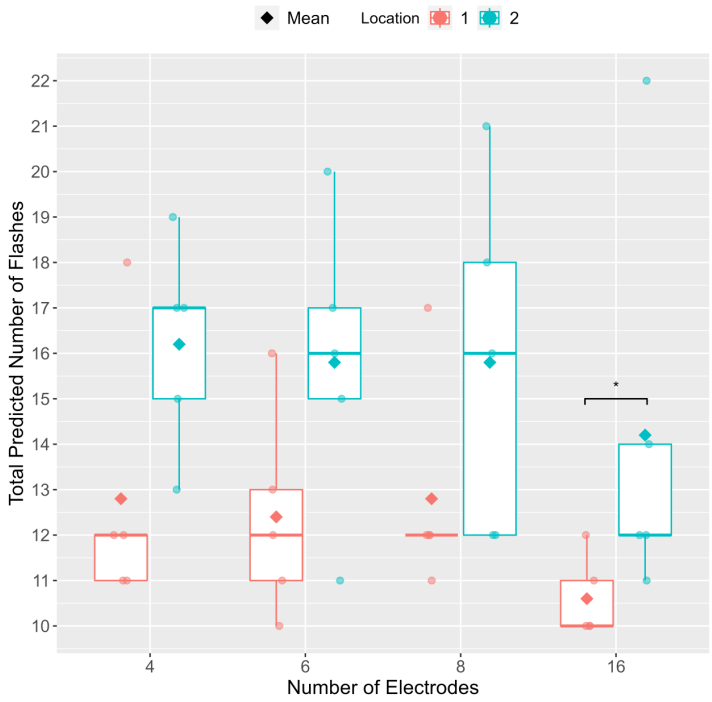


Figure 6.8: Predicted training length, measured by total predicted number of flashes. Statistical analysis by Student's t-tests and Wilcoxon rank-sum tests, $0.01 \leqslant p \leqslant 0.05$ *.

For Location 1, there are no significant differences according to repeated measures ANOVA, neither at Pz ($F_{(3,15)} = 0.78$, $p = 0.529$) nor at POz ($F_{(3,15)} = 0.53$, $p = 0.668$).

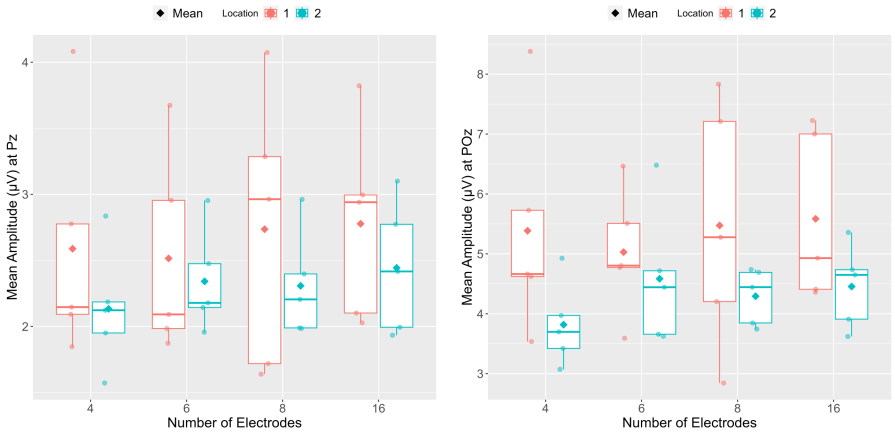
Looking at Location 2 only, there are also no significant differences at Pz ($F_{(3,15)} = 0.61$, $p = 0.620$) or POz ($F_{(3,15)} = 2.55$, $p = 0.105$). However, there are large effect sizes between the 4-electrode set and all other sets (6-electrode set: $d = 1.09$, 8-electrode set: $d = 1.10$, 16-electrode set: $d = 1.93$).

While within-set differences in P300 amplitude between the locations are non-significant at Pz and POz, there are medium to large effect sizes at POz for the 4-electrode set ($d = 1.12$), 8-electrode set ($d = 0.79$) and 16-electrode set ($d = 1.01$).

Figure 6.9 shows the P300 amplitude as measured at Pz and POz. As can be seen in the boxplots, the mean P300 amplitude is quite similar across electrode sets, with higher amplitudes at Location 1. These within-set differences are more pronounced at POz, compared to Pz.

6.3.3.4 *Effect of Location on Data Quality*

Although the difference in mains noise relative to EEG activity (NSR) is non-significant at both Pz ($t_{(8)} = -1.06$, $p = 0.328$) and POz ($t_{(8)} =$

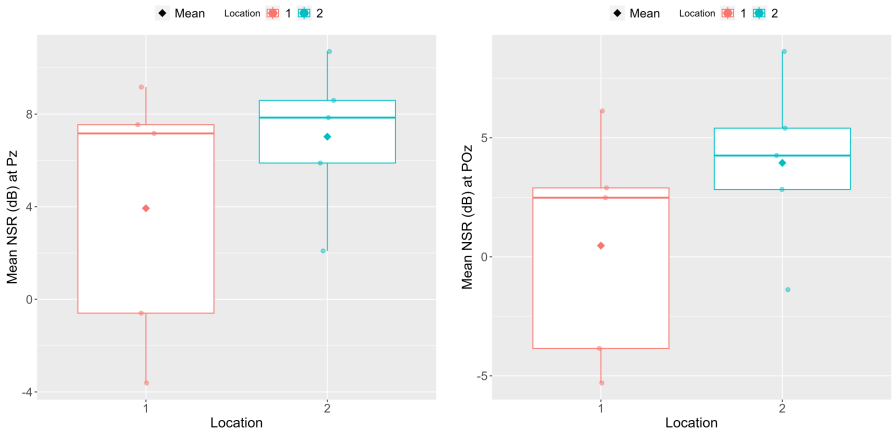


(a) (b)

Figure 6.9: Mean peak-to-peak amplitude in 150 ms to 550 ms window post-stimulus. (a) Measured at Pz. (b) Measured at POz. Statistical analysis by Student's t-tests.

-1.28, $p = 0.239$), the effect size between the locations is medium at Pz ($d = 0.67$) and large at POz ($d = 0.81$).

The mean NSRs can be seen in Figure 6.10, which shows that the noise relative to EEG activity is higher in Location 2.



(a) (b)

Figure 6.10: Mean noise-to-signal ratio (NSR) at both locations. (a) Measured at Pz. (b) Measured at POz. Statistical analysis by Student's t-tests.

6.3.3.5 *Comparison to Offline Analysis Results*

The comparison of spelling accuracy for the 4-electrode set between this experiment and the data analysis study of Section 6.2 reveals no significant equivalence ($W = 135.00$, $p = 0.067$) or difference ($U = 215.00$, $p = 0.682$) between simulated and experimental data. This means that the comparison is inconclusive.

The same is true for the 6-electrode set (equivalence: $W = 129.00$, $p = 0.198$; difference: $U = 207.00$, $p = 0.564$) and the 8-electrode set (equivalence: $W = 79.00$, $p = 0.248$; difference: $U = 174.00$, $p = 0.204$).

Only the comparison between the simulated and experimental data with the 16-electrode set yields conclusive results (equivalence: $W = 115.00$, $p = 0.006$; difference: $U = 267.00$, $p = 0.509$). This means that the simulated and experimental data are statistically and practically the same.

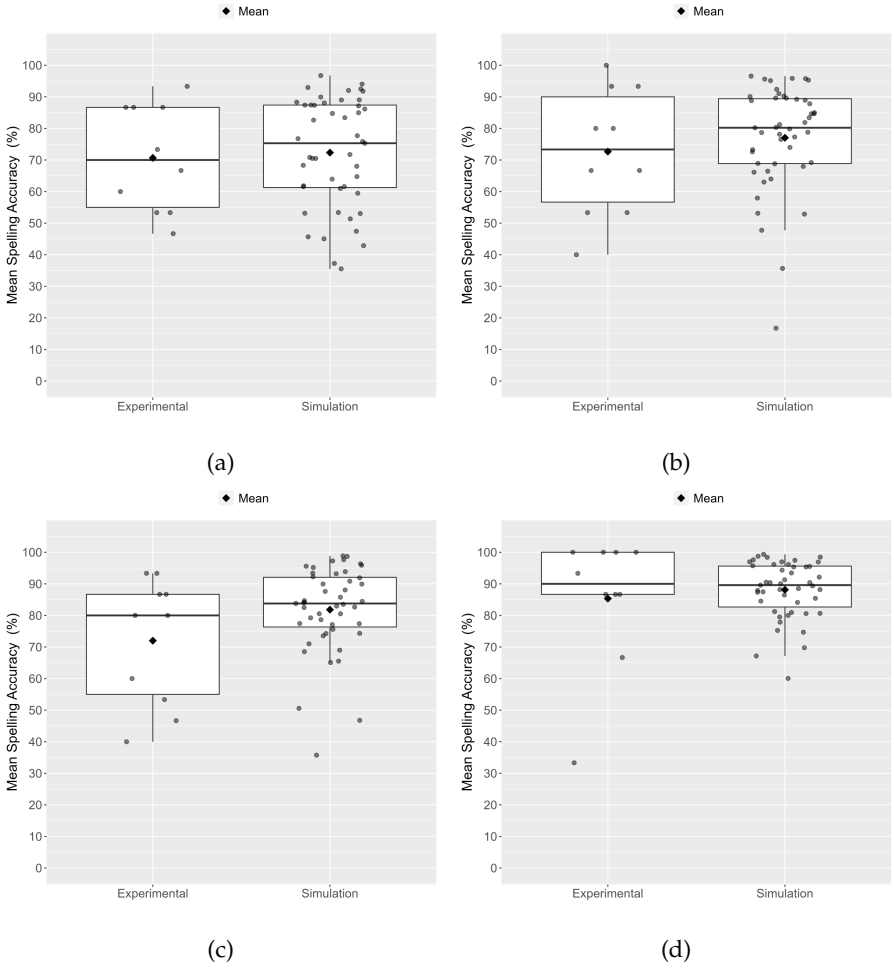


Figure 6.11: Comparison of spelling accuracy with different electrode sets in experiment and simulation. (a) 4-electrode set. (b) 6-electrode set. (c) 8-electrode set. (d) 16-electrode set. Statistical analysis by Wilcoxon rank-sum tests.

6.3.3.6 *Comparison to Performance with 32 Electrodes*

The analysis comparing spelling accuracy using the 4-electrode set with that achieved using 32 electrodes in the previous study (Chapter 5) demonstrates a significant difference ($U = 71.00$, $p < 0.001$), with no equivalence found ($W = 13.00$, $p = 0.841$), suggesting distinct performance outcomes.

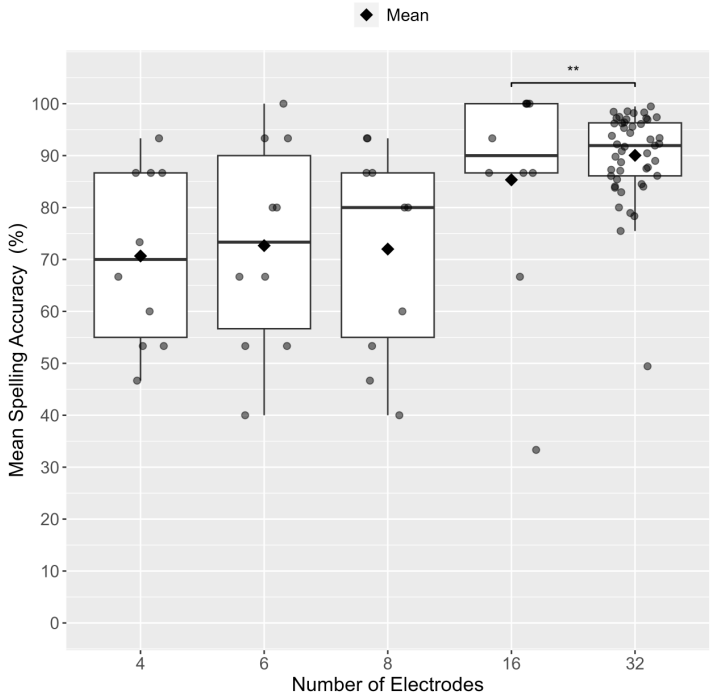


Figure 6.12: Mean spelling accuracy achieved with different electrode sets. Data from the 4-, 6-, 8- and 16-electrode sets comes from the current study, and the data from the 32-electrode set comes from Chapter 5. Equivalence tests by Wilcoxon TOST, $0.001 \leqslant p < 0.01$ **. TOST = two one-sided tests.

The same is true for the 6-electrode set (equivalence: $W = 33.00$, $p = 0.819$; difference: $U = 113.00$, $p = 0.011$) and the 8-electrode set (equivalence: $W = 16.00$, $p = 0.579$; difference: $U = 90.00$, $p = 0.002$).

In contrast, the 16-electrode set is not statistically significantly different ($U = 255.00$, $p = 0.682$) and equivalent ($W = 88.00$, $p = 0.004$) to the spelling accuracy with 32 electrodes.

6.3.4 *Discussion*

This study evaluates P300 speller performance across 4 electrode configurations (4, 6, 8, and 16 electrodes) in two different locations. The study reveals comparable spelling accuracy among the 4-, 6-, and 8-electrode sets, with the 16-electrode configuration demonstrating enhanced performance in both locations. Similarly, the predicted training length, if the ILC controller would have been used to adapt the number of flashes, is similar for the low electrode sets, and lower for the 16-electrode set. Consistently, the P300 amplitude measured at Pz and POz (common electrodes across all sets) shows no significant variation between sets.

Notably, across all metrics, performance is inferior in Location 2 when compared to Location 1. The greater level of electrical noise relative to EEG activity in Location 2 could partly explain this performance disparity. However, it is believed that this is not the only reason for the disparity. Most participants in Location 2 asked for breaks between runs due to eye strain, with one participant requesting the lights to be dimmed. In contrast, no participant in Location 1 asked for any breaks. In Location 2, the bright artificial light reflecting off the white wall behind the monitor likely contributed to increased eye strain among participants. While the increased eye strain and other environmental factors specific to Location 2 might have affected participant attention and therefore performance in the speller, the possibility cannot be discounted that the performance disparity between the locations is incidental and caused by the different participants, other environmental/situational factors such as the time the experiment was conducted, or slight differences in experimental setup. Future studies should ideally employ a within-subjects experiment design with a larger sample size to more definitively ascertain the impact of location on P300 speller performance.

Despite these considerations, the findings clearly indicate superior performance of the 16-electrode set across all evaluated metrics. It is also the only set that is equivalent to both the simulated results and the performance with 32 electrodes. The simulated performance with the 16-electrode set is also equivalent to the 32-electrode set. These results indicate that it can be assumed that the results in Chapter 5 could be replicated using only 16 instead of 32 electrodes, without a loss in performance.

It remains unclear whether the observed reduction in spelling accuracy with smaller electrode sets adversely impacts the efficacy of attention training. Since using less electrodes not only saves time and resources (both in terms of electroconductive gel and computing power) but also reduces participant discomfort, a slight loss in performance might be a worthwhile trade-off. Given the similarity between the 4-, 6- and 8-electrode sets, it is decided to conduct a replication of the attention training study (Chapter 5) using the 4-electrode set, knowing that if the 4-electrode set results in worse outcomes than the previous study, the 16-electrode set could still be used in future studies.

6.4 EVALUATING THE NEW PROTOCOL: STUDY

6.4.1 *Study Design*

To explore the effects of using a smaller electrode set, which is presumed to yield inferior performance in the P300 speller, on the attention training outlined in Chapter 5, the study is replicated with 10 healthy adults. The study was approved by the Maynooth University Ethics Committee (BSRESC-2023-36713). Except for two participants, the study took place in Location 2 due to logistical constraints, as it was the only environment available for conducting the experiments without interruptions, despite the limitations noted in the preceding section. The study procedure is mostly the same as the study of Chapter 5; however, the current study employed 4 electrodes without using the xDAWN spatial filter, in contrast to the previous use of 32 electrodes with the filter.

Moreover, the number of runs and the selection of words for copy-spelling was modified. In the Chapter 5 study, the adaptive part of the training consisted of spelling the word 'BEAUTIFUL' 5 times, which means that the number of flashes was only adapted 4 times. Aiming to increase the adaptation opportunities offered by the ILC controller, 6-letter words were chosen over a 9-letter word, and 8 different words were incorporated into the adaptive training segment. In contrast to the Chapter 5 study, different words were chosen for each run to enhance engagement, following previous participant feedback that repeating the same word was monotonous and varied letters could improve focus. Furthermore, in Chapter 5, there was a post-training word with a fixed number of flashes to allow for comparison between groups. Since there was no comparison between different groups in this study, this post-training run was not necessary. The calibration and evaluation runs remain the same. Consequently, the overall training length, in terms of the total letter count, is comparable between the new (59 letters) and previous study (61 letters). However, 50 of the 59 letters are in the adaptive part of the training in the new study,

compared to only 45 of 61 letters in the previous study. The runs are described in more detail in Section 6.4.1.3.

The remainder of the study procedure remains unchanged. All tasks are briefly described in the following subsections.

6.4.1.1 *Questionnaire*

Participants completed identical questionnaires to those used in Chapter 5. These included the fatigue-boredom questionnaire, where participants rated their fatigue, alertness, boredom, and eye fatigue on a 10-point Likert scale before and after the training, and the NASA TLX questionnaire for assessing mental, physical, and temporal demands, effort, subjective performance, and frustration, completed post-training. Further details on the questionnaires are provided in Section 5.2.1.

6.4.1.2 *Random Dot Motion Task*

Mirroring the study of Chapter 5, participants undertook 40 trials of the RDM task without feedback before, and after, the training. To familiarise participants with the task, 3 introductory runs consisting of 6 trials each were conducted, accompanied by verbal feedback. The RDM task is elaborated upon in Section 5.2.2.

6.4.1.3 *P300 Speller*

This study employed the identical P300 speller setup as used in Chapter 5. Participants were tasked with copy-spelling 11 words, listed in Table 6.1. For data analysis, runs are categorised into calibration, early training, and late training phases, as outlined in the table.

6.4.2 *Data Analysis*

6.4.2.1 *Questionnaire*

Changes in fatigue, boredom, alertness, and eye fatigue are assessed using paired t-tests and Wilcoxon signed-rank tests applied to the fatigue-boredom questionnaire scores. Descriptive statistics are computed for the scores from both the fatigue-boredom questionnaire and the NASA TLX.

6.4.2.2 *Random Dot Motion Task*

The RDM task is analysed in terms of three metrics. The first metric is accuracy, which is the percentage of correct trials in the task run. The second metric is RT, which is the average time between coherent motion onset and button press for correct trials only. Given that accuracy and RT are interrelated and their individual analysis offers an incomplete view of task performance, a composite score is calculated as

Table 6.1: Runs in the P300 speller.

Stage	Run	Word	Number of flashes
Calibration	1	THE	12
	2	QUICK	
	3	DOG	
Early Training	4	WIZARD	10
	5	HUMBLE	varying
	6	JOKERS	
	7	UNLOCK	
	8	THRIVE	
Late Training	9	JUNGLE	varying
	10	SHADOW	
	11	FROZEN	

Note: 'Varying' in the Number of flashes column indicates the adaptation of flashes based on ILC controller adjustments.

the ratio of accuracy to RT. The difference in these metrics between pre- and post-training is analysed by conducting paired t-tests and Wilcoxon signed-rank tests.

