EEG workload prediction in a closed-loop learning environment

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Carina Benigna Walter
aus Mainz

Tübingen
2015

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	16.10.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Martin Bogdan
2. Berichterstatter:	Prof. Dr. Wolfgang Rosenstiel

Acknowledgments

First, I want to gratefully thank Prof. Dr. Wolfgang Rosenstiel for funding and supervising my thesis and providing me with the opportunity to work in his department on this exciting project as part of the ScienceCampus Tübingen.

I especially want to thank my supervisor Prof. Dr. Martin Bogdan for encouraging my research and for allowing me to grow as a research scientist. His support and advice guided me through this work and I greatly appreciated it.

A very special thanks to my colleagues at the Department of Computer Engineering and Embedded Systems in particular the Neuroteam for their support throughout this time:
A special thanks to Anita Dieckgraeff for the pleasant atmosphere in our office and the many delicious homemade cakes as comfort food.
Many thanks to Dr. Martin Spüler, Dr. Armin Walter and Markus Dobler for our time together. I am thankful for their patience, motivation, enthusiasm, as well as for the shared laughs and frustrations that made the daily work so much easier.
Thanks to the TIES colleagues, especially Dustin Peterson, Simon Schulz and Dr. Thomas Schweizer for organizing sport competitions in tabletop soccer and bowling. These events helped to relax after a long working day.
I would like to thank Werner Dreher for the good time in the Basispraktikum and for his care to eat on time - especially white chocolate mousse.
Thanks to Margot Reimold who managed all administrative issues while remaining in a good mood and who was even calm in stressful situations.

I would like to thank the people at the Knowledge Media Research Center, in particular Prof. Dr. Peter Gerjets, Dr. Gabriele Cierniak and Christian Scharinger, for the excellent and intensive collaboration. I am grateful for the enlightening discussions and their immense knowledge in cognitive load.

A special thanks to all people who participated in the experiments, without them this work would not have been possible.

Thanks to the members of the ScienceCampus Tübingen and the LEAD- Graduate School for supporting excellent interdisciplinary research.

A grateful thanks to my parents Gabriele and Gerhard Walter for putting their trust in me and encouraging me during my whole studies. Thanks to Dennis and Miriam Walter, since every once in a while they pulled me out of research towards the other important things in life. A special thanks to Thomas Nestmeyer for his care and patience, as well as his unconditional support during writing my thesis. Without my family I would not have been able to get through this time.

Abstract

The issues of developing an online EEG-based adaptive learning environment are examined in this thesis. The aim is to adapt instructional learning material in real-time, to support learners in their individual learning process and keep them in their optimal workload capacity range during learning. First, suitable learning material is designed, which does not cause artifacts and induces confounds in the EEG data. Second, the most suitable features for an online workload detection in EEG data are determined, by using a variety of pre-processing and feature selection methods, as connectivity and independent component analysis. Third, generalizable classification methods like cross-task classification and cross-subject regression are developed, to enable a workload prediction across a variety of tasks, independently from subjects. In an offline analysis, the cross-subject regression leads to a higher workload prediction accuracy as the cross-task classification. Since the workload prediction across subjects is more precise, this method is used for the subsequent online study. Therefore, the achieved findings and developed classification methods will finally be applied in an online study. The difficulty level of the presented learning material is adapted in real-time, dependent on the predicted workload of each subject. Furthermore, the applicability and efficiency of an online EEG-based adaptive learning environment is investigated and assessed. Comparing the EEG-based learning environment with an error-adaptive learning system, which is state of the art, the induced learning effects are similar. Thus, the learners can successfully be supported in their individual learning process using an EEG-based adaptation of the learning material, by keeping them in their optimal workload range for learning.

Zusammenfassung

In dieser Arbeit werden alle notwendigen Schritte zur erfolgreichen Entwicklung einer EEG-basierten, adaptiven Lernumgebung behandelt. Das Ziel ist es, unterschiedliche Arbeitsgedächtnisbelastung in Echtzeit detektieren und vorhersagen zu können. Abhängig von der Belastungsvorhersage soll das präsentierte Lernmaterial an die individuellen Bedürfnisse eines Lernenden angepasst werden, um ihn/sie während des Lernens in seinem/ihrem optimalen Arbeitsgedächtnisbereich zu halten. So können Lernende mit speziellen Bedürfnissen und unterschiedlichen Vorkenntnissen individuell in ihrer Lerngeschwindigkeit gefördert werden. Zunächst wurde dafür geeignetes Lernmaterial erstellt, das wenig Artefakte erzeugt und möglichst geringe Störungen im EEG-Signal hervorruft. Weiter wurden verschiedene Vorverarbeitungs- und Merkmalsextrationsschritte vergleichend angewendet, um die geeignetsten Features zur Vorhersage der Arbeitsgedächtnisbelastung zu bestimmen. Hierfür wurden unter anderem Konnektivitäts- und Independent-Component-Analysen angewendet.
Für eine effektive und anwendbare Lernumgebung, basierend auf EEG-Signalen, sind generalisierbare Klassifikatoren notwendig. Daher beschäftigt sich ein großer Teil dieser Arbeit mit der Entwicklung von generalisierten Klassifikatoren, wie Cross-Task-Klassifikation oder Cross-Subject-Regression. Diese Methoden ermöglichen die Vorhersage der Arbeitsgedächtnisbelastung über verschiedene Aufgaben und Probanden hinweg. Zunächst werden diese beiden Methoden in Offline-Analysen getestet. Da die Cross-Subject-Regression im Offline-Fall sehr viel genauere Vorhersagewerte liefert, als die Cross-Task-Klassifikation, wird diese am Ende der Arbeit in einer Online-Studie angewendet. In der Online-Studie ist es mit Hilfe der Cross-Subject-Regression möglich, die Arbeitsgedächtnisbelastung in Echtzeit vorherzusagen und basierend darauf, das präsentierte Lernmaterial individuell für den Lernenden anzupassen. Weiter werden der Lerneffekt und die Praktikabilität einer EEG-basierten Lernumgebung untersucht und mit einer Fehler-adaptierten Lernumgebung, die zur Zeit dem aktuellen Stand der Technik entspricht, verglichen. Die induzierten Lerneffekte beider Lernumgebungen sind vergleichbar. Dies ist der Grund, weshalb davon ausgegangen werden kann, dass eine individuelle Förderung von Lernenden mit Hilfe ihrer EEG-Signale möglich und vielversprechend ist. Unter Benutzung von EEG-Signalen kann das Lernmaterial so an die Bedürfnisse eines einzelnen Lernenden angepasst werden, dass sie konstant in ihrem optimalen Arbeitsgedächtnisbereich gehalten werden und eine optimale Unterstützung in ihrem Lernprozess erfahren.

Contents

List of Abbreviations

Acc	Accuracy
Ag	Silver
AgCl	Silver chloride
ANN	Artificial Neural Network
AR	Articulatory Rehearsal
ARM	Autoregressive Model
AUC	Area Under the Curve
BA	Brodmann Areas
BCI	Brain-Computer Interface
CAR	Common Average Reference
CC	Correlation Coefficient
CE	Central Executive
EEG	Electroencephalography
EOG	Electrooculogram
ERD	Event-related Desynchronization
ERS	Event-related Synchronization
FFT	Fast Fourier Transformation
fMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GD	Global Deviation
IC	Independent Components

ICA	Independent Component Analysis
IPS	Intra-Parietal Sulcus
IS	Inner Scribe
kNN	k-Nearest Neighbor
LDA	Linear Discriminant Analysis
LMM	Linear Mixed Models
MEG	Magnetoencephalography
MSE	Mean Squared Error
NIRS	Near Infrared Spectroscopy
PS	Phonological Store
RBF	Radial Basis Function
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
SDIC	System of Differential and Integral Calculus
SNR	Signal to Noise Ratio
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VC	Visual Cache

1 Introduction

Imagine you are a teacher standing in front of a class with 30 students and trying to teach them different types of angle theorems. Everyone will agree: a perfect support of each student in such a situation is nearly impossible. Each subject has its own learning speed, depending on prior knowledge, cognitive capacity and the type of presented learning material. Furthermore, a variety of learning types and learners with special needs, e.g., test anxiety, learning disability or attention deficit disorder occur. So in a classroom with 30 students, 30 different states of working memory capacity as well as types of motivation can be measured. The type and amount of workload during learning is crucial for successful learning and should be held in the optimal working memory capacity range for each learner. When individuals work below their capabilities, they get bored and lose focus more easily. Furthermore, when students are overwhelmed, they are not capable of keeping up the demands. Both scenarios slow down the potential improvement in skills. Therefore, instructions and learning material should be presented in a format that keeps a learner engaged and motivated without overwhelming the limited workload resources. As one person - the teacher - is not able to support each individual's need, it seems advisable to provide individual technological support for learners, which adapts instructional materials and learning tasks to their level of expertise, mental state and working memory capacity, to support them in their learning process. Mainly learners with special needs will benefit from such an individualist system.

1.1 Motivation for electroencephalography-based adaptive learning environments

Until now, learning environments are mainly based on a learner's performance and behavior. These systems [1, 2] aim to diagnose a learner's developing knowledge structure based on question answering, link selection, or errors during problem solving.

Estimating workload as precise as possible is essential for the successful implementation of a learning environment. Online predictions of workload states can be used to adapt the learning material presented to the individual subject [3, 4]. The opportunity of adapting learning environments to current mental states such as inattention, cognitive workload, or mental fatigue might be very helpful for learners. Thus, a non-obtrusive, objective online measurement for workload detection, which can be used to adapt instructional material in real-time, is desirable. These requirements are achieved by measuring electrical currents on the scalp with the electroencephalography (EEG). Further, adapting the learning material online, based on EEG data can be realized by using Brain-Computer Interfaces (BCI).

A BCI is a direct link between a human brain and a technical system. It detects patterns in brain activity and translates them into input commands, given to a machine. The EEG is interpreted by using signal processing and machine learning techniques [5].

1.2 Research questions

The aim of this interdisciplinary thesis is to bridge the gap between psychology, neuroscience, as well as computer science, by developing an online adaptive learning environment, which is able to adjust to the individual workload states of each learner. The online adaptation will be based on specific types of workload via neural signatures in the EEG of learners. To reach this goal, a bundle of challenges of theoretical, methodological and practical nature have to be met during the course of this thesis:

1. How is suitable learning material designed, so that it can successfully be adapted in real-time to individual needs, based on EEG data?
2. Which specific neural signatures are associated with high and low workload and can they be detected in the EEG in real-time, while solving learning material?
3. Can a generalized classification method be developed and applied in a real-world learning scenario?
4. Is workload prediction across subjects possible to support each individual learner in his/her own learning progress successfully, by using an online EEG-adaptive learning environment?

1.3 Outline of the thesis

A precise workload prediction based on EEG data, as well as generalized classification methods working in real-time, are necessary for an applicable online adaptive learning environment. Therefore, the outline of this thesis is as follows:

- At first, the fundamentals about workload, electroencephalography, as well as feature selection methods are explained. Furthermore, the basics of BCIs and common classification methods are introduced in chapter 2.
- The state of the art of current tutoring systems and workload detection, as well as classification methods are reported in chapter 3.
- In chapter 4, the main ideas, challenges and hypotheses for the development of an EEG-based adaptive learning environment will be shown.
- Chapter 5 deals with the analysis of realistic tasks to investigate neural signatures of workload in EEG signals. Common signal processing and machine learning techniques are utilized and discussed.

- The comparison of different feature selection and extraction methods, as connectivity or independent component analysis, is shown in chapter 6, to discover which method is best for finding the neural workload features in EEG data.

- To enable generalized classification methods, the cross-task classification is evolved using a support vector machine (see chapter 7). Therefore, highly controlled working memory tasks are investigated for classifier training, allowing a specific workload manipulation. For classifier validation realistic learning materials are used. Further, the influence of subjective cognitive workload labeling (see chapter 8) and the task order effect during cross-task classification (see chapter 9) are examined in additional analyses.

- In chapter 10 the cross-subject regression is applied as another generalized classification method. Hence, it is investigated, if a precises workload prediction across a variety of subjects is possible, by using linear regression algorithms.

- As a proof of concept, the developed cross-subject regression method is applied in an online adaptive learning system. Realistic learning material, as it might be used in school settings, is adapted in real-time based on EEG data. Thus, each learner is kept in his/her optimal workload capacity range to be supported successfully. Furthermore, the performance of the newly developed learning environment is compared with a state of the art error-based adaptive learning environment. The task design, as well as the results are described in chapter 11.

- At the end, the obtained results will be discussed regarding the benefits and drawbacks of an online EEG-based adaptive learning environment.

2 Fundamentals for developing EEG-adaptive learning environments

The fundamentals about workload, electroencephalography, as well as feature extraction methods are explained in the following sections. Furthermore, the basics of Brain-Computer Interfaces and common classification methods are introduced.

2.1 Development of tutoring systems

Based on the cognitive load theory [6], the type and amount of workload during learning is crucial for successful learning and should be kept within an optimal range of a learner's workload capacity. The optimal workload range is characterized by providing tasks that keep a learner motivated without overwhelming or underchallenging him/her. Therefore, perfect learning material should constantly range between situations that are "too easy" and situations that are "too difficult" for each learner [7].

2.1.1 History of intelligent tutoring systems

Intelligent tutoring systems are computer systems that aim to support learners successfully. The development of tutoring systems has been discussed for centuries. In Figure 2.1 a short time line with some exemplary tutoring systems of the last decades is presented. In 1958 the Autograph was developed as one of the first tutoring systems in the USA. Subjects sat in front of a box and got learning material presented. The answers were not automatically recorded. A few years later it was possible to record the given answers via typewriter (System of Differential and Integral Calculus (SDIC) - 1960) or microphone (Autotutor MarkII - 1962) in a separate answer file. In 1971 the EDUCATOR was designed in Germany. This system permitted to store the answers directly to the questionnaires. The requirements for a current intelligent tutoring system are the capability to react in real-time and adapt the instructional learning material accordingly to the given answers or depending on the mental state of each learner. Corbett and Anderson (1992) developed the LISPITS, which could identify mistakes and provided constructive feedback to the users while performing the exercises.

2.1.2 Intelligent tutoring systems in present and future

So far, intelligent tutoring systems are mainly based on a learner's performance and behavior. The most common systems [1, 2] aim to analyze learners' knowledge based on question answering or the amount of errors during problem solving. The overall goal is

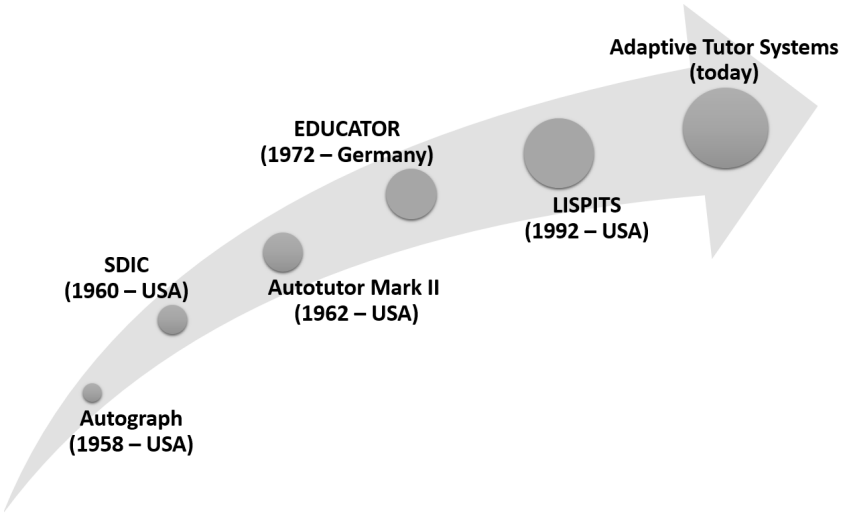


Figure 2.1: Timeline showing the evolution of tutoring systems from the mid of the 20th centuries till today.

to develop a tutoring system that concentrates on the continuous non-obtrusive detection of different levels of workload and adapt the learning material in real time, to keep the amount of workload of each learner within an optimal workload range. The main challenge to achieve this goal is finding appropriate measurements that allow for a continuous and non-obtrusive online tracking of a learner's workload.

2.2 Workload

Throughout this thesis, the term workload will be used as the amount of mental resources that are used to execute a specific task, also known as working memory load.
The concept of workload presented in the following section is based on Gerjets et al. [8]. Workload is related to the executive control of working memory. The small amount of information that can simultaneously be kept in mind during solving a cognitive task is called working memory [9]. Adjusting learning materials to the memory capabilities of learners is necessary to improve learning and education [9].
In cognitive load theory, the workload refers to the multicomponent working memory model by Baddeley [10, 11, 12], which distinguishes verbal and visual temporal storage components from an attentional control system. The storage components are the bottlenecks of learning. They constrain the amount of new information that can be processed simultaneously in order to be integrated into long-term memory. If the amount of information a learner has to process exceeds the capacity of these storage components, no optimal learning can be guaranteed.
Workload during learning is commonly assumed to be the result of an interaction between the complexities of the contents to be learned, the instructional design and a learner's prior knowledge [13]. For more knowledgeable learners, the same learning content can result

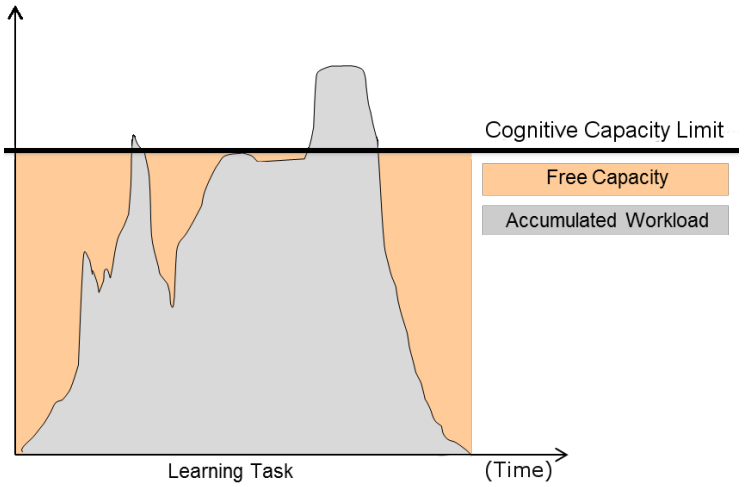


Figure 2.2: Cognitive workload during a learning phase over time. Orange represents free workload capacity and gray the needed workload for solving a task (adapted from [14]).

in lower workload as compared to novices. Furthermore, learning material A can impose higher workload for novice learners compared to a learning material B, whereas the reverse is true for more advanced learners. This is the so-called expertise reversal effect [13].
Workload changes during learning not only due to a learner's increasing level of knowledge, but also due to changing learning materials and task requirements presented to learners. Figure 2.2 shows a schematic diagram of different workload capacities needed during learning. To support a learner perfectly, the cognitive workload should be held around the cognitive capacity limit.