Two study participants were significantly distracted during the post-training run of the RDM task; one participant struggled to stay awake, and another participant's phone rang, which noticeably impacted their performance in the task. Consequently, the analysis of RDM task performance is repeated excluding these two distracted participants.

6.4.2.3 *P300 Speller*

The performance in the P300 speller task is analysed similarly to the study described in Chapter 5. Training length in terms of total number of flashes, as well as spelling accuracy, are assessed to compare these to the results in Section 5.5.3.

Additionally, changes in EEG signals are examined, focusing on P300 amplitude, total power during target and nontarget trials, and alpha power in nontarget trials. For this analysis, epochs from 150 ms to 550 ms post-stimulus, with baseline removal, where the 150 ms period preceding a stimulus is used as the baseline, are extracted. The amplitude is defined as the difference between the positive and negative peak in target epochs. The total power is calculated by averaging the squared samples in target and nontarget epoch averages, respectively. For alpha power analysis, signals are bandpass filtered between 7 and 12 Hz. The 150 ms period following a nontarget stimu-

lus that did not immediately follow a target stimulus is then isolated. The power is then calculated in the same way as the total power.

Paired t-tests are conducted to compare EEG metrics in the calibration stage and early training stage, and calibration stage and late training stage, respectively.

6.4.2.4 *Comparison to Previous Study*

The outcomes of this study are benchmarked against the results from the ILC group in the preceding study (Chapter 5), using a variety of statistical tests for comprehensive comparison.

Fatigue-boredom questionnaire scores are analysed using repeated measures ANOVA, using ranked data for non-normally distributed datasets. For NASA TLX scores, Wilcoxon rank-sum tests and Student's t-tests are used.

Performance metrics of the RDM task are compared by repeated measures ANOVA (on ranked data if necessary).

Wilcoxon TOST and t-TOST for equivalence testing, alongside Wilcoxon rank-sum tests and Student's t-tests for difference testing, are conducted to compare spelling accuracy between this study and the previous one. Consistent with prior analyses, an equivalence bound of 11% is set.

The mean number of flashes is compared to the mean number of flashes in the previous study using Wilcoxon rank-sum tests and Student's t-tests. Since the number of runs is different between the studies, the total number of flashes is not comparable.

EEG changes are compared by Wilcoxon rank-sum tests and Student's t-tests.

6.4.3 *Results*

6.4.3.1 *Questionnaire*

Figure 6.13 shows the pre- and post-training scores for each question of the fatigue-boredom questionnaire. While fatigue scores increased post-training, this increase is not significant, according to a paired t-test ($t_{(9)} = -0.92$, $p = 0.38$). On average, self-reported alertness decreased after the training, but this is also not significant ($t_{(9)} = 1.60$, $p = 0.144$). A non-significant increase in boredom according to a Wilcoxon signed-rank test ($W = 0$, $p = 0.174$) is observed. Only eye fatigue increased significantly post-training ($t_{(9)} = -4.63$, $p = 0.001$).

Figure 6.14 shows the NASA TLX scores for each question. It can be seen that mental demand, performance and frustration yielded the highest scores. It should be noted that a high score for performance means that participants perceived their performance as poor. On the other hand, physical and temporal demand, and frustration were scored low to neutral, on average.

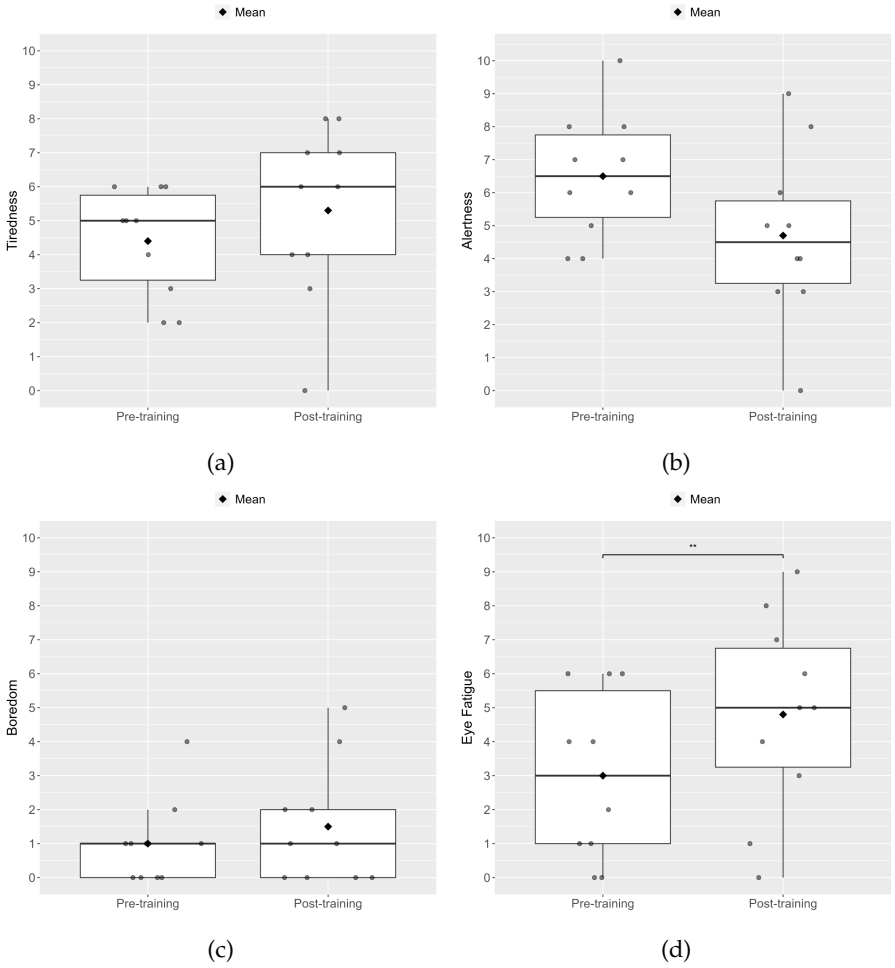


Figure 6.13: Scores of the fatigue-boredom questionnaire. (a) Q1 - Fatigue. (b) Q2 - Alertness. (c) Q3 - Boredom. (d) Q4 - Eye Fatigue. Statistical analysis by paired t-test and Wilcoxon signed-rank tests, $0.001 \leqslant p < 0.01$ **.

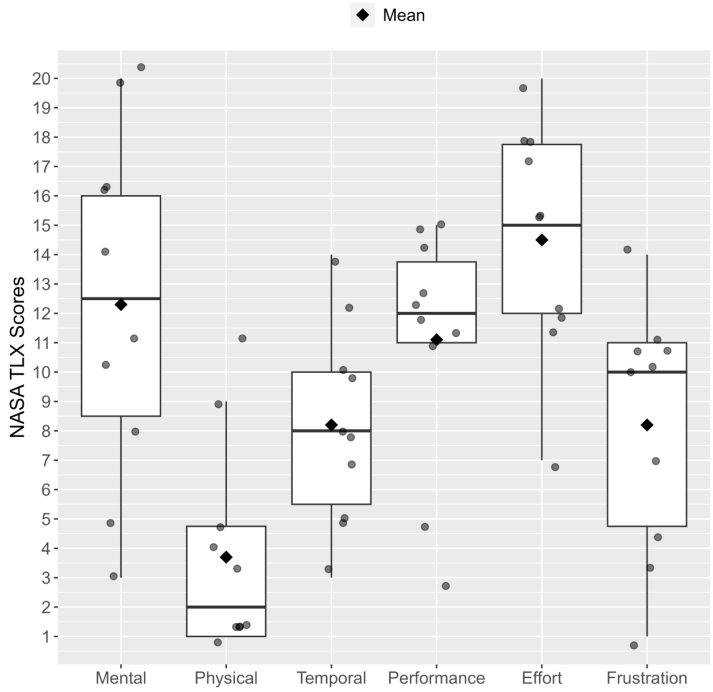


Figure 6.14: Scores of the NASA Task Load Index (TLX) questionnaire.

6.4.3.2 *Random Dot Motion Task*

Figure 6.15 shows the accuracy, RT and score in the RDM task before and after the training. This figure shows all participants. It can be seen that the changes between pre- and post-training are small, and are found to be non-significant, as confirmed by Wilcoxon signed-rank tests and paired t-tests (accuracy: $W = 23.5$, $p = 0.953$, RT: $W = 38$, $p = 0.322$, score: $t_{(9)} = -0.92$, $p = 0.383$).

When the two participants who were distracted in the post-training RDM task are excluded from analysis, these results change. While changes in accuracy are still non-significant ($W = 10.5$, $p = 0.612$), there is now a non-significant but large effect size in RT ($r = 0.55$), and a significant increase in score post-training ($t_{(9)} = -3.39$, $p = 0.012$). These changes can be seen in Figure 6.16.

6.4.3.3 *P300 Speller*

The spelling accuracy over the course of the training (runs 4 to 11) is depicted in Figure 6.17. The training sessions averaged a spelling accuracy of 62.29%, with a standard deviation of 10.60%.

Figure 6.18 shows the number of flashes for each adaptive run (i.e. runs 5 to 11). On average, 7.1 flashes per row and column ($\pm$ 3.8) were used in each run, with an average total of 49.6 flashes ($\pm$ 26.3).

Figure 6.19a presents the average P300 amplitude across different training stages, revealing no significant changes. This observa-

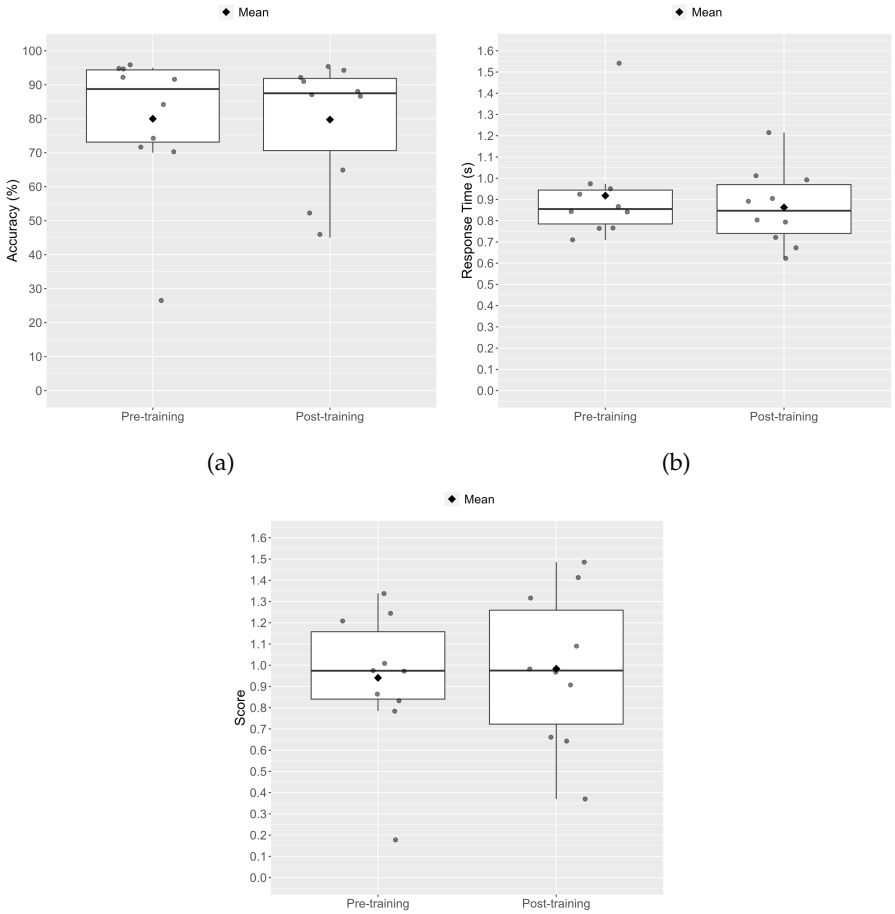


Figure 6.15: Performance in the Random Dot Motion (RDM) task. (a) Accuracy (%). (b) Response time (RT, s). (c) Score, calculated as accuracy divided by response time. Statistical analysis by paired t-tests and Wilcoxon signed-rank tests.

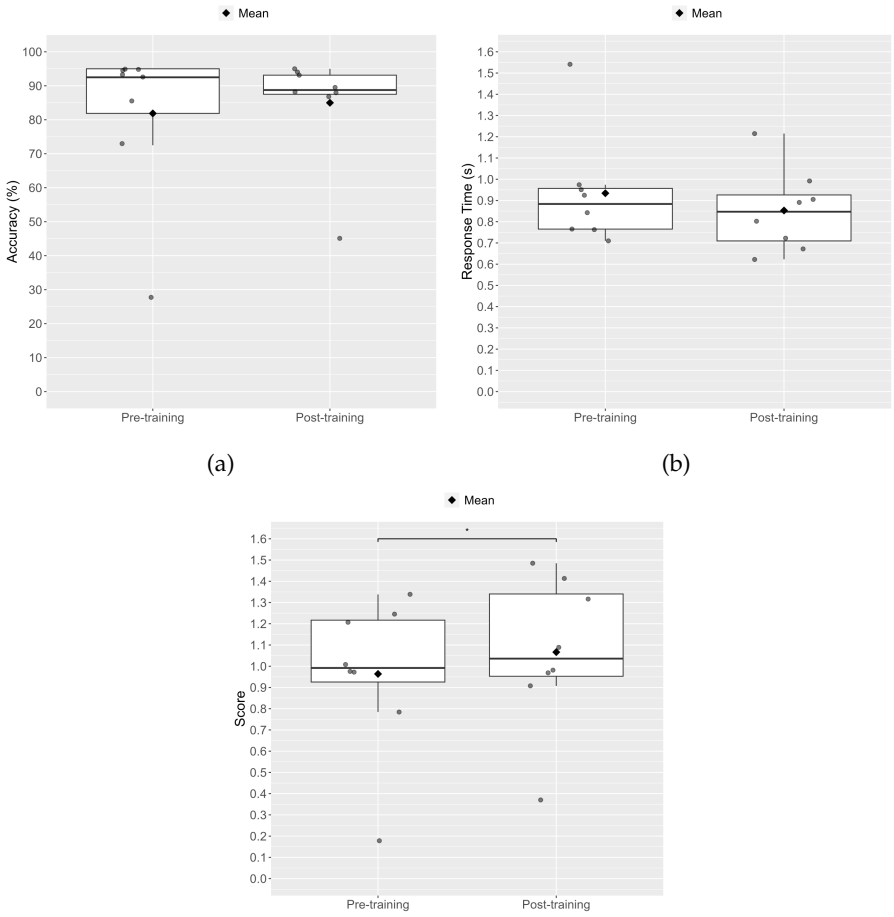


Figure 6.16: Performance in the Random Dot Motion (RDM) task, with two participants that were distracted excluded from analysis. (a) Accuracy (%). (b) Response time (RT, s). (c) Score, calculated as accuracy divided by response time. Statistical analysis by paired t-tests and Wilcoxon signed-rank tests, $0.01 \leqslant p \leqslant 0.05$ *.

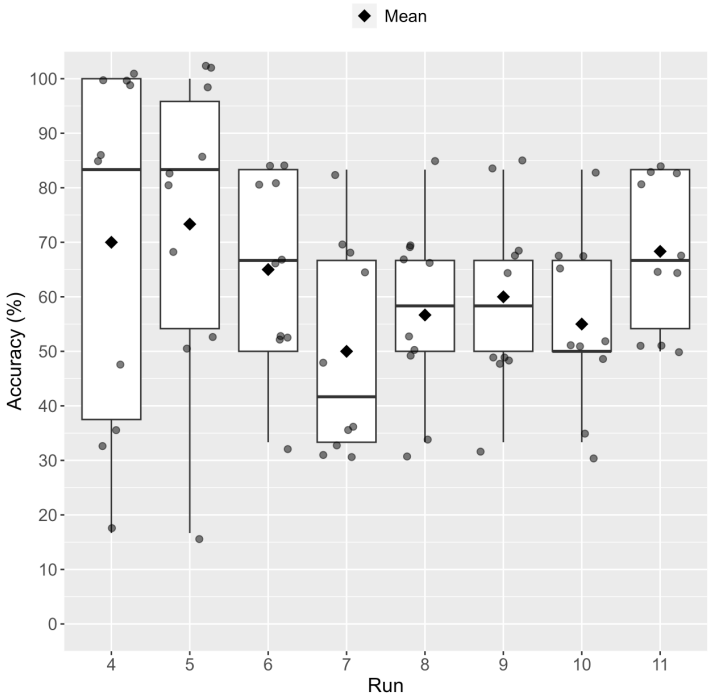


Figure 6.17: P300 spelling accuracy throughout the training.

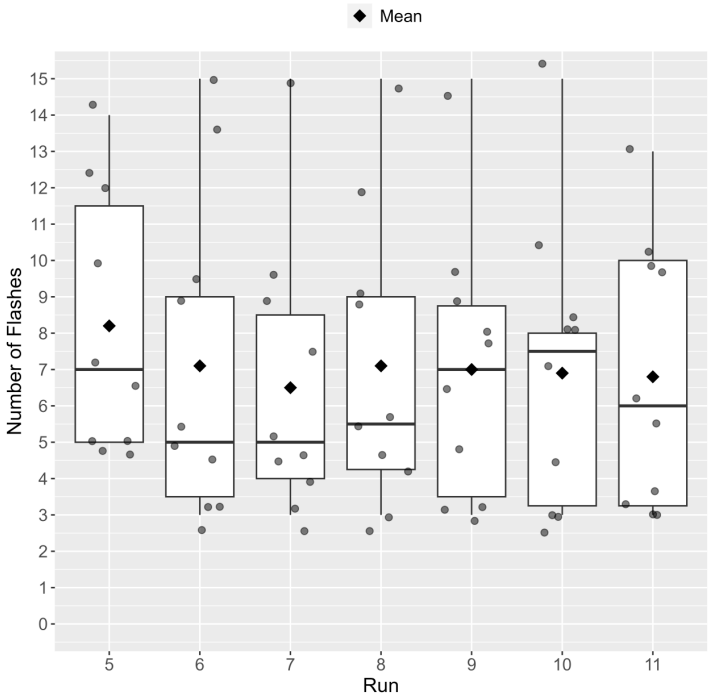
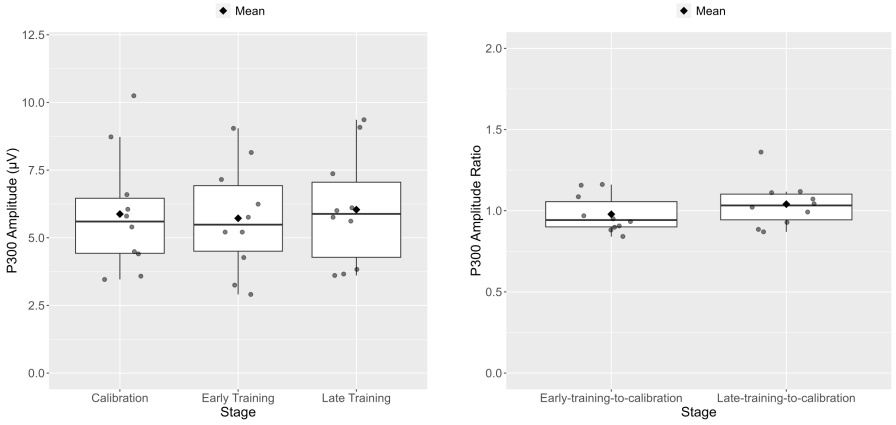


Figure 6.18: Number of flashes per row and column throughout the training.

tion is supported by paired t-tests, which show no significant difference when comparing the amplitude in the early training stage to the calibration stage ($t_{(9)} = 0.75$, $p = 0.475$), as well as the late training stage to the calibration stage ($t_{(9)} = -0.65$, $p = 0.530$). Figure 6.19b illustrates the early-training-to-calibration and late-training-to-calibration ratios. It can be seen that the late-training-to-calibration ratio is slightly above unity on average, indicating a slight (non-significant) increase in P300 strength.