2.2.1 Brain areas involved in workload

Due to the complexity and the extensive range of working memory, a common localization in the brain is not possible. However, individual subsystems and working memory components can be assigned to specific brain regions. Figure 2.3 shows a schematic localization of the main working memory parts. Tasks, claiming the central executive (CE) processes lead to increased activity in the dorsal frontal lobe. The Brodmann areas (BA) 6, 8 and 9 in the pre-frontal cortex are relevant for continuous updating of the working memory [15]. BA 6 describes the pre-motor cortex, which is responsible for the linguistic (articulatory rehearsal = AR) or spatial (inner scribe = IS) repetitions. While the left temporal and parietal lobes are active during processing and storage of auditory information, the right hemisphere is more relevant in processing visual-spatial tasks [16]. The phonological store (PS) can be located at the BA 40 [16]. The visual memory, or visual cache system (VC) is located in the occipital brain areas. Further, Klimesch [17] postulated, that the medial temporal lobe, including the hippocampus, as well as the pre-frontal cortex, are relevant brain areas for working memory.

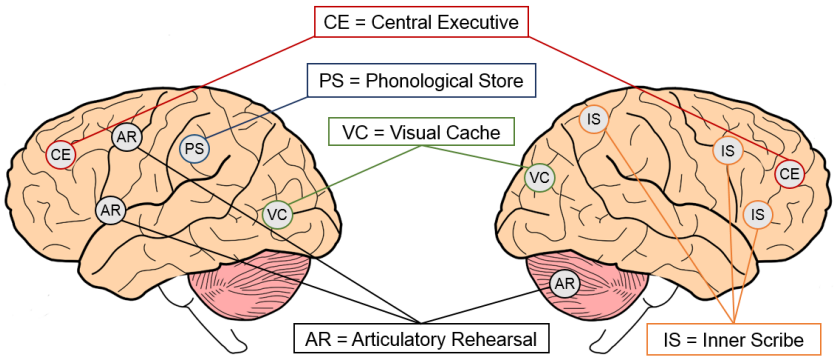


Figure 2.3: Schematic localization of the subsystems of the working memory (modified from [16]). The central executive (CE) is positioned at the pre-frontal cortex. While the articulatory rehearsal (AR) is localized on the left pre-motor cortex, the inner scribe (IS) is located on the right hemisphere. The phonological storage (PS) is located at the left parietal cortex, while the visual cache system (VC) is positioned on occipital brain areas.

Based on functional magnet resonance imaging (fMRI), Klein and colleagues [18] were able to specify brain areas which differ significantly while solving easy and complex learning material. While treating easy mathematical assignments, an activation of the left interparietal sulcus could be detected. Whereas dealing with more complex mathematical assignments a bilateral activation of the intraparietal sulcus, as well as an activation of the gyrus angularis could be measured.

2.2.2 Methods for workload detection

Until now, methods for workload detection mostly rely either on subjective workload ratings [19, 14], or on dual-task procedures [20, 21, 22]. Both procedures are likely to disrupt and annoy learners while studying. Mihalca et al. [23] developed a workload-adaptive learning environment based on subjective rating scales, leading to promising results. Nevertheless, to obtain the subjective workload ratings from each learner, they were frequently interrupted during their learning process. Furthermore, the adaptation of the presented learning material was not instantaneous, as workload could not be assessed continuously.
Performance measures like the amount of errors during problem solving [1, 2] are less obtrusive methods for workload measurement. They do not indicate whether the solution of a task was achieved with a high or low amount of workload. This type of performance measurements might frustrate learners in case of repeated failures and are not usable for a continuous assessment. To conclude, the common state of the art measures, like task performance, rating scales and dual tasks, do not allow a continuous and non-obtrusive measurement of workload, which is necessary for the development of workload-adaptive learning environments.
Additional methods for a continuous workload assessment are physiological measures like pupil dilation or skin conductivity [24, 25, 26]. However, these measures mostly turned out to be not very reliable and specific indicators of workload [20, 14]. A novel methodology

for solving this problem are more direct physiological measures of neural activity to derive more specific indicators of workload during learning. Therefore, EEG signals can be used for workload detection [27] and can be applied to learning scenarios [28, 29, 30].

2.3 Electroencephalography (EEG)

Electroencephalography (EEG) is the recording of electrical activity on the scalp. Since the discovery of the human EEG activity in 1924 by Hans Berger [31], EEG measurements were mainly used for medical reasons and for analyzing the brain function. It is a non-invasive method, which records changes in the electric potential differences produced by brain activity. Thus, mental processes can directly be reflected by the EEG measurements. Compared to other biosignals it has a high temporal resolution which is essential for analyzing users workload [32].

2.3.1 Source of EEG activity

The human head mainly consists of three layers: the scalp, skull and the brain. In Figure 2.4 their approximate resistivities and thicknesses are shown.

The signal which can be recorded on the scalp is generated by extracellular field potentials [34]. Because of the filtering and attenuation from skull and scalp, the electrical activity generated by a single neuron is too small to be recorded by an EEG. Therefore, only large populations of active neurons can generate enough potential to be recordable using the scalp electrodes [33]. The recorded EEG signal reflects the summation of synchronous electrical activity from a population of neurons of several square centimeters of the cortex with a similar spatial order [35].

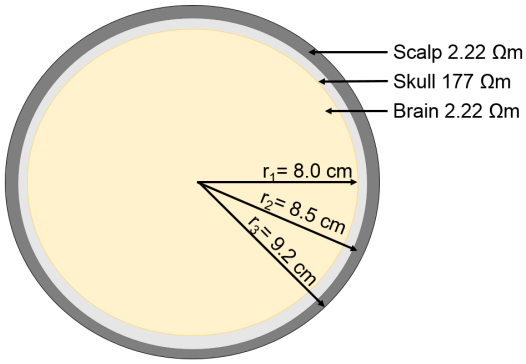


Figure 2.4: The three main layers of the brain including their approximate resistivities and thicknesses (adapted from [33]).

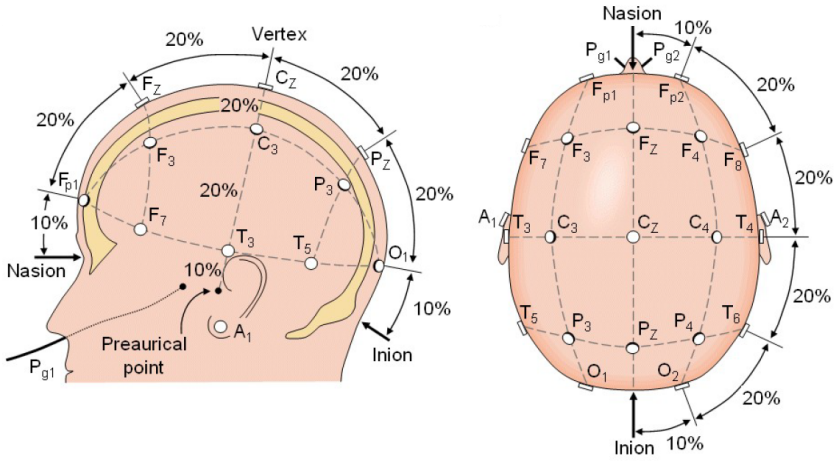


Figure 2.5: A schematic representation of the international 10-20 electrode setting, taken from [37].