(a) (b)

Figure 6.19: P300 peak-to-peak amplitude. (a) Mean amplitude in each stage. (b) Amplitude ratios of the different stages. Statistical analysis by paired t-tests.

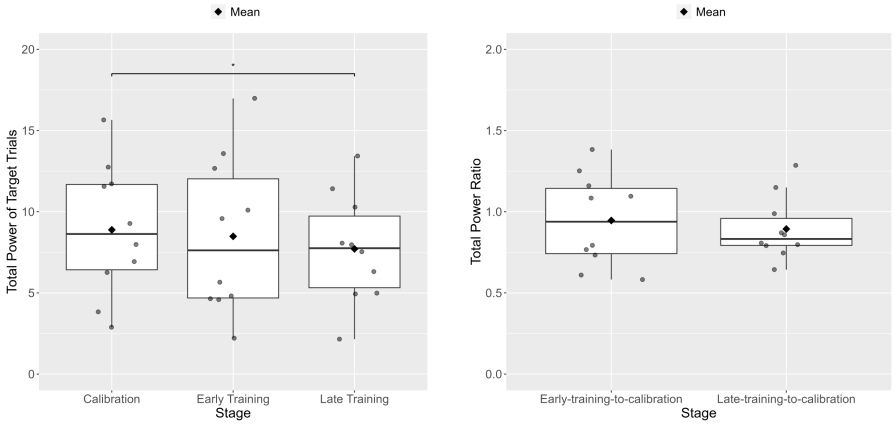
Analysis of total power in target epochs, detailed in Figure 6.20, shows that while the increase from the calibration to the early training stage is not statistically significant ($t_{(9)} = 0.55$, $p = 0.593$), a significant reduction in power is observed from calibration to the late training stage ($t_{(9)} = 2.31$, $p = 0.046$).

The changes in total power of nontarget epochs, which can be seen in Figure 6.21, are not significant (calibration to early training: $t_{(9)} = 1.42$, $p = 0.189$, calibration to late training: $t_{(9)} = 1.76$, $p = 0.112$).

The same goes for alpha power following nontarget stimuli, presented in Figure 6.22, where changes from calibration to early training and calibration to late training are non-significant ($t_{(9)} = 1.00$, $p = 0.342$ and $t_{(9)} = 1.95$, $p = 0.083$, respectively).

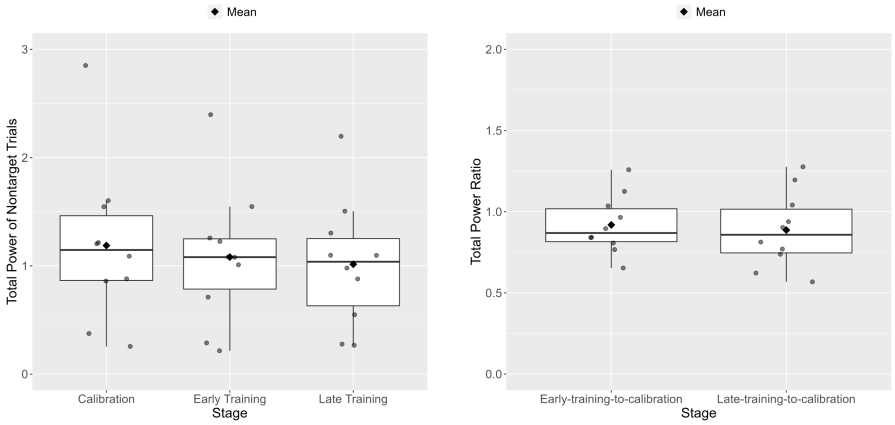
6.4.3.4 *Comparison to Previous Study*

There is no statistically significant difference in fatigue-boredom questionnaire scores between the current study and the one in Chapter 5, according to repeated measures ANOVA. NASA TLX questionnaire performance scores significantly differed ($U = 38.5$, $p = 0.004$), with participants in the current study rating their performance as worse than participants in the previous study (mean score of 11.1 ($\pm$ 4.04)



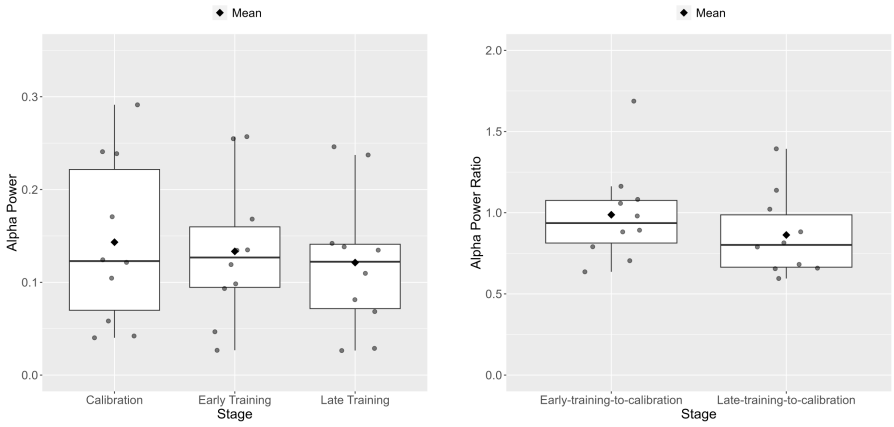
(a) (b)

Figure 6.20: Total power of target trials. (a) Mean power in each stage. (b) Power ratios of the different stages. Statistical analysis by paired t-tests, $0.01 \leqslant p \leqslant 0.05$ *.



(a) (b)

Figure 6.21: Total power of nontarget trials. (a) Mean power in each stage. (b) Power ratios of the different stages. Statistical analysis by paired t-tests.



(a) (b)

Figure 6.22: Alpha power in nontarget trials. (a) Mean power in each stage. (b) Power ratios of the different stages. Statistical analysis by paired t-tests.

and 8.9 ($\pm$ 2.61), respectively). No significant differences exist in the other questions.

Similarly, there are no statistically significant differences in the RDM task (accuracy, RT and score), according to repeated measures ANOVA.

A significant difference in spelling accuracy is observed ($t_{(23)} = 6.83$, $p < 0.001$), with the current study achieving a mean of 62.29% ($\pm$ 10.60%), compared to 88.44% ($\pm$ 7.19%) in the previous study. The accuracies are not equivalent within 11% ($t_{(23)} = 3.96$, $p = 1$).

The mean number of flashes is also statistically significantly different ($U = 13$, $p < 0.001$), with an average of 7.1 ($\pm$ 3.8) flashes in the current study, and only 3.3 ($\pm$ 0.5) flashes in the previous study.

Comparing the training-to-calibration ratios from the previous study, to the early-training-to-calibration ratios from this study, reveals a significant difference for all EEG metrics (P300 amplitude: $t_{(23)} = -3.16$, $p = 0.005$, total target power: $U = 20$, $p = 0.001$, total nontarget power: $U = 15$, $p < 0.001$, alpha power: $t_{(23)} = -3.24$, $p = 0.004$). No significant differences exist between the post-training-to-calibration and late-training-to-calibration ratios.

6.4.4 *Discussion*

This study replicates the experiment detailed in Chapter 5, using a reduced electrode set of only 4 electrodes and incorporating a greater number of shorter words. The study outcomes in terms of questionnaire scores, performance in the RDM task and P300 speller, and EEG changes throughout the training are analysed.

Consistent with previous findings, participants experienced eye fatigue as a result of the training, corroborating feedback from Sec-

tion 6.3 regarding significant eye strain, particularly in Location 2. The questionnaire scores are quite similar to the previous study. This suggests that the perceived training workload remains consistent regardless of electrode count, varying setups, and environmental conditions. Only the perceived performance is different between this study and the study of Chapter 5. This is expected as the spelling accuracy was indeed lower in this study compared to the one in Chapter 5.

In Chapter 5, a significant increase in accuracy in the ILC group was seen, but not in RT. In the current study, accuracy changes were minimal, with a notable trend toward improved RT among participants. This is in line with the results of the study by Arvaneh et al. [96], although a significant change is only observed when two participants who were heavily distracted during the post-training RDM task are excluded from analysis.

As anticipated, spelling accuracy within the P300 speller task is significantly lower in this study compared to Chapter 5, likely attributable to the reduced electrode configuration. The reduced spelling accuracy resulted in a much higher number of flashes in the training. While almost all participants in the ILC group in the Chapter 5 study reached a single flash per row and column at the end of the training, no participant in this study got to less than 3 flashes per row and column.

Contrary to the study of Chapter 5, which demonstrated an increase in target power and a decrease in alpha power post-training, the current study failed to replicate these positive EEG signal changes. This is likely caused by the poor performance eliminating the training effect seen in Chapter 5.

The rationale behind using fewer electrodes is primarily to expedite the experiment setup process, a significant consideration given that the setup duration often exceeded the training time in the Chapter 5 study. Using 4 electrodes sped up the setup significantly, with less than 10 minutes for most participants. However, the current study results show that the training itself was significantly longer as a result of using less electrodes, and the training effect was significantly lower. This indicates that the primary advantage of using fewer electrodes is negated.

Considering these study outcomes and previous findings described in this chapter suggesting equivalence between 16- and 32-electrode configurations, 16 electrodes will be used in future studies.

6.5 SUMMARY

This chapter describes three different studies, comprising both experimental and analytical approaches. Motivated by the lengthy setup time observed in the study of Chapter 5 and limited use of the ILC

controller, these studies aim to refine the attention training protocol detailed in Chapter 5.

- **Electrode Selection: Offline Analysis**
  - Overview: Offline analysis using data from Chapter 5 to identify optimal electrode configurations with fewer than 32 electrodes, based on xDAWN weights. P300 speller performance is assessed with electrode sets ranging from 4 to 32 electrodes, with and without using the xDAWN spatial filter.
  - Key findings: Larger electrode sets (32, 16, and 8 electrodes) achieved higher accuracy with the xDAWN spatial filter, while smaller sets (6 and 4 electrodes) performed better without it.
- **Electrode Selection: Study**
  - Overview: Within-subjects study with 10 healthy adults to compare 4 electrode sets (16, 8, 6, and 4 electrodes) identified in the offline analysis. The study was conducted in two different office locations to assess P300 speller accuracy in real-world conditions.
  - Key findings: The 16-electrode set achieved the best overall performance in both locations, while the 4-, 6-, and 8-electrode sets had comparable but lower accuracy. Performance was generally lower in Location 2, possibly due to environmental factors like lighting causing increased eye strain.
  - Implications: The 16-electrode set was validated as the most efficient configuration, achieving shorter predicted training times and high accuracy even in less controlled environments. However, reduced P300 speller performance does not necessarily compromise training efficacy.
- **Evaluating the New Protocol: Study**
  - Overview: The study in Chapter 5 was replicated with only 4 electrodes to explore the impact of a minimal electrode set on training efficacy.
  - Key findings: The 4-electrode setup resulted in significantly lower spelling accuracy, longer training times, and less pronounced EEG changes, suggesting reduced training efficacy.
  - Implications: Reduced P300 speller performance significantly compromises training effectiveness. This means that the saved time, resources, and cost due to a small electrode set are not a worthwhile trade-off in this case.

Given these outcomes, it can be concluded that future studies should employ a 16-electrode setup, which offers a balance between reduced setup time and maintained P300 speller performance. However, future research should investigate whether an electrode set between 8 and 16 electrodes might provide an even better trade-off, further reducing setup time while still maintaining training efficacy.

7

ATTENTION TRAINING FOR IMPROVED MOTOR SKILL LEARNING IN SURGICAL TRAINING

7.1 MOTIVATION

The P300-based attention training developed in this research project proved effective, as demonstrated in Chapter 5, where the ILC controller significantly accelerated the training process. However, the attention improvements in both Chapter 5 and Section 6.4 were assessed using the RDM task described in Section 5.2.2. While the RDM task is a highly specific and effective metric for evaluating attention in a controlled environment, exploring how attentional improvements following NFB training transfer to real-world scenarios is of significant interest.

This chapter describes a study conducted to assess the efficacy and feasibility of P300-based attention training in a real-world context, specifically within surgical training.

Surgical training must meet high standards, yet teaching opportunities are often limited due to the demanding schedules of experienced surgeons and increasing numbers of surgical trainees [10, 154]. Consequently, developing methods to accelerate surgical training without compromising skill acquisition is crucial.

Attention plays a role in motor learning in several ways. Firstly, selective attention significantly affects motor skill performance, as demonstrated by studies using dual-task paradigms. In these studies, participants perform a motor task alongside a cognitive task (such as mental arithmetic), requiring divided attention, which often leads to a decrease in performance [155]. In the context of surgical skills, dual-task paradigms have shown that divided attention negatively impacts performance, particularly among novice surgeons, affecting either the secondary cognitive task [156, 157] or both the primary and secondary tasks [158]. This suggests that enhancing selective attention during surgical training could enhance skill acquisition and performance.

Additionally, the concept of externally focused attention, where individuals concentrate on visual information and task-relevant cues rather than internal sensations, has been proposed to improve motor learning and skill retention [159]. Studies have shown that directing attention externally enhances the learning and retention of various athletic motor skills [155, 160]. This theory suggests that interventions aimed at improving sustained visual attention, such as P300-based NFB training, could positively impact surgical motor skills.

Table 7.1: Number of participants in each study group.

Group 1	Group 2	Group 3	Group 4	Group 5
17	16	18	18	19

Given these considerations, it is hypothesised that the NFB attention training developed in this thesis will enhance surgical skill performance in surgical trainees.

In this study, medical students are introduced to surgical tasks for the first time, with some students undergoing one or two NFB training sessions. Their performance is assessed both immediately after learning the tasks and again up to 7 weeks later, comparing outcomes between those who received NFB training and those who did not.

The study design is detailed in Section 7.2. Data analysis methods are explained in Section 7.3, followed by the results in Section 7.4. These results are discussed in Section 7.5, and the chapter concludes in Section 7.6.

7.2 STUDY DESIGN

The study involved 5 groups of medical students, with the aim of 20 participants in each group. The sample size was determined based on previous literature, which investigated the use of cognitive training in surgical education [154]. Students were recruited from the Royal College of Surgeons in Ireland and University College Dublin, with experiments conducted at those universities and the Beacon Hospital. Ethical approval was obtained from all participating institutions.

A total of 261 students enrolled in the study, with 115 attending the workshop described below. 84 NFB training sessions were conducted with 65 students (i.e. 19 students participated in two sessions). After losing 27 students to follow-up, 88 students are included in this analysis. The distribution of participants across the groups is shown in Table 7.1.

Figure 7.1 gives an overview of the tasks that were completed by each group. All participants, who were naive to surgical tasks, attended a workshop where they learned suturing and laparoscopic surgery skills from a surgeon. These skills were tested in two sessions: the first test was conducted shortly after the workshop (on the same day for most participants), and the second test occurred between 1 and 7 weeks later (with an average of 3 weeks) to evaluate skill retention. The surgical tasks are described in more detail in Section 7.2.1.

After the first test, participants completed a questionnaire on their sleep hygiene during the 4-week period preceding the study, including questions on sleep length and quality. They also filled out the NASA TLX after both tests to assess their subjective workload during

the surgical tasks. However, the data from these questionnaires is not yet available and is therefore not included in this analysis.

Group 1 participants did not undergo any additional training; they only completed the workshop and the two tests. Group 2 participants performed cognitive simulation techniques between the two tests, as described in Section 7.2.3. Participants in Groups 3, 4, and 5 completed a single NFB training session before the first test, detailed in Section 7.2.2. Additionally, Group 4 completed a second NFB session before the second test. Group 5 participants received both NFB training before the first test and engaged in cognitive simulation between the tests.

Upon completing the study, all participants received a suturing training kit, including custom-made synthetic skin and suturing supplies. They were also offered refreshments after the second test.

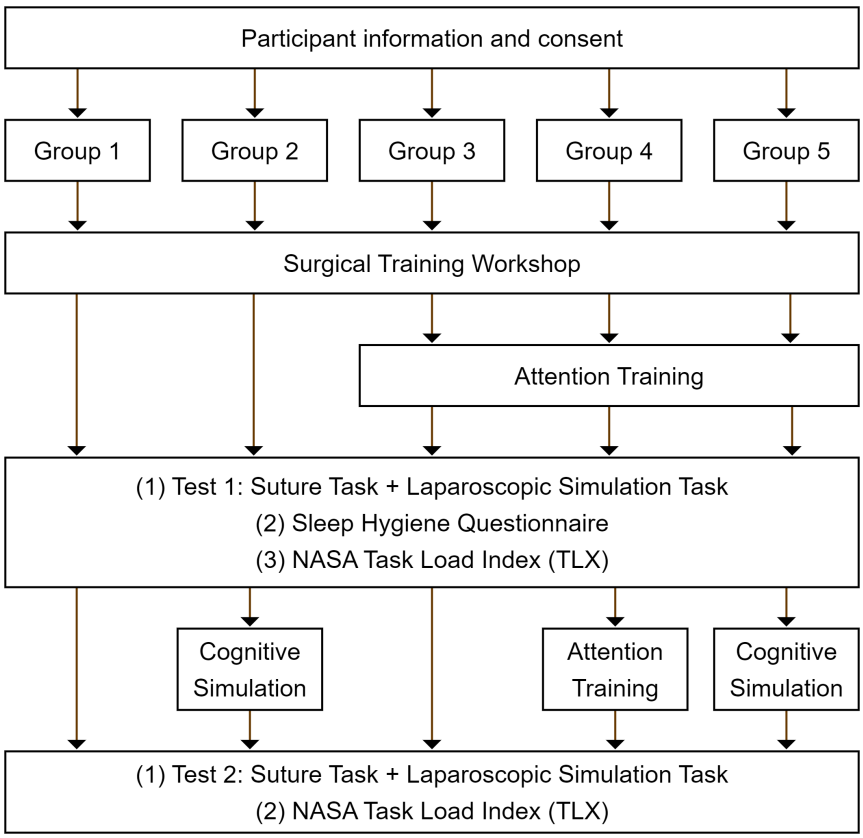


Figure 7.1: Overview of study procedure.

7.2.1 *Surgical Training*

All participants attended a surgical training workshop, lasting up to 3 hours, where an experienced surgeon taught various surgical skills.

After instruction, students practiced these skills, receiving help and feedback from the surgeon, who also answered their questions.