2.3.2 EEG recordings and measurement

In practice, 2 to 128 electrodes are placed on the scalp according to the extended international 10–20 system [36], to record the EEG (see Figure 2.5). The "10" and "20" refer to the actual distances between adjacent electrodes. They are either 10 % or 20 % of the total front–back or right–left distance of the skull.
The EEG signal is recorded relative to a reference, so that no absolute values of electric potentials are shown. This leads to a high temporal (< 1 ms) but limited spatial resolution, depending on the number of used electrodes. A typical human EEG signal measured from the scalp has an amplitude of about $10\,\mu$V to $100\,\mu$V [38]. Due to low signal intensity, amplifiers are used for intensification. In current EEG systems a conductive gel has to be applied to improve the conductance between scalp and electrodes and thus improve the signal to noise ratio (SNR). In the last decade, the classical passive Ag/AgCl electrode was more and more replaced by active electrodes. Active electrodes have the benefit that an additional amplifier is placed right next to each electrode. Thus recordings from active electrodes are less effected from noise induced by cable movements. The electrodes are less prone to picking up the 50 Hz noise induced by the alternating current of European power lines than passive electrodes and deliver good signal quality even at higher impedance [39]. The portability is a major benefit of EEGs. Furthermore, it is a rather cheap option to record brain signals and makes it possible to trace cognitive processes. As it is a non-invasive method to measure brain activity, it is ethically unproblematic. The substantial amount of preparation time is a major drawback of EEGs. This is mostly dependent on the time needed for minimizing the impedance between electrode and skin, which usually is realized by the application of a conductive gel [40]. Current special dry- or water-based electrodes are available, which allow to omit the use of conductive gel and thus lower preparation time. So far the signal quality is not as good as for standard gel-based active electrodes [41, 42]. Further improvements of these electrodes will allow the application of EEG systems in real-world settings.

2.3.3 Non-stationarity in the EEG data

One big challenge in EEG analysis and classification are non-stationarities of the recorded signals. Non-stationarities are changes in EEG signals occurring continuously in association with diverse behavioral and mental states [43, 44], which are characterized by significant day-to-day and subject-to-subject variations. The alternations over time can originate as results of changes in the subjects brain process, e.g., fatigue, modification of recording conditions, as well as changes in operation strategies by the subjects.

2.3.4 Frequency bands of the human EEG

The typical rhythmic activity, which was already described by Hans Berger in 1924, can be divided into frequency bands, which are corresponding to specific mental states. The frequency distribution can be extracted from the EEG recordings through spectral analysis that yields a so called power spectrum (see section 2.4.1).
Based on their frequency and amplitude, the recorded brain activity can be interpreted. Changes in the frequency bands suggest changes in awareness, emotional and attentional states. The five basic wave forms in the human EEG are the following [45]:

- Delta-frequency band: 0.5 Hz – 4 Hz
- Theta-frequency band: 4 Hz – 8 Hz
- Alpha-frequency band: 8 Hz – 13 Hz
- Beta-frequency band: 13 Hz – 40 Hz
- Gamma-frequency band: > 40 Hz

The boundaries of the frequency bands are blurred, so this is a merely rough division and the frequency bands may overlap at the borders. In Figure 2.6 the curves of the main frequency bands are exemplarily shown.

2.3.5 EEG as an index of workload

By using EEG, a continuous measurement of brain activity is possible and allows to estimate the subject's workload. In the research area of workload, EEG activity is well established for determining the workload on individual basis. The physiological measurement method is sensitive to changes in workload demands associated with performance of specific tasks [46, 47, 48, 49]. EEG can be used to estimate the amount of workload of each learner during a learning session. The currently best studied EEG correlates of workload, suitable for continuous online measurements of learning states, are variations in the oscillatory power of theta- and alpha-frequency band activity [17]. Increasing workload leads to an increase in theta-frequency band activity (synchronization) over frontal-midline electrodes [50, 51, 48] and a decrease in alpha-frequency band power (desynchronization) over parietal-occipital electrodes [50, 52, 53]. It is commonly assumed, that a more complex assignment induces higher workload, therefore a stronger alpha-desynchronization as

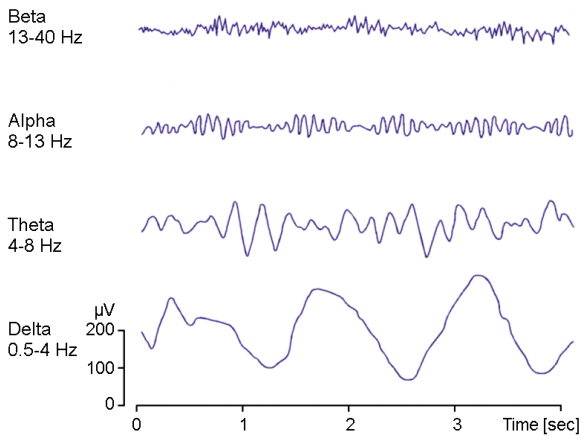


Figure 2.6: Four typical dominant brain rhythms, from high to low frequencies, modified from [37].

well as a theta-synchronization are expected, compared to easy assignments [48, 17, 50]. Klimesch et al. [54] hypothesized these EEG correlates to reflect cognitive achievement. Further, he postulated a strong relationship between these oscillatory variations and memory decline. Based on these findings, Antonenko et al. [29] proposed to use oscillatory power for the measurement of workload in instructional research.

2.3.5.1 Theta-band oscillation

Synchronization in the theta-frequency reflects episodic memory and the encoding of new information [17]. It seems feasible that theta-power reflects attentional demands, task difficulty and workload. As opposed to alpha-frequency, theta-power increases during complex tasks over frontal-midline electrodes [50, 53, 48]. Mecklinger et al. [55] showed higher workload cause increase in theta-band power. Gevins and Smith [27] measured a theta-synchronization during more complex tasks relative to more easy tasks.

2.3.5.2 Alpha-band oscillation

In the literature, alpha-band desynchronization over parietal-occipital electrodes accompany task difficulty [50, 53, 52]. Further EEG studies indicate that alpha-oscillations significantly correlate with processing speed, sensory-semantic memory processes of long-term memory systems and cognitive information processes. For example, alpha-power decreased, while workload and reaction times increased with increasing mental arithmetic task difficulty [56]. Thus, the increased magnitude of an alpha-desynchronization is a result of increasing task complexity or attention [57, 58, 59]. Furthermore, increased desynchronization reflects strong attention and increasing memory performance [17]. Alpha-power desynchonization can be interpreted as electrophysiological correlate of activated cortical areas involved in processing of cognitive information [60].

In the following sections, the focus will be on analyzing, evaluating and interpreting differentiations in frequency bands. The hypothesis of this thesis is based on Klimesch [17] who postulated, good memory performance will be reflected by an increase in theta-power in combination with a large decrease in alpha-power.

2.4 Feature extraction methods for classifying EEG data

EEG signals are continuous variations of potentials as a function over time. After recording EEG data, the raw signal has to be preprocessed to allow quantitative analysis, as well as a classification of the recorded brain activity. Diverse methods can be used to extract features of the recorded brain signals, which display certain characteristics of the recorded brain activity.

2.4.1 Power spectrum

In the research field of workload detection in EEG data, the power spectrum estimation is commonly used. This spectral analysis gives insights into frequency power at different locations in the brain. It provides an estimate of the signal variance as a function of frequencies. For each channel a power frequency vector with n-frequency bins exists, which describes how the energy of a time series is distributed with frequency [61]. For estimating the power spectrum, diverse methods can be applied. Fast Fourier Transformation (FFT) is widely used to convert data from time to frequency domain, by decomposing a sequence of values into components of different frequencies. Another approach is estimating the power spectrum by means of an autoregressive model (ARM) [62, 63]. This method is commonly used, since it is the most robust method for feature extraction in EEG data, when dealing with short time segments [64], because of its superior resolution and smaller statistical fluctuations [65, 66]. As the developed system should react in real-time, fast response times have to be guaranteed. Thus, short time segments of e.g., 400 ms can occur for analyzing the data. Therefore, the ARM is mainly used for feature extraction in this thesis.

2.4.1.1 Autoregressive model

The ARM algorithm is a parametric method, where EEG signals are described in terms of a mathematical model, characterized by a set of parameters. Based on its input data, an ARM is computed with autoregressive coefficients as output, which are actually the weights in a linear filter. The ARM can be expressed as

$$Y_t = \sum_{j=1}^{p} w_{t-j} Y_{t-j} + e \tag{2.1}$$

where Y_t is the predicted signal at time t, p the model order of the ARM to track the changing spectrum properly, w the p weights scaling the input signal Y at time $t-1,\ldots,$ $t-p$ and e the prediction error [67].

The resulting ARM can then be used to estimate the power spectrum by

$$P(\omega) = \frac{1}{\left|1 - \sum_{j=1}^{p} w_j e^{-ij\omega}\right|} \tag{2.2}$$

where $P(\omega)$ is used to denote the signal power spectrum, w_j are the estimated filter weights and i represents the imaginary unit [33].
For generating an ARM, diverse methods could be accomplished. Using the Burg Algorithm [68] based on the Maximum Entropy Method is a common choice for real time closed loop systems, as it is guaranteed to produce a stable model.

2.4.1.2 Quantification of EEG reactivity

Cognitive processes can be detected by changes in the ongoing EEG as event-related desynchronizations (ERD) or event-related synchronizations (ERS). The $\%ERS/ERD$ method [69] converts absolute band power into percentage power by defining the power within a reference interval as 100 % [34].
Thus, the proportion in percent of band power increase (ERS) or decrease (ERD) in any frequency band from a reference interval to an activation interval is determined. The amount of $\%ERS/ERD$ can be calculated according to the formula from Pfurtscheller and colleagues [69]:

$$\%ERS/ERD = \left[\frac{(I_A - I_R)}{I_R}\right] \times 100 \tag{2.3}$$

with I_A being the activation interval, the time interval that strongly imposes the task specific brain activation. I_R serving as reference interval, which imposes only low level of brain activity and is mainly recorded during a fixation or resting phase.
Positive values indicate increases in band power (ERS) and negative values indicate decreases (ERD). For example an increase of task complexity or attention demand results in an increased magnitude of ERD in the alpha-frequency band [17].

2.4.2 Connectivity analysis

Whereas the power spectrum can only be used for one signal and leads to a composition of univariate signal features, connectivity measures can be used to extract information about the interaction or relation between two signals. However, as the primary function of the brain is to organize the behavior, transient patterns of cortical source dynamics that modulate information transmission among non-continuous brain areas play critical roles in cognitive state maintenance and information processing [70, 71, 72, 73]. Connectivity patterns are of high interest for the analysis of the brain behavior during learning. Brain connectivity can be subdivided into neuroanatomical, functional and effective connectivity [74].

In the following, two commonly used connectivity measures are described: coherence as method for evaluating the functional connectivity and Granger causality to measure the effective connectivity.

2.4.2.1 Coherence

The functional connectivity refers to symmetric, undirected correlations among the activities of cortical sources [75]. The earliest functional connectivity studies used coherence between measured EEG scalp signals. The spectral coherence is a statistic that can be used to examine the relation between two signals. Given two signals $s_x(t)$ and $s_y(t)$ (e.g., two EEG channels), the coherence can be calculated from the cross-spectrum $C_{xy}(f)$, which is a measure of the joint spectral properties of two data channels at frequency f. The cross-spectrum can be estimated from pairs of Fourier coefficients as an average over K epochs:

$$C_{xy}(f) = \frac{2}{K} \sum_{k=1}^{K} F_{xk}(f) F_{yk}^{*}(f) \tag{2.4}$$

with $F_{xk}(f)$ being the Fourier transform of $s_x(t)$ in epoch k and $F_{yk}^{*}(f)$ being the complex conjugate of the Fourier transform $F_{yk}(f)$.
The cross-spectrum is a measure of the covariance between two signals at a specific frequency band. Unlike the power spectrum which is real valued, the cross-spectrum is complex valued. The cross-spectrum $C_{xx}(f)$ in between a single signal $s_x(t)$ is equivalent to the power spectrum $P_x(f)$ of that signal. By normalizing the squared magnitude of the cross-spectrum by the power spectrum $P_x(f)$ and $P_y(f)$ of the two corresponding channels, the coherence is obtained [76].

$$Coh_{xy}(f) = \frac{|C_{xy}(f)|^2}{P_x(f) P_y(f)} \tag{2.5}$$

The result is a value between 0 and 1, with 1 meaning the frequency components of both signals do correlate perfectly. A general deficit of functional connectivity is that they cannot be used to identify asymmetric information transfer or causal dependencies between cortical sources.

2.4.2.2 Granger causality

Effective connectivity denotes directed or causal dependencies between brain regions [75]. A popular effective connectivity method is Granger causality. This is a statistical hypothesis test for determining, whether one time series is useful for forecasting another. If a time series $s_i(t)$ contains information in past terms that helps in the prediction of $s_{i+1}(t)$, then $s_i(t)$ is said to cause $s_{i+1}(t)$.

In order to show the improvement of the prediction of one signal by taking the past terms of the second signal into account, bivariate ARMs are fitted to the signals [77]. Formally, a bivariate linear ARM of two variables s_1 and s_2 can be written as

$$s_1(t) = \sum_{j=1}^{p} A_{11}(j)s_1(t-j) + \sum_{j=1}^{p} A_{12}(j)s_2(t-j) + E_1(t) \tag{2.6}$$

$$s_2(t) = \sum_{j=1}^{p} A_{21}(j)s_1(t-j) + \sum_{j=1}^{p} A_{22}(j)s_2(t-j) + E_2(t) \tag{2.7}$$

where p is the model order (i.e., the maximum number of observations included in the model), the matrix A contains the coefficients of the model and E_1 as well as E_2 are prediction errors for each time series. It is said that s_1 causes s_2, if the variance of E_2 is reduced by the inclusion of the s_1 terms in equation (2.7) and vice versa. In other words, s_1 causes s_2, if the coefficients in A_{21} are significantly different from 0. The generalized ARM defined for an arbitrary number of channels is represented as

$$\mathbf{s}(\mathbf{t}) = \sum_{j=1}^{p} A(j)\mathbf{s}(\mathbf{t}-\mathbf{j}) + \mathbf{e}(\mathbf{t}) \tag{2.8}$$

with $\mathbf{s}(\mathbf{t})$ being a data vector at time t and $\mathbf{e}(\mathbf{t})$ indicating the prediction errors.

Equation 2.8 can be rewritten as follows:

$$\mathbf{e}(\mathbf{t}) = \mathbf{s}(\mathbf{t}) - \sum_{j=1}^{p} A(j)\mathbf{s}(\mathbf{t}-\mathbf{j}) \tag{2.9}$$

By using the Fourier method, it is possible to transform the model equation to the frequency domain, to examine the Granger causality in the spectral domain [78].

$$\mathbf{e}(\mathbf{f}) = \mathbf{s}(\mathbf{f}) \left[\delta_{nm} - \sum_{j=1}^{p} a_{nm}(j) e^{-n2\pi f j} \right] \tag{2.10}$$

$\mathbf{e}(\mathbf{f})$ is the transformation of $\mathbf{e}(\mathbf{t})$, where f denotes the frequency, δ_{nm} is one, when $n = m$ and zero elsewhere.

This equation can be presented in a simplified form, as

$$\mathbf{e}(\mathbf{f}) = \mathbf{s}(\mathbf{f})A(f) \tag{2.11}$$

$$\mathbf{s}(\mathbf{f}) = A^{-1}(f)\mathbf{s}(\mathbf{f}) = H(f)\mathbf{e}(\mathbf{f}) \tag{2.12}$$

where $H(f) = A^{-1}(f)$ is the transfer matrix, which contains the information about the interactions between the time series and $h_{nm}(f)$, the element of the n^{th} row and m^{th} column of the transfer matrix.

2.4.3 Source localization

Usually, electrodes placed on the scalp record potential difference of brain activity. As the brain is a three dimensional structure, electrical activity is not only generated on its surface. The recorded potentials at scalp electrodes represent the superposition of cortical potentials and therefore the signals at various electrodes are highly correlated. The projection cannot represent a one-to-one mapping [40]. It is assumed, that the organization of the brain contains regions, processing information for specific tasks.
If activity of the brain can be traced back to one or more of its sources, the type of process, which is underlying the specific signal, can be identified. A common procedure for identifying sources of brain activity is the Independent Component Analysis (ICA) with subsequent dipole fitting (e.g., LORETA).

2.4.3.1 Independent Component Analysis

The Independent Component Analysis (ICA) is a multivariate analysis, that uses statistically independent components (ICs) for reconstructing the source space [79]. This method is utilized to remove the correlations introduced into the measured scalp data by mixing of the component signals. Using ICA, independent processes can be identified by their temporal, spectral and spatial properties.
This linear transformation method requires the following assumptions (from [80]):

- Components have fixed spatial projections during the epoch of interest.
- Signal conduction times are negligible and the potentials from different components combine linearly at the scalp without time delays.
- The number of components that can be determined is equal to the number of scalp electrodes.
- Components are temporally independent of each other across the epoch of evaluation.

The general ICA model is

$$E = Ws \tag{2.13}$$

and consists of E being an epoch of scalp EEG, represented by a $m \times n$ matrix, with m channels and n time stamps, W is a $m \times m$ mixing matrix and s a $m \times n$ component matrix, containing random variables representing the sources.
To find an unmixing matrix W^{-1}, the ICA attempts to find a transform that makes the component output statistically independent. When the unmixing matrix is applied to the measured EEG, the multichannel data E will be decomposed into a sum of temporally independent and spatially fixed components, as

$$u = W^{-1}E \tag{2.14}$$

In (2.14), the variable u is identical to the variable s in (2.13), except for an unknown scaling factor.
The results of an ICA determine how much each component contributes to each EEG channel. However, a one-to-one correspondence does not necessarily exist between components and individual anatomical sources. A single source may produce waveforms that separate into several components. Conversely, multiple sources may contribute to one of the components [80]. These component maps may nearly perfectly match the projection of a single equivalent brain dipole. The problem of finding the location of a signal source generating a dipolar signal at the scalp, can be solved by using dipole fitting algorithms, subsequently.