Participants completed two tests, where their surgical skills were recorded on video for later evaluation. The first test took place shortly after the workshop, and the second test occurred between 1 and 7 weeks later. The surgical tasks that were evaluated in these tests are described in the following subsections.

7.2.1.1 *Suturing Task*

In the suturing task, participants performed various suturing techniques on synthetic skin while their hands and the synthetic skin were recorded. To ensure anonymity, all participants wore white latex gloves and were asked to remove any jewelry.

These recordings were later independently analysed and scored by 3 different experienced surgeons according to a scoring sheet, which evaluates the quality of the suturing technique. At the time of writing this thesis, the scoring of this task has not been completed, and is therefore not included in the analysis and results presented in this chapter.

7.2.1.2 *Laparoscopic Simulation Task*

Participants also completed a laparoscopic simulation task using a laparoscopic training box, similar to the one shown in Figure 7.2. They were instructed to move pegs from one side of the board to the other, passing the pegs between tools in mid-air. Participants were told to ignore any dropped pegs. They were only able to see the board through a screen in front of them, which was recorded.

Figure 7.2: Example of laparoscopic training box with peg board. Source: [161].

The laparoscopic simulation task was scored according to a modified version of the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) [162]. The score is measured in time (seconds) required to complete the task, with a penalty of 15

seconds added for each dropped peg. Thus, a lower score indicates improved performance, reflecting faster task completion and fewer errors. Both task completion time and accuracy have been shown to deteriorate under dual-task conditions [158, 163], which suggests that these metrics are impacted by attention. Due to the objectivity of this score, the recordings for this task were assessed by only one person.

7.2.2 *Attention Training*

Groups 3, 4, and 5 completed one or two NFB training sessions, each lasting less than an hour, including setup. These sessions were conducted shortly after the workshop and before the first test. Due to scheduling issues, some participants had a longer interval between the workshop, training, and testing. Group 4 participants completed the attention training again before the second test.

The attention training followed the protocol described in Section 6.4, with the modification of using 16 electrodes instead of 4. An overview of the training session is provided in Table 7.2 for reference.

Table 7.2: Runs in the P300 speller.

Stage	Run	Word	Number of flashes
Calibration	1	THE	12
	2	QUICK	
	3	DOG	
Early Training	4	WIZARD	10
	5	HUMBLE	varying
	6	JOKERS	
	7	UNLOCK	
	8	THRIVE	
Late Training	9	JUNGLE	varying
	10	SHADOW	
	11	FROZEN	

Note: 'Varying' in the Number of flashes column indicates the adaptation of flashes based on ILC controller adjustments.

7.2.3 *Cognitive Simulation*

Participants in Groups 2 and 5 were instructed to use cognitive simulation techniques, involving the visualisation of the surgical tasks, as frequently as possible between the two tests. Recognising the link

between motor imagery and motor execution [164, 165], cognitive simulation is emerging as an effective training technique in surgical training [10, 154, 166].

Participants were guided by an app called "Surgical CogSim", developed by Amodisc, and their activity levels on the app were monitored to assess adherence to the study protocol.

Although the results of the cognitive simulation groups are included in this chapter for completeness, the cognitive simulation aspect is not the primary focus and is therefore not analysed in detail. Cognitive simulation was included in this study as both an alternative and complementary cognitive training method to NFB training. This approach allowed for a comparison of the effectiveness of each method individually, as well as in combination, to determine whether one is superior or if their combination provides greater benefits.

7.3 DATA ANALYSIS

7.3.1 *Offline EEG Processing*

The EEG signals are processed similarly to the procedures outlined in Section 5.4.1, with one difference. The signals are re-referenced to FC1 instead of Fz, as the Fz electrode is not used in this study. FC1 was chosen because it is the closest available electrode to Fz. Electrode selection (C3, Cz, C4, P3, Pz, P4), filtering, epoch segmentation, and rejection follow the methods detailed in Section 5.4.1.

7.3.2 *Surgical Training*

A repeated measures ANOVA is conducted to analyse the scores across the stages (first test, second test) and between groups. The analysis is performed on ranked data to address non-normality in some conditions. The time interval between the first and second tests is included as a covariate, as it may significantly impact the scores.

Post-hoc Kruskal-Wallis tests are used to explore between-group differences for both tests, while paired t-tests and Wilcoxon signed-rank tests are applied to examine within-group differences.

7.3.3 *Attention Training*

The outcomes analysed in this study mirror those in Section 5.4.4 and Section 6.4.2.3. For Group 4, only the first session is included in this analysis, with a separate analysis of the repeat NFB sessions detailed in Section 7.3.3.1.

Performance in the P300 speller is assessed by analysing the total number of flashes in the session and the mean spelling accuracy

across all runs. Due to non-normality in the data, Kruskal-Wallis tests are used to investigate between-group differences in these outcomes.

Changes in EEG signals are investigated using P300 peak-to-peak amplitude, total power in target and nontarget trials, and alpha power in nontarget trials. Repeated measures ANOVA tests on ranked data are conducted to compare these metrics across different training stages. Post-hoc Kruskal-Wallis tests examine between-group differences, while paired t-tests and Wilcoxon signed-rank tests analyse within-group differences.

7.3.3.1 *Repeat Neurofeedback Session*

To investigate changes in P300 speller performance and EEG signals across NFB sessions, paired t-tests and Wilcoxon signed-rank tests are conducted on mean spelling accuracy, total number of flashes, and the early-training-to-calibration and late-training-to-calibration ratios of the previously mentioned EEG metrics.

Agreement between the xDAWN spatial filter weights across the two sessions is assessed using intraclass correlation coefficients (ICCs) [167] for each participant. An ICC of 0 indicates no agreement, while an ICC of 1 signifies perfect agreement.

7.3.4 *Correlation between Attention Training and Surgical Training*

Following the approach outlined in Section 5.4.5, potential correlations between attention training and surgical task performance are investigated.

The metric used for the surgical training task is the difference in scores between the first and second tests. Attention training performance is quantified by mean and minimum spelling accuracy. Changes in EEG signals are quantified using the early-training-to-calibration and late-training-to-calibration ratios of P300 amplitude, total power of target and nontarget trials, and alpha power of nontarget trials.

Pearson's correlation coefficient is applied to normally distributed data, while Spearman's correlation coefficient is used for non-normal data. These tests are conducted without considering group assignments, as there are no significant between-group differences in the metrics used.

7.4 RESULTS

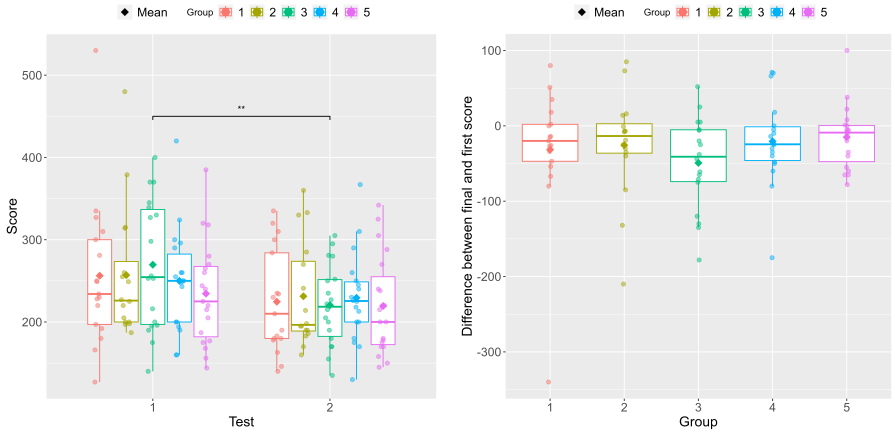
7.4.1 *Surgical Training*

A repeated measures ANOVA on the laparoscopic simulation task scores, with time difference between tests as a covariate, reveals a significant main effect of stage ($F_{(1,87)} = 10.00$, $p = 0.002$). The time

difference does not significantly affect the scores ($F_{(1,87)} = 0.12$, $p = 0.729$), so it is excluded from subsequent analyses.

There are no significant between-group difference in the first test ($\chi^2_{(4)} = 1.95$, $p = 0.745$) or the second test ($\chi^2_{(4)} = 0.85$, $p = 0.932$) according to Kruskal-Wallis tests. Paired t-tests and Wilcoxon signed-rank tests indicate that participants in Group 3 had significantly lower scores in the second test compared to the first ($t_{(17)} = 3.40$, $p = 0.003$), whereas no significant differences are observed between tests in the other groups.

Figure 7.3a shows the scores for both tests in each group, while the difference in scores between the tests is presented in Figure 7.3b to highlight the individual improvement (negative difference) or deterioration (positive difference) of scores in each group. It can be seen that all groups improved on average, with the largest average improvement seen in Group 3.



(a) (b)

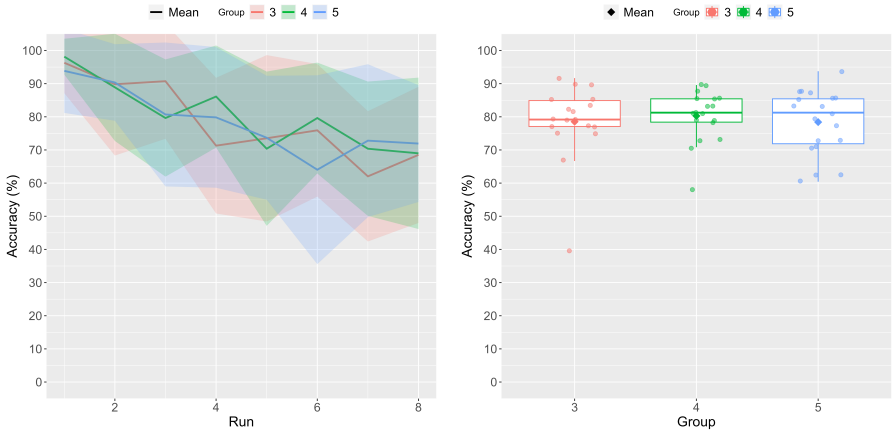
Figure 7.3: Scores of the laparoscopic simulation task. (a) Score in each test. (b) Difference in score between second and first test.Spelling accuracy in the P300 speller. Statistical analysis by Kruskal-Wallis tests and paired t-tests/Wilcoxon signed-rank tests, $0.001 \leqslant p < 0.01$ **.

7.4.2 *Attention Training*

7.4.2.1 *P300 Speller Performance*

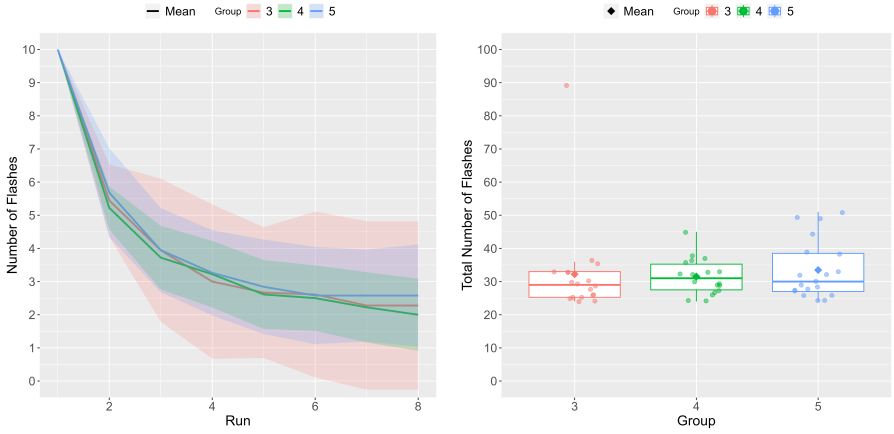
Figure 7.4 and Figure 7.5 show the spelling accuracy and total number of flashes for each group. The data indicate that most participants maintained relatively high spelling accuracy throughout the experiment, even as the number of flashes progressively decreased.

Kruskal-Wallis tests confirm that there are no significant between-group differences in spelling accuracy ($\chi^2_{(2)} = 0.26$, $p = 0.877$) or number of flashes ($\chi^2_{(2)} = 1.35$, $p = 0.508$).



(a) (b)

Figure 7.4: Spelling accuracy in the P300 speller. (a) Mean accuracy (%) in each run, standard deviation illustrated by shading. (b) Mean accuracy (%) over all runs. Statistical analysis by Kruskal-Wallis tests.

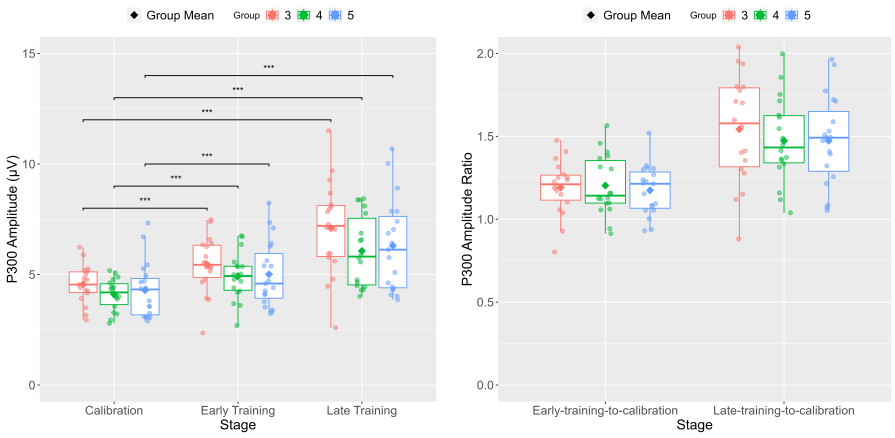


(a) (b)

Figure 7.5: Number of flashes in the P300 speller. (a) Mean number of flashes in each run, standard deviation illustrated by shading. (b) Total number of flashes over all runs. Statistical analysis by Kruskal-Wallis tests.

7.4.2.2 *EEG Signals*

A repeated measures ANOVA on mean P300 amplitude across different training stages reveals a significant main effect of stage ($F_{(2,108)} = 480.99$, $p < 0.001$). A Kruskal-Wallis test confirms that there are no significant between-group differences in any of the three stages. All groups experienced a significant increase from the calibration stage, in both the early (Group 3: $t_{(17)} = -5.23$, $p < 0.001$, Group 4: $t_{(17)} = -4.83$, $p < 0.001$, Group 5: $W = 12$, $p < 0.001$) and late training stage (Group 3: $t_{(17)} = -6.76$, $p < 0.001$, Group 4: $W = 0$, $p < 0.001$, Group 5: $W = 0$, $p < 0.001$). Mean P300 amplitude in each stage, along with the ratios of different stages compared to calibration, are presented in Figure 7.6.

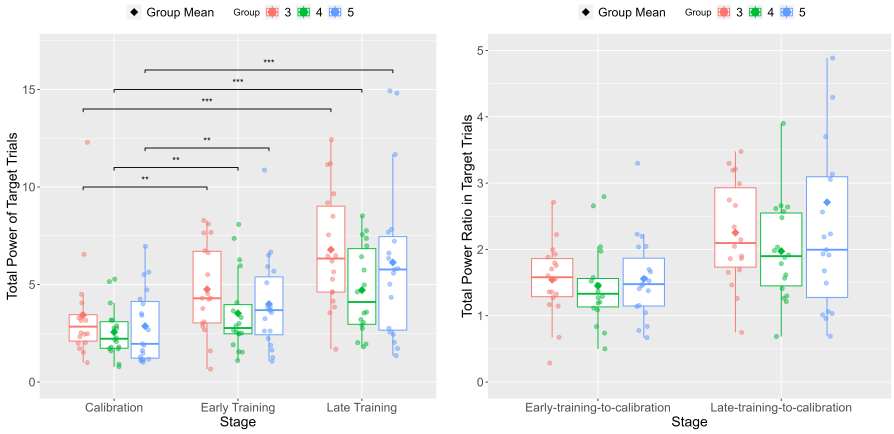


(a) (b)

Figure 7.6: P300 peak-to-peak amplitude. (a) Mean amplitude in each stage. (b) Amplitude ratios of the different stages. Statistical analysis by Kruskal-Wallis tests and paired t-tests/Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.

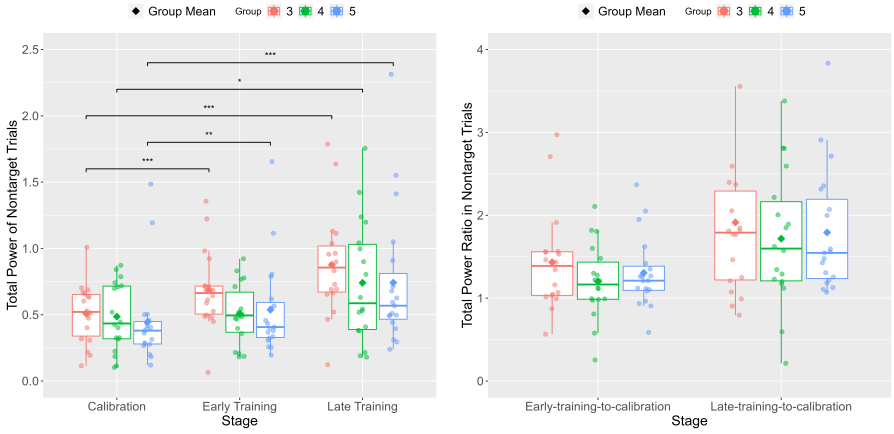
Similarly to the P300 amplitude, the total power in target trials increased significantly from the calibration stage ($F_{(2,108)} = 74.14$, $p < 0.001$) in all groups, in both the early training (Group 3: $W = 27$, $p = 0.008$, Group 4: $W = 25$, $p = 0.007$, Group 5: $W = 21$, $p = 0.002$) and the late training stages (Group 3: $W = 8$, $p < 0.001$, Group 4: $t_{(17)} = -5.03$, $p < 0.001$, Group 5: $W = 8$, $p < 0.001$). The total target power is illustrated in Figure 7.7.

A repeated measures ANOVA test on total power of nontarget trials also reveals a significant main effect of stage ($F_{(2,108)} = 121.23$, $p < 0.001$). A significant increase from calibration to early training stages is seen in Group 3 ($t_{(17)} = -4.07$, $p < 0.001$) and Group 5 ($W = 30$, $p = 0.007$), and in all groups from calibration to late training stages (Group 3: $t_{(17)} = -4.95$, $p < 0.001$, Group 4: $t_{(17)} = -2.55$, $p = 0.021$, Group 5: $W = 0$, $p < 0.001$). Figure 7.8 shows the total nontarget power in the different stages.