2.4.3.2 Dipole fitting

Finding the intracranial sources generating a known distribution of scalp potentials is a major obstacle by using EEG data to visualize brain dynamics. To overcome the inverse problem, an EEG feature is assumed to be generated by one or more intracranial dipole sources. The problem of finding the locations is most reasonable for relatively small sources, but unreasonable for potentials generated by cerebral cortical neurons, because of many simultaneous active dipoles. A well-known EEG source imaging algorithm is the Laplacian weighted minimum norm (LORETA) [81]. It is a method for locating the electrical activity in the brain, based on scalp potentials from multiple EEG channels. Firstly, an ICA decomposition has to be conducted, then the components which should be fitted have to be selected and finally the dipole model can be matched to the ICA component. The method selects the solution with a smooth spatial distribution by minimizing the Laplacian of the weighted sources, as the activity in neurons in neighboring areas of the cortex correlates. LORETA generally provides rather blurred solutions. An analysis of limited frequency bands can be used to determine which regions of the brain are activated during different mental tasks.

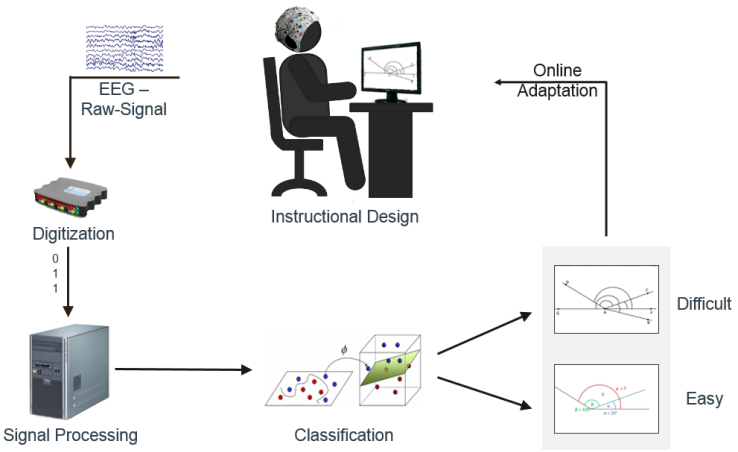


Figure 2.7: Schematic workflow of a BCI application - an online adaptive learning environment based on EEG data.

2.5 Brain-Computer Interface

A Brain-Computer Interface (BCI) is a technical system allowing a direct communication between the human brain and an external computer device. It allows a non-muscular interaction with the environment because it does not use the common brain pathways of the muscles. The commands can be directly encoded, based on the brain activity, as they provide information about mental states that can be translated into human behavior.

2.5.1 Operation of a BCI

As common communication systems, the BCI consists of input data (e.g., brain activity), components that translate the input into output signals and output data (e.g., control device). The brain activity can be translated into a signal for a control device, by passing through four operations of the BCI-system, as shown in Figure 2.7:

- Data recording and digitization
- Signal processing and feature extraction
- Classification
- Application

At first, the brain signal is acquired with one of the available recording techniques. A variety of recording techniques exists, which rely on different underlying physiological processes. This includes methods as functional magnetic resonance imaging (fMRI) or near infrared spectroscopy (NIRS), which are based on the hemodynamic response of the brain, which lead to a bad temporal resolution. Another method measuring the electrical brain activity indirectly is the Magnetoencephalography (MEG), where the magnetic field

induced by ionic currents can be recorded. This causes high spatial and high temporal resolution, but the recording device is not portable and requires a shielded room. Although these methods can be used for BCI, EEG is the most common technique to measure the brain activity, as it has a high temporal resolution and measures the electrical brain activity directly. Further, the simple, low cost equipment as well as the portability offers the best possibility to ensure a non-muscular communication and control in real-world BCI applications. After selecting the desired recording method, the data will be recorded, amplified and digitized. To improve the signal quality, the signal is preprocessed, thus it can be filtered (e.g., high-pass, low-pass) or artifacts can be removed. To allow an easier translation of the recorded brain signal to a "control" signal, feature extraction methods are applied to the digitized, preprocessed signal. Features can be from time domain (e.g., event related potentials) as well as from frequency domain (e.g., frequency bands). After feature extraction, the features are classified by using appropriate classification methods. The classifier result serves as input for a computer application, e.g., a spelling device, mathematical tasks or an online adaptive learning environment.

2.5.2 Application scenarios for BCIs

The basic idea of a BCI-based application was to provide communication and control channels for severely disabled persons. As the reliability and usability of BCI systems has improved over the past decade, BCI approaches are also developed for applications beyond assistive technologies addressing more general problems of human-computer interaction [82]. Cognitive user states can be used to support a given interaction. With this more general use, new methods for real time cognitive state assessment became available. By using the two definitions from Zander and Kothe [83], two types of BCI systems can be distinguished: active BCI and passive BCI.

Active BCI

An active BCI is an interacting system, which derives its outputs from brain activity which is consciously controlled by the user. Such a system is independent from external events, for controlling an application. The definition describes a classical BCI application, a direct control of an application for patients.

Passive BCI

A passive BCI is an interacting system, which derives its outputs from brain activity without the purpose of voluntary control, for enriching a human-computer interaction with implicit information. The passive BCI can be considered as an additional, implicit communication channel containing information about the current user state.

In this thesis passive BCIs for designing workload-adaptive learning technologies based on EEG signals are used, to prevent learners from getting overwhelmed and confused due to complex learning materials.

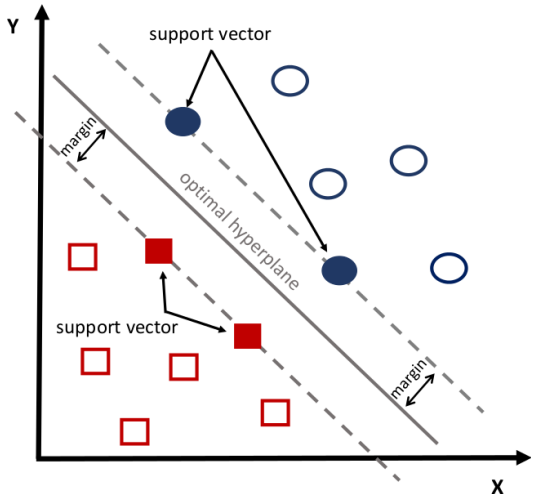


Figure 2.8: Optimal hyperplane, margin and support vectors for a two class problem found by a SVM.

2.6 Classification methods used for EEG data

For developing an efficient adaptive learning environment based on passive BCI technology, it is crucial to be able to classify diverse workload states of each learner. Classification is used to predict individual workload states (e.g., excessive demand or optimal cognitive load) based on features derived from brain signals. For the categorization of unknown features, the classification algorithm needs to be trained on previously collected training data. There are two major methods to train the classifier, supervised and unsupervised. For supervised learning methods, labeled training data is required, meaning each data point has a corresponding class label. In contrast, unsupervised learning algorithms operate on unlabeled data. After a training phase, the classifier is, in both modes, able to classify new data points with unknown labels. The classification output is a corresponding class which can then be used as a "control signal" for the passive BCI applications. This thesis will concentrate on two methods for workload classification: supervised non-linear Support Vector Machines (SVMs) as well as supervised linear Regression.

2.6.1 Support Vector Machine

A Support Vector Machine (SVM) is a supervised classification tool introduced by Vapnik [84] in 1963. This machine learning method is based on statistical learning theory, aiming to minimize the likelihood of classification error. A feasible training algorithm for non-linear classification was not developed until 1992 [85].
Given n training vectors $\mathbf{x_i} \in \mathbb{R}^m$, $i = 1,\ldots,n$ and corresponding class labels $y_i \in \{-1,1\}$, the SVM finds the optimal hyperplane that maximizes the margin, i.e., the distance between the closest data points (support vectors) of the two classes, to increase generalization capabilities.

Assuming linear separability, this is done by solving the following optimization problem:

$$\arg\min_{\mathbf{w},b} \|\mathbf{w}\| \quad \text{subject to } y_i(\mathbf{w}\cdot\mathbf{x_i}-b)\geq 1 \quad \text{for all } i=1,\ldots,n \tag{2.15}$$

where b is the bias, depending on $\mathbf{x_i}$ and y_i and $\mathbf{w}$ the normal of the hyperplane.
If no hyperplane can be found to linearly separate the data points, the SVM uses slack variables ξ to allow misclassification and approximation. ξ measure the degree of misclassification of data $\mathbf{x_i}$, leading to the following modification:

$$y_i(\mathbf{w}\cdot\mathbf{x_i}-b)\geq 1-\xi_i \quad \text{for all } i=1,\ldots,n \tag{2.16}$$

The SVM can be used as non-linear classifier to enable the separation of non-linear data, by replacing every dot product by a non-linear kernel function. In the case of non-linear separable classes, the approach used by SVMs is to map the input data into a higher dimensional space. This allows the algorithm to fit the maximum margin in a transformed feature space. Thus, data from two different classes can be linearly separated by a hyperplane in this higher dimensional feature space. The transformation into the high dimensional space is achieved by kernels [86]. As the radial basis function kernel (RBF) is mainly used for classifying EEG data with SVMs [87] this kernel is used in this thesis as well. The RBF kernel on two samples $\mathbf{x_i}$ and $\mathbf{x_j}$ is defined as:

$$k(\mathbf{x_i},\mathbf{x_j}) = e^{-\gamma\left\|\mathbf{x_i}-\mathbf{x_j}\right\|^2} \tag{2.17}$$

with $\gamma>0$ being a kernel parameter.
Using Lagrage multipliers $\alpha_i,\beta_i\geq 0$ for finding the local maximum and minimum in (2.16) yields the following optimization problem:

$$\arg\min_{\mathbf{w},\xi,b}\max_{\alpha,\beta}\left\{\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w}\cdot\mathbf{x_i}-b\right)-1+\xi_i\right]-\sum_{i=1}^{n}\beta_i\xi_i\right\} \tag{2.18}$$

where C is the regularization parameter, with $0<C<\infty$. C controls the cost of misclassification on the training data. A larger C penalizes the cost of misclassification stronger, leading to potential overfitting.

Sequential Minimal Optimization

Sequential Minimal Optimization (SMO) [88] is an iterative algorithm, which can be used to efficiently solve the optimization problem shown in (2.18) and is also implemented in LIBSVM[1] [89], which is used in this thesis. The optimization problem is divided into smaller sub-problems, involving two Lagrange multipliers (α_i, α_j).

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Therefore, (2.18) has to be reformulated into the dual form, with $k(\mathbf{x_i}, \mathbf{x_j})$ being the kernel function [90]:

$$\min_{\alpha} f(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x_i}, \mathbf{x_j}) \tag{2.19}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{for any } i = 1, \ldots, n \tag{2.20}$$

In each iteration, a pair (α_i, α_j) is found that violates the Karush-Kuhn-Tucker conditions [91] for the given optimization problem. By reducing the constrains to:

$$\begin{aligned} 0 &\leq \alpha_i \\ \alpha_j &\leq C \\ \alpha_i y_i + \alpha_j y_j &= -\sum_{\substack{\ell=1 \\ \ell \neq i,j}}^{n} \alpha_\ell y_\ell \end{aligned} \tag{2.21}$$

the optimization problem can be solved regarding (α_i,α_j) for n training vectors. Furthermore, the function value $f(\alpha)$ strictly decreases with each iteration and the SMO algorithm converges to an optimum.

2.6.2 Linear regression

Linear regression is a supervised learning algorithm. It is an approach for modeling the relationship between a dependent variable, $\mathcal{Y}$ (i.e., response variable) and an independent variable $\mathcal{X}$ (i.e., predictor variable). Compared to the classification problem, where discrete values are predicted, the regression problem predicts real-valued output. In linear regression, unknown model parameters are estimated from the data. Given n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ a general linear regression model is described by:

$$y_i = \alpha + x_i \beta + \varepsilon_i \quad i = 1, \ldots, n \tag{2.22}$$

where y_i is the i-th response, α and β are unknown parameters, which define the linear relationship between x_i and y_i. ε_i is the i-th noise term, based on the random error in the measurement [92, 93].

Linear regression models are often fitted using the least squares approach, where β is regularized to have small values (i.e., ridge regression).

Thus, the linear ridge regression can be formulated as

$$\arg\min_{\alpha,\beta} \sum_{i=1}^{n} (\alpha + x_i\beta - y_i)^2 + \omega \left(\alpha^2 + \|\beta\|^2\right) \tag{2.23}$$

where ω is the weighting factor of the regularization.

2.6.3 Linear mixed models

A linear mixed model (LMM) is an extension of linear regression models that can be used to describe the relationship in clustered data as a mathematical formula [94]. They are particularly useful in settings where repeated measurements are made on the same statistical units. Compared to a linear regression, where each observation has to be independent, LMMs allow continuous dependent and independent variables, as well as interactions between any combination of discrete and continuous variables. Assuming, one subject rates a number of assignments with a certain degree of difficulty. These multiple ratings from the same subject cannot be regarded as independent from each other. Every person has a slightly different experience of difficulty. Inter-dependence in the responses in relation to some factor can be made by LMMs. Therefore, this model consists of two parts: fixed and random effects. Fixed effects are the independent variables as used in a linear regression. Random effects are the variables which are specific for the data. This effect accounts for the correlation in the data. The individual differences in relation to each factor are modeled by assuming difference random intercepts for each response. This allows to resolve this non-independence by assuming a different "difficulty baseline" value for each subject.
The standard form of a LMM is:

$$\mathbf{y} = X\beta + Z\mathbf{b} + \varepsilon \tag{2.24}$$

where $\mathbf{y}$ is the known response vector of observations. β is an unknown vector of fixed effects, which are common to all subjects, while $\mathbf{b}$ is the subject specific random effect, an unknown vector. ε is the observation error, which is independent from the random effect vector $\mathbf{b}$. X and Z are known design matrices relating the observations $\mathbf{y}$ to β and $\mathbf{b}$, respectively [94].

2.6.4 Performance measurements for evaluating prediction methods

To be able to compare the performance of diverse classification methods, a wide range of performance measurements are known. In this section all metrics which are used in the following studies will be introduced.

Correlation coefficient

The Pearson's correlation coefficient (*CC*) is used to observe the statistical relationship between two variables X and Y. The correlation coefficient applied to samples can be obtained from the population Pearson's correlation coefficient:

$$p_{x,y} = \frac{cov(X,Y)}{\sigma_X, \sigma_Y} \tag{2.25}$$

$$= \frac{E\left[(X-\mu_x)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y} \tag{2.26}$$

$$= \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}} \tag{2.27}$$

where *cov* is the covariance, σ_X is the standard deviation of X, μ_X is the mean of X and E is the expected value. The *CC* of two datasets $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$, containing n values each, is calculated by substituting covariances and variances in (2.27) with their estimates on data points x_i and y_i. Therefore, the *CC*, mostly given as r, is defined as follows:

$$r = \frac{n\sum_{i=1}^{k} x_i y_i - \sum_{i=1}^{k} x_i \sum_{i=1}^{k} y_i}{\sqrt{n\sum_{i=1}^{k} x_i^2 - \left(\sum_{i=1}^{k} x_i\right)^2}\sqrt{n\sum_{i=1}^{k} y_i^2 - \left(\sum_{i=1}^{k} y_i\right)^2}} \tag{2.28}$$

The *CC* can be positive, as well as negative. If the *CC* is close to 0, there is no linear correlation between the actual and the predicted variables. A *CC* of 1 means the two variables correlate perfectly.

Squared Correlation Coefficient

The squared correlation coefficients (r^2) are used for performance estimation and feature selection [95, 96]. These values can be assumed to be a correlation measurement. The r^2-value describes the variance dimension for a feature, which is explained by the class membership. It is located between 0 and 1, where 0 stands for no correlation and 1 for perfect correlation. Using this value as performance estimation, it enables to compare diverse prediction methods.

Root Mean Square Error

The Root Mean Square Error (*RMSE*) is the square root of the mean squared error (*MSE*). The *MSE* is a measure of how close a fitted line is to data points. The squaring is done in order that negative values do not cancel positive values. Furthermore, to punish data points which are further away of the fitted line stronger, compared to those which are almost on the fitted line. The smaller the *MSE*, the closer the fit is to the data.

The *MSE* is calculated as follows,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{2.29}$$

where x_i and y_i are data points from the datasets $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$ containing n values.
Calculating the square root of the *MSE* leads to the *RMSE*, described as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.30}$$

The *RMSE* is thus the average distance, of a data point from the fitted line, measured along a vertical line.

Global deviation

The global deviation (*GD*) is an additional performance measurement to observe the statistical relationship between actual values and the corresponding predicted variables. *GD* is defined by the average squared difference [97]:

$$GD(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)\right)^2 \tag{2.31}$$

where y_i denotes the actual value at time instance i and x_i is the corresponding predicted value. The smaller the *GD*-value, the smaller the predicted bias error. Compared to *GD*, the *RMSE* and the *CC* captures noise. Furthermore, it is the only method allowing a reasonable estimation of the prediction bias.

Accuracy

The classification accuracy (*Acc*) is widely used as performance measurement, when the number of classes and the time of a trial is constant.

$$Acc = \frac{\#\text{ of correctly classified trials}}{\#\text{ of total trials}} \tag{2.32}$$

Acc is a straightforward measure but has some limitations based on the facts, that the classification accuracy of less frequent classes have smaller impact.

Cross-validation

Cross-validation is a split-validation technique where a dataset is partitioned in the following way: one subset is not used for model training, but reserved for the evaluation of the classification performance, resulting in a single entry of accuracy statistics. k-fold cross-validation divides the data into k subsets; $k-1$ subsets are used for classifier training and

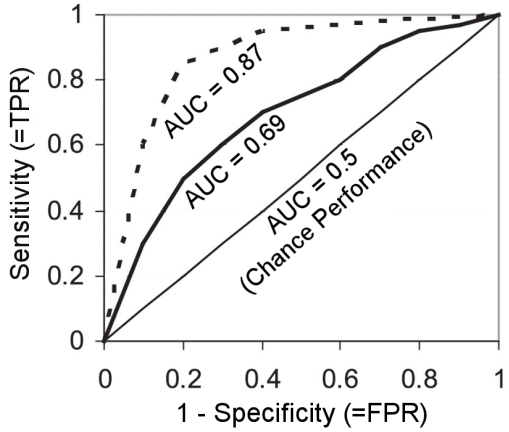


Figure 2.9: Three ROC curves representing an excellent, good and useless classifier for a binary classification problem, with the respective *AUC*-values.

the retained subset is used for classifier validation. This procedure is repeated k times so that every subset has been used once for testing and the other times for training. The final *Acc* result is reported as an average of all folds [98].