(a) (b)

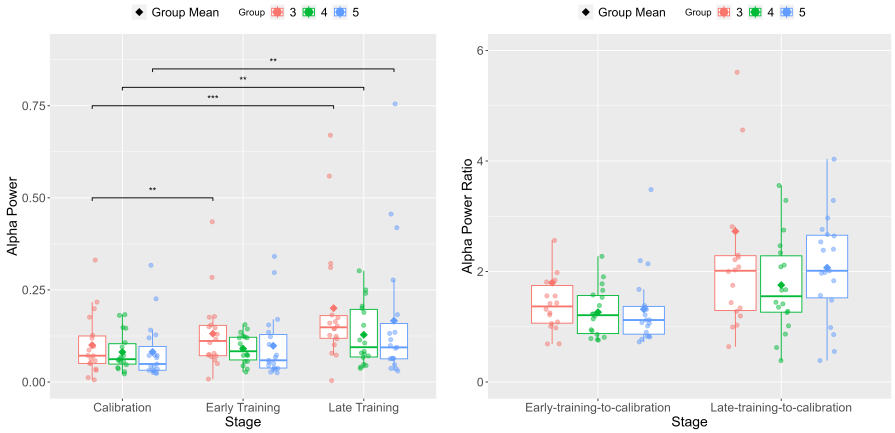
Figure 7.7: Total power in target trials. (a) Mean total power in each stage. (b) Total power ratios of the different stages. Statistical analysis by Kruskal-Wallis tests and Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.



(a) (b)

Figure 7.8: Total power in nontarget trials. (a) Mean total power in each stage. (b) Total power ratios of the different stages. Statistical analysis by Kruskal-Wallis tests and paired t-tests/Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.

A significant effect of stage on alpha power is also observed ($F_{(2,108)} = 235.47$, $p < 0.001$). While all groups experienced a significant increase in alpha power in the late training stage compared to the calibration stage (Group 3: $W = 3$, $p < 0.001$, Group 4: $W = 23$, $p = 0.005$, Group 5: $W = 15$, $p = 0.002$), only Group 3 experienced a significant increase from the calibration to the early training stage ($W = 22$, $p = 0.006$). The alpha power can be seen in Figure 7.9.



(a) (b)

Figure 7.9: Alpha power in nontarget trials. (a) Mean alpha power in each stage. (b) Alpha power ratios of the different stages. Statistical analysis by Kruskal-Wallis tests and paired t-tests/Wilcoxon signed-rank tests, $p < 0.001$ ***, $p < 0.01$ **, $p \leqslant 0.05$ *.

7.4.2.3 *Repeat Neurofeedback Sessions*

The only outcome where significant differences between the first and second NFB session are revealed by a Wilcoxon signed-rank test is mean spelling accuracy ($W = 115$, $p = 0.016$). The mean spelling accuracy achieved by each participant in Group 4 in the two session is illustrated in Figure 7.10.

Figure 7.11 shows the xDAWN weights and ICC for each participant in Group 4. The agreement between weights across sessions is highly variable, with even the highest ICC values indicating only moderate agreement between xDAWN weights across sessions.

7.4.3 *Correlation between Attention Training and Surgical Training*

No significant correlations are found between spelling accuracy, EEG changes, and scores on the laparoscopic simulation task.

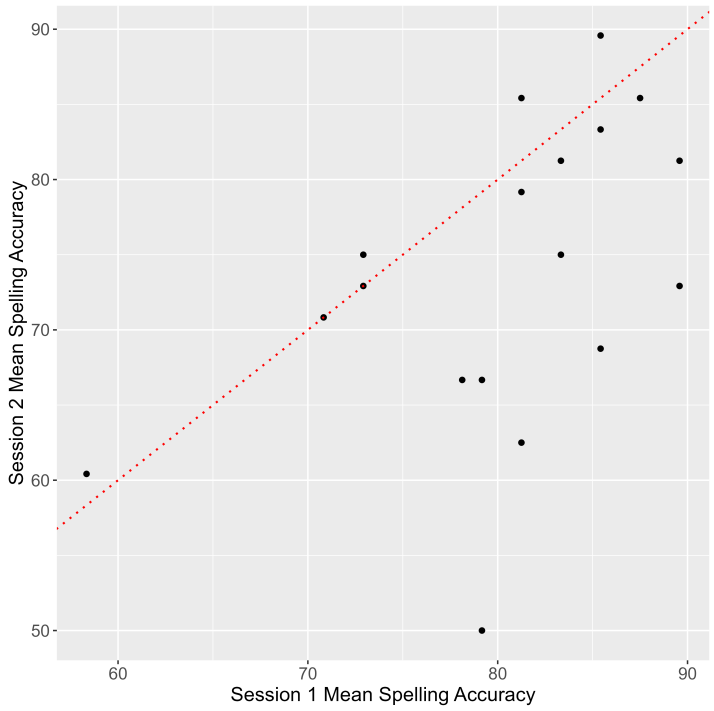


Figure 7.10: Mean spelling accuracy (%) over all runs in both neurofeedback (NFB) training sessions. Red line is line of equality.

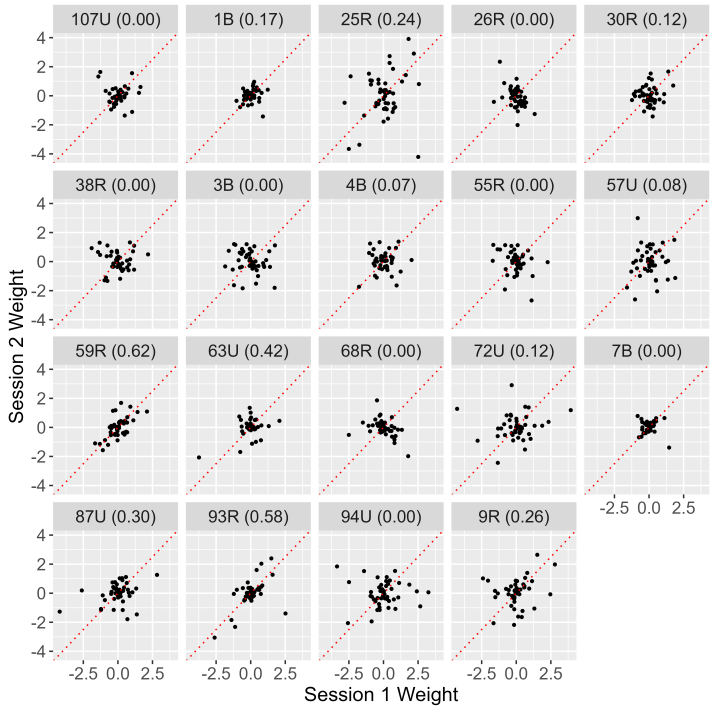


Figure 7.11: xDAWN weights for each participant in both neurofeedback (NFB) training sessions. Red line is line of equality, Intraclass Correlation Coefficient (ICC) is shown in brackets after the participant ID.

7.5 DISCUSSION

This study explores the effects of NFB training and cognitive simulation on surgical skills. However, the results presented here are based on preliminary data, as only a portion of the study data was available at the time of writing. Consequently, these findings may not fully represent the outcomes of the study.

No significant differences in laparoscopy scores are found between groups in either of the two tests. Only Group 3, which received a single NFB training session, showed a significant improvement in the second test. Given that the two other NFB groups did not show similar improvements, the observed effect in Group 3 is likely coincidental rather than a result of the NFB training.

There is also no significant correlation between performance and EEG changes in the attention training and performance in the laparoscopy task. This correlation is explored to determine whether attention training may only enhance surgical skills in participants who performed well, were sufficiently challenged, or showed positive EEG responses to the training. The lack of correlation suggests that this was not the case.

These findings could suggest that the attention improvements resulting from NFB training may not transfer to the skills required for the laparoscopy task. Another possible explanation is that one or two training sessions may be insufficient to produce such transfer effects. While a single session was sufficient to show improvements in the RDM task, as demonstrated in Chapter 5, NFB training is typically conducted over 10 or more sessions spanning several weeks or months to achieve lasting changes [50]. This extended training period might be especially important for facilitating far transfer effects, such as those investigated in this study. Additionally, the laparoscopy task may have been too simple to benefit from the enhanced attentional control fostered by NFB training. Given that this was the easier of the two tasks included in the study, the suturing task, which involves a more detailed assessment of technique, may provide a better measure of the effects of NFB training on surgical skills. The more detailed assessment categories in the suturing task may offer stronger evidence of transfer effects or provide insight into how enhanced attention might influence suturing skills.

Given the study findings, which did not support the hypothesis of transfer effects, it might have been beneficial to include the RDM task alongside the surgical tasks to better understand the relationship between NFB-related attention improvements and surgical skill acquisition. Including the RDM task as an additional measure could have allowed for direct correlations between improvements in visual attention and performance in the surgical task, offering a more comprehensive view of the transferability of training effects.

Interestingly, the analysis of EEG signals reveals that most participants in the NFB groups exhibited a stronger P300 response, as well as higher total power in target trials during the later stages of training. These results are consistent with those observed in Chapter 5 and Section 6.4. However, unlike participants in previous studies, these participants also showed increased alpha power and total power in nontarget trials. This increase in nontarget activity may suggest that the NFB training led to heightened overall brain activity.

Repeat NFB sessions in Group 4 provided an opportunity to analyse performance and EEG changes across sessions. While most metrics remained stable, a slight decrease in spelling accuracy was observed in the second session for some participants. This decline could be attributed to reduced motivation, as the novelty of the experiment may have diminished. Additionally, the timing of the sessions may have played a role, with the first session conducted during the semester and the second session typically occurring during or after exam time, potentially affecting participant focus and engagement.

The stability of xDAWN spatial filter weights across NFB sessions is also examined. In Section 6.2, these weights were used to identify electrode sets, and their stability across sessions could have implications for personalising electrode configurations in long-term training. However, the agreement of xDAWN weights is low for most participants and moderate for a few, suggesting that xDAWN weights may not be reliable for this purpose.

The deployment of the NFB system in this study is notable because it involved a real-world setting with a large number of users: 84 sessions with 65 participants, whose primary interest was surgical training rather than merely participating in an NFB experiment. This real-world application study demonstrates the feasibility of integrating NFB training into professional training environments. By reducing the number of electrodes to 16, the setup time was halved compared to the initial study in Chapter 5, enabling the training of up to 7 participants per day. This scalability is crucial for practical application in real-world settings and could be further enhanced by the use of dry electrodes, which would further reduce setup time and might improve participant comfort.

Not all participants undergoing the NFB training were equally motivated or interested in the training, which likely limited its training effects. Motivation is known to be a critical factor in the success of NFB training, as it directly impacts engagement and the ability to achieve the desired cognitive changes [18]. The observation that motivation varied among participants suggests that incorporating gamification elements and introducing a reward system could enhance participant engagement and motivation. This would likely make the training more effective, especially in real-world settings where maintaining participant interest can be challenging.

A significant advantage of the NFB training over the self-guided cognitive simulation was its structured and supervised nature, which ensured consistent adherence to the study protocol. This consistency allows for more reliable comparisons within the NFB groups. In contrast, the self-guided cognitive simulation lacked direct oversight, making it challenging to verify whether participants were fully engaged or correctly following the prescribed training. This lack of control potentially affects the outcomes and may contribute to the absence of significant improvements in that group.

7.6 SUMMARY

This chapter presents the preliminary results from a study investigating the effects of P300-based NFB training on surgical skills. While these preliminary findings do not support the hypothesis that attention training directly improves surgical skills or their retention, they provide valuable insights into the application of NFB in a real-world setting.

The study successfully demonstrates the feasibility of deploying the NFB training system outside a controlled laboratory environment, involving a large number of users. By reducing the number of electrodes from 32 to 16, the system was able to achieve practical setup times, making it scalable and suitable for integration into professional training environments.

These achievements highlight the system's potential for real-world applications, even though further analysis, particularly of the suturing task, is needed to draw definitive conclusions about the transfer of attention training to surgical skills.

Part IV

CONCLUSIONS

8

CONCLUSIONS

8.1 SUMMARY OF ACHIEVEMENTS

This thesis presents the development and rigorous evaluation of an ERP-based NFB training system for attention enhancement in healthy adults, using a P300 speller and ILC to personalise task difficulty. As discussed in Chapter 2, ERP-based NFB is not widely used, despite showing promising results, largely due to concerns about the lack of well-controlled studies, limited research outside the lab, and the long training time required. The aim of this thesis is to address these concerns and demonstrate the system's effectiveness and potential beyond lab environments. The primary objectives outlined in Chapter 1 are to establish the effectiveness, efficiency, practicality, and applicability of the NFB system. In pursuit of these objectives, this thesis addresses the following research questions:

1. Can ILC enhance the efficiency of NFB training to improve attention in healthy adults?
2. How does the number of electrodes affect the usability, accuracy, and effectiveness of the NFB system?
3. Can attentional improvements gained from NFB training transfer to motor skill learning?

These objectives and research questions guided the research process and are systematically addressed throughout this thesis.

To improve efficiency, an ILC controller is chosen to dynamically adapt task difficulty, providing a straightforward yet effective method for optimising the training process. The controller is designed based on the specific requirements of the training, and is initially tested and compared to existing approaches in simulation. The results demonstrate that the ILC controller can enhance training efficiency by accelerating the NFB training without significantly impacting performance.

These results are then verified in a clinical trial described in Chapter 5, where healthy adults were tested in a cognitive task before and after undergoing the NFB training. Three different task difficulty adaptation approaches are compared. The study shows that P300-based training effectively enhances attention, with all groups demonstrating improved performance in a cognitive task. This supports the use of P300-based NFB training for attention enhancement. Additionally, the group with ILC personalisation completed the training in

the least time, without compromising training efficacy, further supporting the simulation results and demonstrating that using ILC for task difficulty adaptation accelerates training. This directly addresses Research Question 1, confirming that ILC-based adaptation improves training speed and efficiency in ERP-based NFB systems.

To enhance the practicality and usability of the system, additional studies explore the feasibility of reducing the number of electrodes used. It is found that as few as 4 electrodes can achieve good performance in the P300 speller, comparable to 6- and 8-electrode sets, although performance is highly dependent on the environment. However, the positive training effects observed in Chapter 5 could not be replicated with only 4 electrodes, likely due to diminished performance and the increased number of flashes required during training. Consequently, a 16-electrode setup is used in subsequent studies, effectively halving the setup time compared to the initial study outlined in Chapter 5. This finding addresses Research Question 2 by demonstrating that, while reducing the number of electrodes improves usability, a minimum of 16 electrodes is necessary to retain system effectiveness and achieve robust training outcomes.

To demonstrate the system's applicability in real-world settings, the NFB training system is evaluated in a more realistic context: surgical skills training. Preliminary results from this study suggest that the attention improvements observed in Chapter 5 may not directly transfer to surgical skill performance, thereby addressing Research Question 3. However, these conclusions are tentative, as more comprehensive analysis, including potential effects on suturing tasks, is required. Nonetheless, the surgical training study successfully demonstrates the large-scale deployment of the system outside a lab environment.

In summary, this thesis demonstrates that an ERP-based NFB system with task difficulty controlled by ILC can effectively enhance attention in healthy adults. This contributes to the growing body of evidence supporting ERP-based NFB, an area that remains underexplored. By achieving the objectives of demonstrating effectiveness, efficiency, practicality, and real-world applicability, this research lays a foundation for further exploration of NFB training in real-world settings and its potential impact on specific skill transfers.

8.2 FUTURE DIRECTIONS

There are some limitations to the studies conducted in this thesis, which open up future directions for this research:

- **Further System Development:** While the current system demonstrates the benefits of applying ILC for task difficulty adaptation in NFB training, there are opportunities for further refinement. The speller model and simulation developed in Section 4.3.1 were based on a single train-test split from a sin-

gle dataset. Although this was sufficient for tuning and initial testing of the ILC controller, incorporating more datasets, such as those collected in this thesis, and applying cross-validation could improve the model's generalisability and robustness. This would also support the implementation of more advanced control strategies, such as NOILC [110]. The well-defined properties and algorithms associated with ILC provide a rich framework that can be further exploited to optimise this NFB system.

Additionally, incorporating features like early stopping [101] in the speller, or experimenting with different classifiers, could enhance the system efficiency and user experience.

Another area for improvement is the further investigation of electrode configurations, which could significantly enhance the practicality of the training by reducing setup time and increasing user comfort. Based on the results discussed in Chapter 6, the optimal number of electrodes to balance setup time with performance is likely somewhere between 8 and 16 electrodes when using the xDAWN spatial filter. Intermediate configurations between 8 and 16 electrodes were not explored, as EEG caps typically come preconfigured with either 8 or 16 electrodes. The decision to use 16 electrodes was made to fully use the available resources and avoid underusing the equipment. However, future work could investigate intermediate configurations, which may offer a more efficient balance between signal quality and practical considerations like setup time and user comfort. Additionally, exploring other spatial filtering or signal processing methods beyond xDAWN could enable the use of smaller electrode sets while maintaining performance. Alternatively, dry electrodes could be explored to further improve the system. Dry electrodes eliminate the need for electroconductive gels, making the system more user-friendly and easier to deploy in real-world settings. Future research should explore the feasibility of dry electrodes in ERP-based NFB training, including their impact on signal quality, user experience, and overall training outcomes.

- **Further Investigation of Training Transfer Effects:** The preliminary results from the surgical skills study in Chapter 7 suggest limited transfer effects of the attention training to surgical tasks. Given these outcomes, it is crucial to conduct more comprehensive studies to investigate the broader transfer effects of NFB training. Future research should investigate the limits of training effects by testing its impact on a range of different, more realistic, cognitive tasks beyond the RDM task, particularly to better understand its potential in real-world applications.

- **Alternative Stimulus Modalities:** While the current NFB training uses visual stimuli to elicit P300 responses, exploring alternative stimulus modalities, such as auditory stimuli, could offer additional benefits. The auditory P300, for example, might provide a more immersive experience and could be more suitable for some training environments or individuals who respond better to auditory cues. Incorporating multiple modalities, or allowing customisation based on participant preference, could enhance both the effectiveness and engagement of NFB training. Furthermore, combining visual and auditory stimuli may potentially improve attentional control and cognitive processing by leveraging multi-sensory integration [168]. Future research should investigate the impact of these different modalities on training outcomes, as well as on user experience and engagement levels.

- **Cognitive Rehabilitation:** The NFB training in this thesis is solely focused on attention enhancement in healthy adults, meaning that no insights are gained on whether the system could be effective for cognitive rehabilitation in neurodiverse or diseased populations, such as people with dementia or Parkinson's disease. While [93] reports interesting results for the use of P300-based NFB in children with ADHD, other studies investigating P300-based NFB are also limited to healthy adults [94–96]. Therefore, investigating the efficacy of the system for cognitive rehabilitation in these patient populations, rather than focusing solely on cognitive enhancement in healthy individuals, is a major future direction for ERP-based NFB research.

- **Long-term Training:** Apart from the surgical skills study described in Chapter 7, only one session of NFB training was completed, with no long-term follow-up. In the surgical skills study, some participants completed two sessions and a follow-up was conducted between 1 and 7 weeks post-training, which is still relatively short-term. Consequently, no conclusions can be drawn about the efficacy of long-term training and how long the training effects last post-training. Since [95] reports a plateau effect after just three sessions of P300-based NFB, investigating long-term training, and its effects post-training, is an important endeavour. This could be done separately or in conjunction with cognitive rehabilitation of individuals with cognitive deficits. An interesting aspect of long-term training is the possibility of further training personalisation, such as customising electrode sets to allow for fast setup with high classification accuracy. Another potential area for personalisation is tuning the ILC controller, where the maximum step size or penalty function could be adjusted for each individual, and over time.

- **Gamification:** As discussed in Section 5.5.1 and Section 6.4.3.1, participants reported increased boredom after the training. Furthermore, while some participants were competitive and aimed for the best spelling accuracy, others did not care about receiving good or bad feedback, since it was inconsequential. Both boredom and lack of motivation can significantly impact the training efficacy [18]. This is why gamification of the NFB system should be considered in future work. For instance, [95] uses a matching pairs card game, where cards are flipped by focusing on their flashes, to gamify the oddball paradigm. This type of gamification could enhance participant engagement. Additionally, introducing rewards along with feedback could motivate participants to actively seek positive feedback.

Part V

APPENDIX

A

STATISTICAL METHODS

This appendix provides an overview of the statistical methods used throughout this thesis. These methods are grouped according to their function, and justifications are provided for why specific tests were chosen based on the characteristics of the data (e.g. normality, repeated measures, and the number of groups).

A.1 COMPARING GROUP MEANS

A.1.1 *Two Groups*

The starting point for comparing the means of two independent groups was Student's t-test, which tests the null hypothesis that the means of the two groups are equal, taking into account the variances of the groups. This test assumes that the data are normally distributed [169].

When the assumption of normality was not met, the non-parametric Wilcoxon rank-sum test (also known as the Mann-Whitney U test) was used. The Wilcoxon test compares the ranks of the data to assess whether one group tends to have higher or lower values than the other, making it appropriate for non-normal distributions [169].

A.1.2 *Three or More Groups*

When comparing three or more independent groups, ANOVA was used to assess whether there were any significant differences between the group means. ANOVA is an extension of the Student's t-test and assumes normally distributed data [169].

When the assumption of normality was violated, the Kruskal-Wallis test was used as a non-parametric alternative to ANOVA. Similar to the Wilcoxon rank-sum test, the Kruskal-Wallis test compares the ranks of data across groups rather than their actual values [169].

When significant differences were detected by ANOVA or the Kruskal-Wallis test, post-hoc pairwise comparisons were conducted to determine which specific group means were significantly different. For ANOVA, Tukey's Honest Significant Difference (HSD) test, commonly referred to as the Tukey test, was used [169]. For the Kruskal-Wallis test, the non-parametric Tukey-Kramer-Nemenyi test was performed for pairwise comparisons [170].

A.2 COMPARING REPEATED MEASURES

A.2.1 *Two Measures*

For repeated measures, such as pre- and post-training measures in the same participant, the paired t-test was applied, provided the data met normality assumptions. The paired t-test tests the null hypothesis that the mean difference between paired observations is zero [169].

If the data were not normally distributed, the Wilcoxon signed-rank test (the non-parametric alternative to the paired t-test) was used. This test ranks the differences between paired observations and is appropriate for handling non-normal data in repeated measures [169].

A.2.2 *Three or More Measures*

For repeated measures data with more than two measures, repeated measures ANOVA was used to assess significant differences between time points or conditions within the same group. This test extends the paired t-test to accommodate more than two conditions [171].

When normality assumptions were violated, the Friedman test served as a non-parametric alternative to repeated measures ANOVA. The Friedman test ranks each time point individually within subjects, making it appropriate for non-normal data in repeated-measures designs involving a single group [170].

A.3 COMPARING REPEATED MEASURES ACROSS GROUPS

When repeated measures were compared across groups, repeated measures ANOVA was employed to examine both main effects and interactions. This analysis determines whether differences between time points or conditions are consistent across groups, thus accounting for both within-group and between-group variations [171].

For data that were non-normally distributed, a non-parametric approach was taken by performing repeated measures ANOVA on ranked data [172]. Since the Friedman test does not extend to multi-group comparisons, this ranked ANOVA served as an alternative to capture interaction effects across groups.

A.4 COMPARING GROUP MEANS WITH BASELINE ADJUSTMENT

When comparisons needed to account for both group differences and an additional categorical variable, such as baseline measurements, two-way ANOVA was used. This approach tested for interactions between the two factors, group and baseline, while allowing for comparisons of main effects across both factors. The two-way ANOVA

assumes normally distributed data within each group and baseline category [169].

When normality assumptions were not met, the non-parametric alternative, ART ANOVA, was used. ART ANOVA preserves main effects and interactions by ranking data within each factor level, making it a suitable choice for factorial designs with non-normal data distributions [173].

A.5 EQUIVALENCE TESTING

Equivalence testing was used to determine whether two groups were statistically equivalent within a predefined margin. This was done using the TOST procedure. The null hypothesis in equivalence testing is the opposite of traditional tests: it posits that the means of the two groups are not equivalent, i.e. the difference between the two group means is greater than a pre-specified equivalence margin. The alternative hypothesis is that the difference between the group means falls within the equivalence margin, indicating that the groups are statistically equivalent [151].

For data meeting normality assumptions, the t-TOST procedure was applied, which uses the Student's t-test described above. For non-normally distributed data, the Wilcoxon-TOST method was used as a non-parametric alternative, applying the Wilcoxon rank-sum test to assess equivalence [151].

A.6 EFFECT SIZE ESTIMATION

To complement hypothesis testing, effect sizes were sometimes reported to quantify the magnitude of differences observed between groups. For parametric tests, Cohen's D was used to measure the effect size, expressing the difference between two means in terms of standard deviations [152]. This provided insight into the practical significance of results beyond p-values. For non-parametric tests, Spearman's correlation coefficient was used as an alternative to Cohen's D, to estimate the effect size based on ranked data [153].

A.7 CORRELATION AND ASSOCIATION

When exploring relationships between two continuous variables, Pearson's correlation coefficient was applied for normally distributed data to measure the strength and direction of the linear relationship between the variables [169].

For non-normally distributed data, Spearman's correlation coefficient was used, which assesses monotonic relationships based on ranked data rather than actual values. This non-parametric measure

is appropriate when the relationship between variables may not be strictly linear but still follows a consistent pattern [169].

In addition, the ICC was used to assess reliability in measurements, particularly for evaluating test-retest stability over time. ICC is valuable for determining consistency in repeated measurements or assessing inter-rater reliability in subjective evaluations [167].

B

SUMMARY OF KEY RESULTS FROM CHAPTER 5

Table B.1: Summary of Key Results.

Outcome	Statistical Test	Result	Implication
Pre-training boredom scores	Kruskal-Wallis	Higher in random difficulty group than benchmark group ($p = 0.03$)	Baseline boredom differs between groups, impacting spelling accuracy (as shown in sensitivity analysis).
Pre-training eye fatigue scores	Kruskal-Wallis	Higher in random difficulty group than ILC and benchmark groups ($p = 0.004$, $p = 0.018$)	Baseline eye fatigue differs between groups, impacting perceived mental and physical demand, as well as P300 amplitude (as shown in sensitivity analysis).
Post-training tiredness increase	Paired t-test, Wilcoxon signed-rank	Significant increase in all groups ($p < 0.05$)	Training induced significant fatigue, as expected for a challenging task.
Post-training eye fatigue increase	Paired t-test, Wilcoxon signed-rank	Significant increase in ILC and benchmark groups ($p \leqslant 0.001$)	Training induced significant eye strain, likely due to extended screen exposure.
Post-training boredom increase	Wilcoxon signed-rank	Significant increase in ILC group ($p = 0.029$)	Training may become boring for some participants.

Table B.1: Summary of Key Results (Continued).

Outcome	Statistical Test	Result	Implication
Perceived Physical Demand	Kruskal-Wallis	Higher in random difficulty group than benchmark group ($p = 0.024$)	Perceived physical demand may reflect mental fatigue from training.
RDM task accuracy	Paired t-test, Wilcoxon signed-rank	Significant increase in all groups post-training ($p < 0.05$)	All adaptation methods improve RDM task accuracy, supporting training efficacy.
RDM task score	Paired t-test, Wilcoxon signed-rank	Significant increase in ILC group only ($p = 0.015$)	Only the ILC group achieved a significant improvement in score (accuracy/RT), suggesting enhanced training efficacy.
Training length	Kruskal-Wallis	Lower in ILC group than benchmark and random difficulty groups ($p = 0.007$, $p < 0.001$)	ILC controller significantly accelerates training.
P300 amplitude	One-way ANOVA	Higher in ILC group compared to random difficulty group during training ($p = 0.044$)	Supports hypothesis that reduced flashes drive attentional improvement.

Table B.1: Summary of Key Results (Continued).

Outcome	Statistical Test	Result	Implication
Total power (post-training)	-	Higher in target trials and lower in nontarget trials compared to baseline in ILC and benchmark groups; lower in both target and nontarget trials in random difficulty group	Indicates improved focus and reduced distractibility with personalised task difficulty.
Alpha power (post-training)	-	Lower compared to baseline in ILC and benchmark groups; higher in random difficulty group	Suggests more focused brain state post-training with personalised task difficulty.
RDM accuracy and spelling accuracy	Pearson's correlation	Negative correlation in benchmark group ($p = 0.01$)	Higher spelling accuracy correlates with lesser RDM improvement, implying task may not be challenging enough for high performers.
RDM RT and alpha power	Spearman's correlation	Positive correlation in ILC group ($p = 0.032$)	Indicates more focused participants had faster reaction times.

BIBLIOGRAPHY

[1] United Nations Department of Economic and Social Affairs, Population Division, "World Population Prospects 2022: Summary of Results," 2022.

[2] X. Li, X. Feng, X. Sun, N. Hou, F. Han, and Y. Liu, "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019," *Front. Aging Neurosci.*, vol. 14, p. 937 486, 2022. DOI: 10.3389/FNAGI.2022.937486.

[3] Z. Ou, J. Pan, S. Tang, D. Duan, D. Yu, H. Nong, and Z. Wang, "Global Trends in the Incidence, Prevalence, and Years Lived With Disability of Parkinson's Disease in 204 Countries/Territories From 1990 to 2019," *Front. Public Heal.*, vol. 9, p. 776 847, 2021. DOI: 10.3389/FPUBH.2021.776847.

[4] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C. E. Teunissen, J. Cummings, and W. M. van der Flier, "Alzheimer's disease," *Lancet*, vol. 397, no. 10284, pp. 1577–1590, 2021. DOI: 10.1016/S0140-6736(20)32205-4.

[5] C. Sun and M. J. Armstrong, "Treatment of Parkinson's Disease with Cognitive Impairment: Current Approaches and Future Directions," *Behav. Sci.*, vol. 11, no. 4, p. 54, 2021. DOI: 10.3390/BS11040054.

[6] T. J. Krivanek, S. A. Gale, B. M. McFeeley, C. M. Nicastri, and K. R. Daffner, "Promoting Successful Cognitive Aging: A Ten-Year Update," *J. Alzheimer's Dis.*, vol. 81, no. 3, pp. 871–920, 2021. DOI: 10.3233/JAD-201462.

[7] E. M. Arenaza-Urquiljo *et al.*, "Whitepaper: Defining and investigating cognitive reserve, brain reserve and brain maintenance Reserve, Resilience and Protective Factors PIA Empirical Definitions and Conceptual Frameworks Workgroup HHS Public Access," *Alzheimer's Dement.*, vol. 16, no. 9, pp. 1305–1311, 2020. DOI: 10.1016/j.jalz.2018.07.219.

[8] N. Gates and M. Valenzuela, "Cognitive exercise and its role in cognitive function in older adults," *Curr. Psychiatry Rep.*, vol. 12, no. 1, pp. 20–27, 2010. DOI: 10.1007/S11920-009-0085-Y.

[9] E. Kalbe, D. Aarsland, and A. K. Folkerts, "Cognitive Interventions in Parkinson's Disease: Where We Want to Go within 20 Years," *J. Parkinsons. Dis.*, vol. 8, no. s1, S107–S113, 2018. DOI: 10.3233/JPD-181473.

[10] M. J. Anderson, A. J. DeMeireles, D. P. Trofa, D. Kovacevic, C. S. Ahmad, and T. S. Lynch, "Cognitive Training in Orthopaedic Surgery," *J. Am. Acad. Orthop. Surg. Glob. Res. Rev.*, vol. 5, no. 3, e21.00021, 2021. DOI: 10.5435/JAAOSGLOBAL-D-21-00021.

[11] K. J. Blacker, J. Hamilton, G. Roush, K. A. Pettijohn, and A. T. Biggs, "Cognitive Training for Military Application: a Review of the Literature and Practical Guide," *J. Cogn. Enhanc.*, vol. 3, no. 1, pp. 30–51, 2019. DOI: 10.1007/S41465-018-0076-1.

[12] M. Slimani, N. L. Bragazzi, D. Tod, A. Dellal, O. Hue, F. Cheour, L. Taylor, and K. Chamari, "Do cognitive training strategies improve motor and positive psychological skills development in soccer players? Insights from a systematic review," *J. Sports Sci.*, vol. 34, no. 24, pp. 2338–2349, 2016. DOI: 10.1080/02640414.2016.1254809.

[13] N. Omejc, B. Rojc, P. P. Battaglini, and U. Marusic, "Review of the therapeutic neurofeedback method using electroencephalography: EEG Neurofeedback," *Bosn. J. basic Med. Sci.*, vol. 19, no. 3, pp. 213–220, 2019. DOI: 10.17305/BJBMS.2018.3785.

[14] J. C. Da Silva and M. L. De Souza, "Neurofeedback Training for Cognitive Performance Improvement in Healthy Subjects: A Systematic Review," *Psychol. Neurosci.*, vol. 14, no. 3, pp. 262–279, 2021. DOI: 10.1037/PNE0000261.

[15] R. T. Thibault, M. Lifshitz, N. Birbaumer, and A. Raz, "Neurofeedback, Self-Regulation, and Brain Imaging: Clinical Science and Fad in the Service of Mental Disorders," *Psychother. Psychosom.*, vol. 84, no. 4, pp. 193–207, 2015. DOI: 10.1159/000371714.

[16] W. K. Norris, M. K. Allison, S. Fisher, and G. M. Curran, "Implementation Science Application to EEG Neurofeedback Research: A Call to Action," *NeuroRegulation*, vol. 11, no. 2, p. 211, 2024. DOI: 10.15540/NR.11.2.211.

[17] C. Simon, D. A. Bolton, N. C. Kennedy, S. R. Soekadar, and K. L. Ruddy, "Challenges and Opportunities for the Future of Brain-Computer Interface in Neurorehabilitation," *Front. Neurosci.*, vol. 15, p. 699428, 2021. DOI: 10.3389/FNINS.2021.699428.

[18] R. Sitaram *et al.*, "Closed-loop brain training: the science of neurofeedback," *Nat. Rev. Neurosci.*, vol. 18, no. 2, pp. 86–100, 2017. DOI: 10.1038/NRN.2016.164.

[19] J. Kamiya, "The First Communications About Operant Conditioning of the EEG," *J. Neurother.*, vol. 15, no. 1, pp. 65–73, 2011. DOI: 10.1080/10874208.2011.545764.

[20] M. B. Sterman, "Basic concepts and clinical findings in the treatment of seizure disorders with EEG operant conditioning," *Clin. Electroencephalogr.*, vol. 31, no. 1, pp. 45–55, 2000. DOI: 10.1177/155005940003100111.

[21] R. Onagawa, Y. Muraoka, N. Hagura, and M. Takemi, "An investigation of the effectiveness of neurofeedback training on motor performance in healthy adults: A systematic review and meta-analysis," *Neuroimage*, vol. 270, p. 120 000, 2023. DOI: 10.1016/J.NEUROIMAGE.2023.120000.

[22] E. Niedermeyer and F. Lopes Da Silva, *Electroencephalography : Basic Principles, Clinical Applications, and Related Fields*. Wolters Kluwer, 2004.

[23] J. N. Acharya, A. Hani, J. Cheek, P. Thirumala, and T. N. Tsuchida, "American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature," *J. Clin. Neurophysiol.*, vol. 33, no. 4, pp. 308–311, 2016. DOI: 10.1097/wnp.0000000000000316.

[24] A. S. Malik and H. U. Amin, "Designing an EEG Experiment," in *Designing EEG Experiments for Studying the Brain*, Academic Press, 2017, ch. 1, pp. 1–30. DOI: 10.1016/B978-0-12-811140-6.00001-1.

[25] H. Gevensleben, B. Albrecht, H. Lütcke, T. Auer, W. I. Dewiputri, R. Schweizer, G. Moll, H. Heinrich, and A. Rothenberger, "Neurofeedback of slow cortical potentials: Neural mechanisms and feasibility of a placebo-controlled design in healthy adults," *Front. Hum. Neurosci.*, vol. 8, p. 990, 2014. DOI: 10.3389/FNHUM.2014.00990.

[26] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. MIT Press, 2014.

[27] A. Bryniarska, J. A. Ramos, and M. Fernández, "Machine Learning Classification of Event-Related Brain Potentials during a Visual Go/NoGo Task," *Entropy*, vol. 26, no. 3, p. 220, 2024. DOI: 10.3390/E26030220.

[28] H. U. Amin, R. Ullah, M. F. Reza, and A. S. Malik, "Single-trial extraction of event-related potentials (ERPs) and classification of visual stimuli by ensemble use of discrete wavelet transform with Huffman coding and machine learning techniques," *J. Neuroeng. Rehabil.*, vol. 20, no. 1, p. 70, 2023. DOI: 10.1186/S12984-023-01179-8.

[29] S. Van Voorhis and S. A. Hillyard, "Visual evoked potentials and selective attention to points in space," *Percept. Psychophys.*, vol. 22, no. 1, pp. 54–62, 1977. DOI: 10.3758/BF03206080.

[30] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, "Electrophysiological Studies of Face Perception in Humans," *J. Cogn. Neurosci.*, vol. 8, no. 6, pp. 551–565, 1996. DOI: 10.1162/JOCN.1996.8.6.551.

[31] S. J. LUCK and S. A. HILLYARD, "Electrophysiological correlates of feature analysis during visual search," *Psychophysiology*, vol. 31, no. 3, pp. 291–308, 1994. DOI: 10.1111/J.1469-8986.1994.TB02218.X.

[32] A. Pfefferbaum, J. M. Ford, B. J. Weller, and B. S. Kopell, "ERPs to response production and inhibition," *Electroencephalogr. Clin. Neurophysiol.*, vol. 60, no. 5, pp. 423–434, 1985. DOI: 10.1016/0013-4694(85)91017-X.

[33] S. Sutton, M. Braren, J. Zubin, and E. R. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, 1965. DOI: 10.1126/SCIENCE.150.3700.1187.

[34] D. Regan, "Some characteristics of average steady-state and transient responses evoked by modulated light," *Electroencephalogr. Clin. Neurophysiol.*, vol. 20, no. 3, pp. 238–248, 1966. DOI: 10.1016/0013-4694(66)90088-5.

[35] G. Bin, X. Gao, Y. Wang, B. Hong, and S. Gao, "VEP-based brain-computer interfaces: Time, frequency, and code modulations," *IEEE Comput. Intell. Mag.*, vol. 4, no. 4, pp. 22–26, 2009. DOI: 10.1109/MCI.2009.934562.

[36] M. Kuba, Z. Kubová, J. Kremláček, and J. Langrová, "Motion-onset VEPs: Characteristics, methods, and diagnostic use," *Vision Res.*, vol. 47, no. 2, pp. 189–202, 2007. DOI: 10.1016/J.VISRES.2006.09.020.

[37] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "ERP components on reaction errors and their functional significance: a tutorial," *Biol. Psychol.*, vol. 51, no. 2-3, pp. 87–107, 2000. DOI: 10.1016/S0301-0511(99)00031-9.

[38] M. W. Eysenck and M. T. Keane, "Attention and Performance," in *Cognitive Psychology: a Student's Handbook*, 8th ed., Taylor and Francis Group, 2020, ch. 5, pp. 178–234.

[39] R. M. Klein and M. A. Lawrence, "On the Modes and Domains of Attention," in *Cognitive Neuroscience of Attention*, M. I. Posner, Ed., Guilford Publications, 2011, ch. 2, pp. 11–28.

[40] A. Johnson, R. W. Proctor, E. De Haan, and R. Kessels, "Disorders of Attention," in *Attenion: Theory and Practice*, Sage Publications, 2004, ch. 12, pp. 367–395.

[41] A. M. Re and A. Capodieci, "Signs and Symptoms," in *Understanding ADHD: A Guide to Symptoms, Management and Treatment*, Taylor & Francis Group, 2020, ch. 1, pp. 1–17.

[42] A. Ridderinkhof, E. I. de Bruin, S. van den Driesschen, and S. M. Bögels, "Attention in Children With Autism Spectrum Disorder and the Effects of a Mindfulness-Based Program," *J. Atten. Disord.*, vol. 24, no. 5, pp. 681–692, 2020. DOI: 10.1177/1087054718797428.

[43] F. Bernard, J. M. Lemée, A. Ter Minassian, and P. Menei, "Right Hemisphere Cognitive Functions: From Clinical and Anatomic Bases to Brain Mapping During Awake Craniotomy Part I: Clinical and Functional Anatomy," *World Neurosurg.*, vol. 118, pp. 348–359, 2018. DOI: 10.1016/J.WNEU.2018.05.024.

[44] S. Vossel, J. J. Geng, and G. R. Fink, "Dorsal and Ventral Attention Systems: Distinct Neural Circuits but Collaborative Roles," *Neurosci.*, vol. 20, no. 2, pp. 150–159, 2014. DOI: 10.1177/1073858413494269.

[45] E. Magosso, F. De Crescenzio, G. Ricci, S. Piastra, and M. Ursino, "EEG Alpha Power Is Modulated by Attentional Changes during Cognitive Tasks and Virtual Reality Immersion," *Comput. Intell. Neurosci.*, vol. 2019, p. 7051079, 2019. DOI: 10.1155/2019/7051079.

[46] J. Polich, "Updating P300: An Integrative Theory of P3a and P3b," *Clin. Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, 2007. DOI: 10.1016/j.clinph.2007.04.019.

[47] A. Bluschke, E. Eggert, J. Friedrich, R. Jamous, A. Prochnow, C. Pscherer, M. L. Schreiter, B. Teufert, V. Roessner, and C. Beste, "The Effects of Different Theta and Beta Neurofeedback Training Protocols on Cognitive Control in ADHD," *J. Cogn. Enhanc.*, vol. 6, no. 4, pp. 463–477, 2022. DOI: 10.1007/S41465-022-00255-6.

[48] K. Benchenane, P. H. Tiesinga, and F. P. Battaglia, "Oscillations in the prefrontal cortex: a gateway to memory and attention," *Curr. Opin. Neurobiol.*, vol. 21, no. 3, pp. 475–485, 2011. DOI: 10.1016/J.CONB.2011.01.004.

[49] C. S. Herrmann and R. T. Knight, "Mechanisms of human attention: event-related potentials and oscillations," *Neurosci. Biobehav. Rev.*, vol. 25, no. 6, pp. 465–476, 2001. DOI: 10.1016/S0149-7634(01)00027-6.

[50] H. Marzbani, H. R. Marateb, and M. Mansourian, "Neurofeedback: A Comprehensive Review on System Design, Methodology and Clinical Applications," *Basic Clin. Neurosci.*, vol. 7, no. 2, pp. 143–158, 2016. DOI: 10.15412/J.BCN.03070208.

[51] Neurofeedback Collaborative Group, "Double-Blind Placebo-Controlled Randomized Clinical Trial of Neurofeedback for Attention-Deficit/Hyperactivity Disorder With 13-Month

Follow-up.," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 60, no. 7, pp. 841–855, 2021. DOI: 10.1016/j.jaac.2020.07.906.

[52] Neurofeedback Collaborative Group, "Neurofeedback for Attention-Deficit/Hyperactivity Disorder: 25-Month Follow-up of Double-Blind Randomized Controlled Trial.," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 62, no. 4, pp. 435–446, 2023. DOI: 10.1016/j.jaac.2022.07.862.

[53] M. Ryoo and C. Son, "Effects of Neurofeekback Training on EEG, Continuous Performance Task (CPT), and ADHD Symptoms in ADHD-prone College Students.," *J. Korean Acad. Nurs.*, vol. 45, no. 6, pp. 928–938, 2015. DOI: 10.4040/jkan.2015.45.6.928.

[54] A. R. Bakhshayesh, S. Hänsch, A. Wyschkon, M. J. Rezai, and G. Esser, "Neurofeedback in ADHD: a single-blind randomized controlled trial.," *Eur. Child Adolesc. Psychiatry*, vol. 20, no. 9, pp. 481–491, 2011. DOI: 10.1007/s00787-011-0208-y.

[55] S. Jirayucharoensak, P. Israsena, S. Pan-Ngum, S. Hemrungrojn, and M. Maes, "A game-based neurofeedback training system to enhance cognitive performance in healthy elderly subjects and in patients with amnestic mild cognitive impairment," *Clin. Interv. Aging*, vol. 14, pp. 347–360, 2019. DOI: 10.2147/CIA.S189047.

[56] J.-H. Jang, J. Kim, G. Park, H. Kim, E.-S. Jung, J.-Y. Cha, C.-Y. Kim, S. Kim, J.-H. Lee, and H. Yoo, "Beta wave enhancement neurofeedback improves cognitive functions in patients with mild cognitive impairment: A preliminary pilot study.," *Medicine (Baltimore).*, vol. 98, no. 50, e18357, 2019. DOI: 10.1097/MD.0000000000018357.

[57] Y.-S. Lee, S.-H. Bae, S.-H. Lee, and K.-Y. Kim, "Neurofeedback training improves the dual-task performance ability in stroke patients.," *Tohoku J. Exp. Med.*, vol. 236, no. 1, pp. 81–88, 2015. DOI: 10.1620/tjem.236.81.

[58] Z. Guleken, G. Eskikurt, and S. Karamürsel, "Investigation of the effects of transcranial direct current stimulation and neurofeedback by continuous performance test.," *Neurosci. Lett.*, vol. 716, p. 134648, 2020. DOI: 10.1016/j.neulet.2019.134648.

[59] S. Kober, M. Witte, M. Stangl, A. Väljamäe, C. Neuper, and G. Wood, "Shutting down sensorimotor interference unblocks the networks for stimulus processing: An SMR neurofeedback training study," *Clin. Neurophysiol.*, vol. 126, no. 1, pp. 82–95, 2015. DOI: 10.1016/j.clinph.2014.03.031.

[60] T Egner and J. H. Gruzelier, "EEG biofeedback of low beta band components: frequency-specific effects on variables of attention and event-related brain potentials.," *Clin. Neurophysiol.*, vol. 115, no. 1, pp. 131–139, 2004. DOI: 10.1016/s1388-2457(03)00353-5.

[61] L. Morales-Quezada, D. Martinez, M. M. El-Hagrassy, T. J. Kaptchuk, M. B. Sterman, and G. Y. Yeh, "Neurofeedback impacts cognition and quality of life in pediatric focal epilepsy: An exploratory randomized double-blinded sham-controlled trial," *Epilepsy Behav.*, vol. 101, p. 106 570, 2019. DOI: https://doi.org/10.1016/j.yebeh.2019.106570.

[62] E.-J. Lee and C.-H. Jung, "Additive effects of neurofeedback on the treatment of ADHD: A randomized controlled study.," *Asian J. Psychiatr.*, vol. 25, pp. 16–21, 2017. DOI: 10.1016/j.ajp.2016.09.002.

[63] D. Mahmood, H. Nisar, and C. Y. Tsai, "Exploring the efficacy of neurofeedback training in modulating alpha-frequency band and its effects on functional connectivity and band power," *Expert Syst. Appl.*, vol. 254, p. 124 415, 2024. DOI: 10.1016/J.ESWA.2024.124415.

[64] A. M. Berger and E. J. Davelaar, "Frontal Alpha Oscillations and Attentional Control: A Virtual Reality Neurofeedback Study," *Neuroscience*, vol. 378, pp. 189–197, 2018. DOI: 10.1016/J.NEUROSCIENCE.2017.06.007.

[65] M. Navarro Gil, C. Escolano Marco, J. Montero-Marín, J. Minguez Zafra, E. Shonin, and J. García Campayo, "Efficacy of Neurofeedback on the Increase of Mindfulness-Related Capacities in Healthy Individuals: a Controlled Trial," *Mindfulness*, vol. 9, no. 1, pp. 303–311, 2018. DOI: 10.1007/S12671-017-0775-1.

[66] J. Mishra, M. Lowenstein, R. Campusano, Y. Hu, J. Diaz-Delgado, J. Ayyoub, R. Jain, and A. Gazzaley, "Closed-Loop Neurofeedback of α Synchrony during Goal-Directed Attention," *J. Neurosci.*, vol. 41, no. 26, pp. 5699–5710, 2021. DOI: 10.1523/JNEUROSCI.3235-20.2021.

[67] T. Ros, J. Théberge, P. A. Frewen, R. Kluetsch, M. Densmore, V. D. Calhoun, and R. A. Lanius, "Mind over chatter: Plastic up-regulation of the fMRI salience network directly after EEG neurofeedback," *Neuroimage*, vol. 65, pp. 324–335, 2013. DOI: 10.1016/J.NEUROIMAGE.2012.09.046.

[68] M. P. Deiber, C. Ammann, R. Hasler, J. Colin, N. Perroud, and T. Ros, "Electrophysiological correlates of improved executive function following EEG neurofeedback in adult attention deficit hyperactivity disorder," *Clin.*

Neurophysiol., vol. 132, no. 8, pp. 1937–1946, 2021. DOI: 10.1016/J.CLINPH.2021.05.017.

[69] Z. Hao, C. He, Y. Ziqian, L. Haotian, and L. Xiaoli, "Neurofeedback training for children with ADHD using individual beta rhythm," *Cogn. Neurodyn.*, vol. 16, no. 6, pp. 1323–1333, 2022. DOI: 10.1007/S11571-022-09798-Y.

[70] J. Bielas and Ł. Michalczyk, "Beta Neurofeedback Training Improves Attentional Control in the Elderly.," *Psychol. Rep.*, vol. 124, no. 1, pp. 54–69, 2021. DOI: 10.1177/0033294119900348.

[71] Z. Yuan *et al.*, "Effect of BCI-Controlled Pedaling Training System With Multiple Modalities of Feedback on Motor and Cognitive Function Rehabilitation of Early Subacute Stroke Patients," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2569–2577, 2021. DOI: 10.1109/TNSRE.2021.3132944.

[72] A. M. Neuhäußer, A. Bluschke, V. Roessner, and C. Beste, "Distinct effects of different neurofeedback protocols on the neural mechanisms of response inhibition in ADHD," *Clin. Neurophysiol.*, vol. 153, pp. 111–122, 2023. DOI: https://doi.org/10.1016/j.clinph.2023.06.014.

[73] J.-R. Wang and S. Hsieh, "Neurofeedback training improves attention and working memory performance.," *Clin. Neurophysiol.*, vol. 124, no. 12, pp. 2406–2420, 2013. DOI: 10.1016/j.clinph.2013.05.020.

[74] K. Xiong, M. Wan, D. Cai, and W. Nan, "Down-regulation of theta amplitude through neurofeedback improves executive control network efficiency in healthy children.," *Int. J. Psychophysiol.*, vol. 197, p. 112 301, 2024. DOI: 10.1016/j.ijpsycho.2024.112301.

[75] 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," *Appl. Psychophysiol. Biofeedback*, vol. 39, no. 3-4, pp. 193–202, 2014. DOI: 10.1007/S10484-014-9257-6.

[76] J. Ghaziri, A. Tucholka, V. Larue, M. Blanchette-Sylvestre, G. Reyburn, G. Gilbert, J. Lévesque, and M. Beauregard, "Neurofeedback Training Induces Changes in White and Gray Matter," *Clin. EEG Neurosci.*, vol. 44, no. 4, pp. 265–272, 2013. DOI: 10.1177/1550059413476031.

[77] F. H. Wang, L. Y. Sun, X. M. Cui, H. D. Zhao, L. F. Yang, Z. Wang, and T. K. Shi, "Comparative efficacy of targeted structural patterns of electroencephalography neurofeedback in children with inattentive or combined attention deficit hy-

peractivity disorder," *Brain Behav.*, vol. 12, no. 6, e2572, 2022. DOI: 10.1002/BRB3.2572.

[78] L. Li, L. Yang, C.-j. Zhuo, and Y.-F. Wang, "A randomised controlled trial of combined EEG feedback and methylphenidate therapy for the treatment of ADHD.," *Swiss Med. Wkly.*, vol. 143, w13838, 2013. DOI: 10.4414/smw.2013.13838.

[79] R. T. Thibault and A. Raz, "The psychology of neurofeedback: Clinical intervention even if applied placebo," *Am. Psychol.*, vol. 72, no. 7, pp. 679–688, 2017. DOI: 10.1037/AMP0000118.

[80] T. Ros *et al.*, "Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist)," *Brain*, vol. 143, no. 6, pp. 1674–1685, 2020. DOI: 10.1093/BRAIN/AWAA009.

[81] O. M. Bazanova, T. Auer, and E. A. Sapina, "On the Efficiency of Individualized Theta/Beta Ratio Neurofeedback Combined with Forehead EMG Training in ADHD Children," *Front. Hum. Neurosci.*, vol. 12, no. 3, 2018. DOI: 10.3389/FNHUM.2018.00003.

[82] U. Strehl, "Slow Cortical Potentials Neurofeedback," *J. Neurother.*, vol. 13, no. 2, pp. 117–126, 2009. DOI: 10/btn5q2.

[83] S. Baumeister, I. Wolf, S. Hohmann, N. Holz, R. Boecker-Schlier, T. Banaschewski, and D. Brandeis, "The impact of successful learning of self-regulation on reward processing in children with ADHD using fMRI," *Atten. Defic. Hyperact. Disord.*, vol. 11, no. 1, pp. 31–45, 2019. DOI: 10.1007/S12402-018-0269-6.

[84] K. Mayer, S. N. Wyckoff, U. Schulz, and U. Strehl, "Neurofeedback for Adult Attention-Deficit/Hyperactivity Disorder: Investigation of Slow Cortical Potential Neurofeedback — Preliminary Results," *J. Neurother.*, vol. 16, no. 1, pp. 37–45, 2012. DOI: 10.1080/10874208.2012.650113.

[85] B. Barth, K. Mayer-Carius, U. Strehl, S. N. Wyckoff, F. B. Haeussinger, A. J. Fallgatter, and A. C. Ehlis, "A randomized-controlled neurofeedback trial in adult attention-deficit/hyperactivity disorder," *Sci. Rep.*, vol. 11, no. 1, p. 16 873, 2021. DOI: 10.1038/S41598-021-95928-1.

[86] P. M. Aggensteiner *et al.*, "Slow cortical potentials neurofeedback in children with ADHD: comorbidity, self-regulation and clinical outcomes 6 months after treatment in a multicenter randomized controlled trial," *Eur. Child Adolesc. Psychiatry*, vol. 28, no. 8, pp. 1087–1095, 2019. DOI: 10.1007/S00787-018-01271-8.

[87] J. Hasslinger, S. Bölte, and U. Jonsson, "Slow Cortical Potential Versus Live Z-score Neurofeedback in Children and Adolescents with ADHD: A Multi-arm Pragmatic Randomized Controlled Trial with Active and Passive Comparators," *Res. child Adolesc. Psychopathol.*, vol. 50, no. 4, pp. 447–462, 2022. DOI: 10.1007/S10802-021-00858-1.

[88] S. Wangler, H. Gevensleben, B. Albrecht, P. Studer, A. Rothenberger, G. H. Moll, and H. Heinrich, "Neurofeedback in children with ADHD: specific event-related potential findings of a randomized controlled trial.," *Clin. Neurophysiol.*, vol. 122, no. 5, pp. 942–950, 2011. DOI: 10.1016/j.clinph.2010.06.036.

[89] T. Mehta, N. Mannem, N. K. Yarasi, and P. C. Bollu, "Biomarkers for ADHD: the Present and Future Directions," *Curr. Dev. Disord. Reports*, vol. 7, pp. 85–92, 2020. DOI: 10.1007/S40474-020-00196-9.

[90] K. Rieger, M. H. Rarra, L. Diaz Hernandez, D. Hubl, and T. Koenig, "Neurofeedback-Based Enhancement of Single-Trial Auditory Evoked Potentials: Treatment of Auditory Verbal Hallucinations in Schizophrenia," *Clin. EEG Neurosci.*, vol. 49, no. 6, pp. 367–378, 2018. DOI: 10.1177/1550059418765810.

[91] M. Musso, D. Hübner, S. Schwarzkopf, M. Bernodusson, P. Levan, C. Weiller, and M. Tangermann, "Aphasia recovery by language training using a brain–computer interface: a proof-of-concept study," *Brain Commun.*, vol. 4, no. 1, fcac008, 2022. DOI: 10.1093/BRAINCOMMS/FCAC008.

[92] G. Pei *et al.*, "Enhancing Working Memory Based on Mismatch Negativity Neurofeedback in Subjective Cognitive Decline Patients: A Preliminary Study," *Front. Aging Neurosci.*, vol. 29, no. 12, p. 263, 2020. DOI: 10.3389/FNAGI.2020.00263.

[93] M. Fouillen, "P300-based Brain-Computer Interfaces for attention training in children with ADHD," Ph.D. dissertation, 2019.

[94] X. J. Li, L. M. Tang, Z. L. Zhang, J. Wu, and Q. Li, "Attention and Memory Training System Based on Neural Feedback," in *2022 3rd Int. Conf. Comput. Vision, Image Deep Learn. Int. Conf. Comput. Eng. Appl. CVIDL ICCEA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 794–798. DOI: 10.1109/CVIDLICCEA56201.2022.9824137.

[95] J. D. Jacoby, "Event-related potentials as a form of neurofeedback using low-cost hardware," Ph.D. dissertation, 2016.

[96] M. Arvaneh, I. H. Robertson, and T. E. Ward, "A P300-Based Brain-Computer Interface for Improving Attention," *Frontiers in Human Neuroscience*, vol. 12, no. 524, 2019. DOI: 10.3389/FNHUM.2018.00524.

[97] T. Nierhaus, C. Vidaurre, C. Sannelli, K. R. Mueller, and A. Villringer, "Immediate brain plasticity after one hour of brain–computer interface (BCI)," *J. Physiol.*, vol. 599, no. 9, pp. 2435–2451, 2021. DOI: 10.1113/JP278118.

[98] M. Arvaneh, T. E. Ward, and I. H. Robertson, "Effects of feedback latency on P300-based brain-computer interface," in *2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, IEEE, 2015, pp. 2315–2318. DOI: 10.1109/EMBC.2015.7318856.

[99] T. W. P. Janssen, M. Bink, W. D. Weeda, K. Geladé, R. van Mourik, A. Maras, and J. Oosterlaan, "Learning curves of theta/ beta neurofeedback in children with ADHD.," *Eur. Child Adolesc. Psychiatry*, vol. 26, no. 5, pp. 573–582, 2017. DOI: 10.1007/s00787-016-0920-8.

[100] Z. Gu, Z. Chen, J. Zhang, X. Zhang, and Z. L. Yu, "An Online Interactive Paradigm for P300 Brain-Computer Interface Speller," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 2, pp. 152–161, 2019. DOI: 10.1109/TNSRE.2019.2892967.

[101] L. Bianchi, C. Liti, and V. Piccialli, "A New Early Stopping Method for P300 Spellers," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1635–1643, 2019. DOI: 10.1109/TNSRE.2019.2924080.

[102] J. Jin, B. Z. Allison, E. W. Sellers, C. Brunner, P. Horki, X. Wang, and C. Neuper, "An adaptive P300-based control system," *J. Neural Eng.*, vol. 8, no. 3, p. 036 006, 2011. DOI: 10.1088/1741-2560/8/3/036006.

[103] A. N. Antle, E. S. McLaren, H. Fiedler, and N. Johnson, "Evaluating the impact of a mobile neurofeedback app for young children at school and home," in *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, 2019, p. 36. DOI: 10.1145/3290605.3300266.

[104] J. C. Whitehead, R. Neeman, and G. M. Doniger, "Preliminary Real-World Evidence Supporting the Efficacy of a Remote Neurofeedback System in Improving Mental Health: Retrospective Single-Group Pretest-Posttest Study.," *JMIR Form. Res.*, vol. 6, no. 7, e35636, 2022. DOI: 10.2196/35636.

[105] M. Uchiyama, "Formation of High-Speed Motion Pattern of a Mechanical Arm by Trial," *Transaction of the Society of Instrument and Control Engineers*, vol. 14, pp. 706–712, 1978. DOI: 10.9746/sicetr1965.14.706.

[106] S. Arimoto, S. Kawamura, and F. Miyazaki, "Bettering operation of Robots by learning," *Journal of Robotic Systems*, vol. 1, no. 2, pp. 123–140, 1984. DOI: 10.1002/rob.4620010203.

[107] H. S. Ahn, Y. Q. Chen, and K. L. Moore, "Iterative learning control: Brief survey and categorization," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 37, no. 6, pp. 1099–1121, 2007. DOI: 10.1109/TSMCC.2007.905759.

[108] L. Xia, Y. Feng, L. Zheng, C. Wang, and X. Wu, "Development of an adaptive iterative learning controller with sensorless force estimator for the hip-type exoskeleton," in *IEEE Int. Conf. Robot. Biomimetics, ROBIO 2019*, 2019, pp. 2516–2521. DOI: 10.1109/ROBIO49542.2019.8961508.

[109] C. T. Freeman, A. M. Hughes, J. H. Burridge, P. H. Chappell, P. L. Lewin, and E. Rogers, "Iterative learning control of FES applied to the upper extremity for rehabilitation," *Control Eng. Pract.*, vol. 17, no. 3, pp. 368–381, 2009. DOI: 10.1016/J.CONENGPRAC.2008.08.003.

[110] D. H. Owens and J. Hätönen, "Iterative learning control - An optimization paradigm," *Annual Reviews in Control*, vol. 29, no. 1, pp. 57–70, 2005. DOI: 10.1016/j.arcontrol.2005.01.003.

[111] D. H. Owens, *Iterative Learning Control - An Optimization Paradigm* (Advances in Industrial Control). Springer London, 2016. DOI: 10.1007/978-1-4471-6772-3.

[112] D. Shen and X. Li, "A survey on iterative learning control with randomly varying trial lengths: Model, synthesis, and convergence analysis," *Annual Reviews in Control*, vol. 48, pp. 89–102, 2019. DOI: 10.1016/j.arcontrol.2019.10.003.

[113] T. D. Son, G. Pipeleers, and J. Swevers, "Optimal iterative learning control design with trial-varying initial conditions," in *2013 European Control Conference*, 2013, pp. 1181–1186. DOI: 10.23919/ecc.2013.6669535.

[114] X. D. Li, T. F. Xiao, and H. X. Zheng, "Adaptive discrete-time iterative learning control for non-linear multiple input multiple output systems with iteration-varying initial error and reference trajectory," *IET Control Theory and Applications*, vol. 5, no. 9, pp. 1131–1139, 2011. DOI: 10.1049/iet-cta.2010.0379.

[115] J Hätönen, T. J. Harte, D. H. Owens, J Ratcliffe, P Lewin, and E Rogers, "Discrete-Time Arimoto ILC-Algorithm Revisited," *IFAC Proceedings Volumes*, vol. 37, no. 12, pp. 541–546, 2004. DOI: 10.1016/S1474-6670(17)31525-2.

[116] Z. Zhang and Q. Zou, "Data-driven robust iterative learning control of linear systems," *Automatica*, vol. 164, p. 111 646, 2024. DOI: 10.1016/J.AUTOMATICA.2024.111646.

[117] R. Chi, D. Wang, Z. Hou, and S. Jin, "Data-driven optimal terminal iterative learning control," *J. Process Control*, vol. 22, no. 10, pp. 2026–2037, 2012. DOI: 10.1016/J.JPROCONT.2012.08.001.

[118] T. Meng and W. He, "Iterative Learning Control of a Robotic Arm Experiment Platform with Input Constraint," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 664–672, 2018. DOI: 10.1109/TIE.2017.2719598.

[119] X. Jin, "Fault-tolerant iterative learning control for mobile robots non-repetitive trajectory tracking with output constraints," *Automatica*, vol. 94, pp. 63–71, 2018. DOI: 10.1016/J.AUTOMATICA.2018.04.011.

[120] Y. Chen, B. Chu, and C. T. Freeman, "Iterative Learning Control for Robotic Path Following With Trial-Varying Motion Profiles," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 6, pp. 4697–4706, 2022. DOI: 10.1109/TMECH.2022.3164101.

[121] J. Lu, Z. Cao, R. Zhang, and F. Gao, "Nonlinear Monotonically Convergent Iterative Learning Control for Batch Processes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5826–5836, 2018. DOI: 10.1109/TIE.2017.2782201.

[122] T. Liu and F. Gao, "Robust two-dimensional iterative learning control for batch processes with state delay and time-varying uncertainties," *Chem. Eng. Sci.*, vol. 65, no. 23, pp. 6134–6144, 2010. DOI: 10.1016/J.CES.2010.08.031.

[123] Q. Zhu, F. Song, J. X. Xu, and Y. Liu, "An Internal Model Based Iterative Learning Control for Wafer Scanner Systems," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 5, pp. 2073–2084, 2019. DOI: 10.1109/TMECH.2019.2929565.

[124] H. Wei, F. Viti, and C. M. Tampere, "An iterative learning approach for signal control in urban traffic networks," in *16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC 2013)*, 2013. DOI: 10.1109/ITSC.2013.6728275.

[125] T. Lan, F. Yan, and H. Lin, "Iterative Learning Control with Forgetting Factor for Urban Road Network," *J. Control Sci. Eng.*, vol. 2017, p. 9 269 187, 2017. DOI: 10.1155/2017/9269187.

[126] L. Van Nguyen and D. H. Dao, "Iterative Learning Control for Autonomous Driving Vehicles," in *2019 Int. Conf. Adv. Technol. Commun.*, IEEE Computer Society, 2019. DOI: 10.1109/ATC.2019.8924520.

[127] H. Chen, Z. Xiong, and Y. Ji, "Iterative learning control for automatic train operation with discrete gears," in *2019 IEEE 8th Data Driven Control Learn. Syst. Conf.*, 2019. DOI: 10.1109/DDCLS.2019.8909057.

[128] D. Huang, Y. Chen, D. Meng, and P. Sun, "Adaptive iterative learning control for high-speed train: A multi-agent approach," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 7, pp. 4067–4077, 2021. DOI: 10.1109/TSMC.2019.2931289.

[129] C. L. Lynch and M. R. Popovic, "Functional Electrical Stimulation," *IEEE Control Syst. Mag.*, vol. 28, no. 2, pp. 40–50, 2008. DOI: 10.1109/MCS.2007.914689.

[130] P. Sampson, C. Freeman, S. Coote, S. Demain, P. Feys, K. Meadmore, and A. M. Hughes, "Using Functional Electrical Stimulation Mediated by Iterative Learning Control and Robotics to Improve Arm Movement for People with Multiple Sclerosis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 2, pp. 235–248, 2016. DOI: 10.1109/TNSRE.2015.2413906.

[131] S. Sa-e, C. T. Freeman, and K. Yang, "Iterative learning control of functional electrical stimulation in the presence of voluntary user effort," *Control Eng. Pract.*, vol. 96, p. 104 303, 2020. DOI: 10.1016/J.CONENGPRAC.2020.104303.

[132] T. Seel, C. Werner, J. Raisch, and T. Schauer, "Iterative learning control of a drop foot neuroprosthesis — Generating physiological foot motion in paretic gait by automatic feedback control," *Control Eng. Pract.*, vol. 48, pp. 87–97, 2016. DOI: 10.1016/J.CONENGPRAC.2015.11.007.

[133] W. Guan, L. Zhou, and Y. S. Cao, "Joint Motion Control for Lower Limb Rehabilitation Based on Iterative Learning Control (ILC) Algorithm," *Complexity*, vol. 2021, p. 6 651 495, 2021. DOI: 10.1155/2021/6651495.

[134] W. Ting and S. Aiguo, "An Adaptive Iterative Learning Based Impedance Control for Robot-Aided Upper-Limb Passive Rehabilitation," *Front. Robot. AI*, vol. 6, 2019. DOI: 10.3389/frobt.2019.00041.

[135] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988. DOI: 10.1016/0013-4694(88)90149-6.

[136] Y. Renard, F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand, and A. Lécuyer, "OpenViBE: An open-source software platform to design, test, and use brain-computer interfaces in real and virtual environments," *Presence Teleoperators Virtual Environ.*, vol. 19, no. 1, pp. 35–53, 2010. DOI: 10.1162/PRES.19.1.35.

[137] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN Algorithm to Enhance Evoked Potentials: Application to Brain-Computer Interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, 2009. DOI: 10.1109/tbme.2009.2012869.

[138] R. O. Duda, P. E. Hart, and D. G. Stork, "Linear Discriminant Functions," in *Pattern Classification*, 2nd ed., New York: Wiley, 2001, pp. 215–281.

[139] C. Ledesma-Ramirez, E. Bojorges-Valdez, O. Janez-Suarez, C. Saavedra, L. Bougrain, and G. Gentiletti, "An Open-Access P300 Speller Database," in *4th Int. BCI Meet.*, Pacific Grove, CA, USA, 2010.

[140] M. Lövdén, L. Bäckman, U. Lindenberger, S. Schaefer, and F. Schmiedek, "A Theoretical Framework for the Study of Adult Cognitive Plasticity," *Psychol. Bull.*, vol. 136, no. 4, pp. 659–676, 2010. DOI: 10.1037/a0020080.

[141] S. Noble, E. Woods, T. Ward, and J. Ringwood, "Adaptive p300-based brain-computer interface for attention training: Protocol for a randomized controlled trial," *JMIR Research Protocols*, vol. 12, no. e46135, 2023. DOI: 10.2196/46135.

[142] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Adv. Psychol.*, vol. 52, pp. 139–183, 1988. DOI: 10.1016/S0166-4115(08)62386-9.

[143] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, 2019. DOI: 10.3758/S13428-018-01193-Y.

[144] ANT Neuro, *eego (TM) rt*. [Online]. Available: https://www.ant-neuro.com/products/eego_rt (visited on 05/06/2024).

[145] ANT Neuro, *waveguard (TM) original*. [Online]. Available: https://www.ant-neuro.com/products/waveguard_original (visited on 05/06/2024).

[146] C. C. von Bastian and A. Eschen, "Does working memory training have to be adaptive?" *Psychol. Res.*, vol. 80, no. 2, pp. 181–194, 2016. DOI: 10.1007/S00426-015-0655-Z.

[147] J. J. Foxe and A. C. Snyder, "The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention," *Front. Psychol.*, vol. 2, no. 154, 2011. DOI: 10.3389/FPSYG.2011.00154.

[148] K. A. Colwell, D. B. Ryan, C. S. Throckmorton, E. W. Sellers, and L. M. Collins, "Channel selection methods for the P300 Speller," *J. Neurosci. Methods*, vol. 232, pp. 6–15, 2014. DOI: 10.1016/J.JNEUMETH.2014.04.009.

[149] V. Bhandari, N. D. Londhe, and G. B. Kshirsagar, "A Systematic Review of Computational Intelligence Techniques for Channel Selection in P300-Based Brain Computer Interface Speller," *Artif. Intell. Appl.*, vol. 2, no. 3, pp. 169–178, 2024. DOI: 10.47852/BONVIEWAIA42021390.

[150] W. Speier, A. Deshpande, and N. Pouratian, "A method for optimizing EEG electrode number and configuration for signal acquisition in P300 speller systems," *Clin. Neurophysiol.*, vol. 126, no. 6, pp. 1171–1177, 2015. DOI: 10.1016/J.CLINPH.2014.09.021.

[151] D. Lakens, A. M. Scheel, and P. M. Isager, "Equivalence Testing for Psychological Research: A Tutorial," *Adv. Methods Pract. Psychol. Sci.*, vol. 1, no. 2, pp. 259–269, 2018. DOI: 10.1177/2515245918770963.

[152] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988.

[153] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited. An overview of some recommended measures of effect size," *Trends Sport. Sci.*, vol. 1, no. 21, 2014.

[154] L. Wallace, N. Raison, F. Ghumman, A. Moran, P. Dasgupta, and K. Ahmed, "Cognitive training: How can it be adapted for surgical education?" *Surg.*, vol. 15, no. 4, pp. 231–239, 2017. DOI: 10.1016/J.SURGE.2016.08.003.

[155] J. H. Song, "The role of attention in motor control and learning," *Curr. Opin. Psychol.*, vol. 29, pp. 261–265, 2019. DOI: 10.1016/J.COPSYC.2019.08.002.

[156] K. E. Hsu, F. Y. Man, R. A. Gizicki, L. S. Feldman, and G. M. Fried, "Experienced surgeons can do more than one thing at a time: Effect of distraction on performance of a simple laparoscopic and cognitive task by experienced and novice surgeons," *Surg. Endosc.*, vol. 22, no. 1, pp. 196–201, 2008. DOI: 10.1007/S00464-007-9452-0.

[157] A. T. Meneghetti, G. Pachev, B. Zheng, O. N. Panton, and K. Qayumi, "Objective assessment of laparoscopic skills: dual-task approach," *Surg. Innov.*, vol. 19, no. 4, pp. 452–459, 2012. DOI: 10.1177/1553350611430673.

[158] M. A. Ghazanfar, M. Cook, B. Tang, I. Tait, and A. Alijani, "The effect of divided attention on novices and experts in laparoscopic task performance," *Surg. Endosc.*, vol. 29, no. 3, pp. 614–619, 2015. DOI: 10.1007/S00464-014-3708-2.

[159] G. Wulf and R. Lewthwaite, "Optimizing performance through intrinsic motivation and attention for learning: The OPTIMAL theory of motor learning," *Psychon. Bull. Rev.*, vol. 23, no. 5, pp. 1382–1414, 2016. DOI: 10.3758/S13423-015-0999-9.

[160] L. K. Chua, J. Jimenez-Diaz, R. Lewthwaite, T. Kim, and G. Wulf, "Superiority of external attentional focus for motor performance and learning: Systematic reviews and meta-analyses.," *Psychol. Bull.*, vol. 147, no. 6, pp. 618–645, 2021. DOI: 10.1037/BUL0000335.

[161] Limbs & Things LTD, *Fls trainer box with tv camera*. [Online]. Available: https://fls-products.com/fls/products/50302/50302-fls-trainer-box-with-tv-camera (visited on 10/06/2024).

[162] M. C. Vassiliou, G. A. Ghitulescu, L. S. Feldman, D. Stanbridge, K. Leffondré, H. H. Sigman, and G. M. Fried, "The MISTELS program to measure technical skill in laparoscopic surgery: Evidence for reliability," *Surg. Endosc. Other Interv. Tech.*, vol. 20, no. 5, pp. 744–747, 2006. DOI: 10.1007/S00464-005-3008-Y.

[163] H. M. Mentis, A. Chellali, K. Manser, C. G. Cao, and S. D. Schwaitzberg, "A Systematic Review of the Effect of Distraction on Surgeon Performance: Directions for Operating Room Policy and Surgical Training," *Surg. Endosc.*, vol. 30, no. 5, pp. 1713–1724, 2016. DOI: 10.1007/S00464-015-4443-Z.

[164] M. Jeannerod, "Neural Simulation of Action: A Unifying Mechanism for Motor Cognition," *Neuroimage*, vol. 14, no. 1, S103–S109, 2001. DOI: 10.1006/NIMG.2001.0832.

[165] J. Munzert, B. Lorey, and K. Zentgraf, "Cognitive motor processes: The role of motor imagery in the study of motor representations," *Brain Res. Rev.*, vol. 60, no. 2, pp. 306–326, 2009. DOI: 10.1016/J.BRAINRESREV.2008.12.024.

[166] J. Cragg, F. Mushtaq, N. Lal, A. Garnham, M. Hallissey, T. Graham, and U. Shiralkar, "Surgical cognitive simulation improves real-world surgical performance: randomized study," *BJS Open*, vol. 5, no. 3, zrab003, 2021. DOI: 10.1093/BJSOPEN/ZRAB003.

[167] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016. DOI: 10.1016/J.JCM.2016.02.012.

[168] M. Marucci, G. Di Flumeri, G. Borghini, N. Sciaraffa, M. Scandola, E. F. Pavone, F. Babiloni, V. Betti, and P. Aricò, "The impact of multisensory integration and perceptual load in virtual reality settings on performance, workload and presence," *Sci.*

Rep., vol. 11, no. 1, p. 4831, 2021. DOI: 10.1038/s41598-021-84196-8.

[169] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability & statistics for engineers & scientists*, 9th. Pearson, 2011.

[170] D. J. Denis, *Univariate, Bivariate, and Multivariate Statistics Using R: Quantitative Tools for Data Analysis and Data Science*. John Wiley & Sons, Inc, 2020.

[171] L. N. Muhammad, "Guidelines for repeated measures statistical analysis approaches with basic science research considerations," *J. Clin. Invest.*, vol. 133, no. 11, e171058, 2023. DOI: 10.1172/JCI171058.

[172] W. J. Conover and R. L. Iman, "Rank transformations as a bridge between parametric and nonparametric statistics," *Am. Stat.*, vol. 35, no. 3, pp. 124–129, 1981. DOI: 10.1080/00031305.1981.10479327.

[173] H. Mansouri, R. L. Paige, and J. G. Surles, "Aligned Rank Transform Techniques for Analysis of Variance and Multiple Comparisons," *Commun. Stat. - Theory Methods*, vol. 33, no. 9, pp. 2217–2232, 2004. DOI: 10.1081/STA-200026599.