Bootstrapping

To provide a more robust statement of *Acc*, an approach that uses confidence intervals can be used, e.g., the non-parametric method bootstrapping [98]. This method can be used to assess variations in the estimated model accuracy. For bootstrapping no assumption is made regarding the populations of the input variables. Given a training set $\mathcal{D}$ of size n, bootstrapping generates m new training sets $\mathcal{D}'$ each of size n, by sampling from $\mathcal{D}$ uniformly and with replacement. Sampling with replacement is done multiple times to estimate the mean variability and variance of model outputs. Finally, the m models are combined by averaging the classification output.

ROC-Curve

When having an unbalanced number of trials, the receiver operating characteristics (*ROC*) is a more accepted measurement to evaluate the behavior of the classifier as using the *Acc*. The *ROC* illustrates the performance of a binary classifier system as its discrimination threshold is varied. The *ROC* is created by plotting the false positive rate (FPR) on the x-axis against the true positive rate (TPR) on the y-axis.

$$TPR = \frac{TP}{TP+FN} \tag{2.33}$$

$$FPR = \frac{FP}{FP+TN} \tag{2.34}$$

TP represents true positives, where FN stands for false negatives, FP means false positives, TN means true negative. To compare diverse classification models as well as to estimate the performance of the classier the Area Under the Curve (AUC) has to be calculated. The AUC is the area between the ROC curve and the x-axis (see Figure 2.9) which can be calculated by the definite integral. A classifier that separates the classes perfectly has an AUC equal to 1. Conversely a classifier that separates the classes not better than random guessing has an AUC close to 0.5. Thus, the higher the AUC value the better the performance of the classifier.

3 State of the art in workload classification and EEG-based tutoring systems

The state of the art in workload detection methods, as well as classification methods used for workload detection and prediction are reported in the following sections. This chapter is partially based on [8, 99, 100].

3.1 EEG-based measurement of workload during learning

As reported in section 2.3.5, EEG has widely been used for determining workload in individuals. Gevins and colleagues [50] did research on the influence of task difficulty in EEG signals and workload. In another study, Antonenko and Niederhauser [28] revealed differences in theta- and alpha- frequencies when reading hypertext with and without link previews. The researchers Gerlic and Jausovec [30] found that learning about planets from spoken text combined with music and pictorial information (i.e., high workload) yielded to alpha-desynchronization in temporal and occipital electrodes. Learning from written text alone (i.e., low workload), an alpha-desynchronization in frontal and central electrodes occurred. However, in all studies, it remains unclear, due to the complex learning materials used for experimentation, whether the observed EEG differences between more and less demanding learning materials really go back to differences in workload or whether they might be mostly artifacts of perceptual or motor differences between experimental conditions. These problems of perceptual-motor confounds seem inevitable when using standard EEG power analysis in comparing realistic learning materials varying in levels of difficulty, instead of comparing more controlled experimental tasks without perceptual-motor confounds.

To summarize the widespread literature, the reported findings suggest that alpha- and theta-frequencies are stable indicators of cognitive performance and are useful as parameter for detecting different workload states. All reported studies were analyzed offline. Furthermore it is unclear, whether the classifications are really based on differences in cognitive workload or on some of the perceptual-motor confounds of the different instructional conditions. Therefore, modified analysis and classification methods are necessary, to answer the residual questions.

3.2 Offline classification of workload levels based on EEG data

A continuous measurement of brain activity using EEG signals is possible and allows the estimation of subjects' cognitive workload without interrupting the learning process. By using appropriate classification methods [86, 101], estimations should be automated. To achieve this goal, efficient classification methods are necessary. In this section a short overview about common classification methods will be given, followed by modified classification techniques.

3.2.1 Common workload classification methods

To the day of writing, only scattered studies [102, 103, 104] demonstrated applications of BCIs addressing cognitive states. Having a closer look at the tasks used for the workload classification based on EEG data reveals, that problems of muscle artifacts and perceptual confounds seem to be quite common. Interestingly, these problems even occur in studies using low-level tasks from working memory research for classification.
Berka and colleagues [105] used Linear Discriminant Analysis (LDA) to classify different levels of workload in motor tasks, as well as in cognitive tasks. They used EEG data to distinguish four classes of "vigilance". An increasing workload leads to a classification into the high vigilance class for all types of tasks. However, as their classifier mainly relies on alpha-band power in combination with behavioral measures, it remains quite unclear whether the classifier is sensitive for increased workload, increased motor activity and eye movements, or both.
Similar arguments apply to the work of Chaouachi et al. [106]. They used two digit span tasks (requiring subjects to memorize sequences of digits of different length) and a logic task (requiring subjects to induce rules of different difficulties describing sequences of numbers like 2 - 4 - 6) for classification by means of Gaussian process regression. The NASA-TLX rating scale [19] was used to obtain subjective data on experienced cognitive workload. Their results yielded over 90 % accuracy with regard to the prediction which tasks were subjectively rated as simple, intermediate or difficult. Again, the difficulty levels of the used tasks were systematically confounded with the amount of motor activities, so that results cannot unambiguously be interpreted as classification of cognitive workload.
SVMs were used for workload classification by Putze and colleagues [107], as well as by Brouwer et al. [104]. The first work [107] recorded different physiological parameters (EEG, pulse, respiration) during accomplishing a driving simulator task. These parameters could be differentiated in diverse levels of workload using SVMs. However, even if there are no motoric confounds of task difficulty, there might be perceptual confounds. In the second study, Brouwer and colleagues [104] classified cognitive workload based on a letter n-back task, implemented in three levels of difficulty (0-back, 1-back, 2-back). No obvious perceptual-motor confounds, as well as no differences between difficulty levels with regard to eye-movements were detected. As features for classification, Brouwer et al. [104] used three frequency bands (theta, alpha and beta), on seven EEG channels on frontal, central and parietal brain areas. Applying a SVM for classification they were able to reach a classification accuracy of 85 % correct on average, for distinguishing 0-back and 2-back

tasks. However, as the 0-back task involves no updating process at all, it might not really be considered a working-memory task, compared to 1-back or 2-back tasks. Classifying these two tasks leads to classification results of approximately 75 % on average correct.
To conclude, the postulated studies showed cognitive workload classification is indeed possible in non-complex workload tasks with standard classification methods.

3.2.2 Modified classification methods towards zero training

A major assumption in many machine learning algorithms is that the training and testing data have to be in the same feature space and the same distribution. However, in many real-world applications, this assumption does not hold. The recording of calibration data is a time consuming prepositional step at the beginning of every new session. Especially, for subjects with low concentration ability, this initial calibration reduces the valuable remaining time in the application phase. But even for ordinary users, the calibration is an annoying procedure. Furthermore, the normal within-classification method cannot be transferred to realistic learning environments, since data from realistic learning tasks should not be used for training a classifier. Besides the time consuming calibration phase, realistic learning tasks are not reproducible as performance tasks as they induce learning effects. Considering these reasons, transfer learning [108, 109], cross-task classification [110, 111, 32], as well as cross-subject classification [112, 113] might be an appropriate approach to solve these challenges.

3.2.2.1 Transfer learning for workload detection and motoric control

In many real-world applications (e.g., in a learning environment), it is not possible to collect a large amount of training data. In such cases transfer learning would be desirable. Transfer learning uses the information contained in other subjects' training data [109]. The aim of transfer learning is to reduce the re-calibration effort, by adapting models trained in one time period to a new time period. Thus, it would be helpful if the classification knowledge could be transferred into a new domain.
Wu et al. [108] postulated a study where they combine transfer learning and active class selection to reduce the needed training samples and increase the classification accuracy. 18 subjects solved a virtual reality Stroop task. The authors trained a classifier to differentiate between three scenarios: scenarion I: low threat, color naming; scenario II: high threat, color naming; and scenarion III: high threat, Stroop inference). For classification, Wu and colleagues compared the performance of k-Nearest Neighbor (kNN) and SVM by using just a small amount of training samples of each individual subject and made use of additional data from other subjects to optimize the classifier parameters (e.g., kernel-parameter). 29 features, like spectral power of EEG signals at alpha-, beta- and theta-frequencies, as well as additional physiological measurements were used for classification. For individual subjects, a classification accuracy of 95.82 % was achieved, with 30 user-specific training samples. By using the postulated transfer learning method, they further demonstrated that the proposed approach can significantly reduce the number of user-specific training data samples. Transfer learning in combination with kNN or SVM

can save over 7 % primary training samples over the baseline approach.
In order to avoid the calibration session, Jin et al. [113] derived an online generic classification model by directly combining online user-specific training data and offline training data from other users. Eleven subjects had to write 20 characters by using a N400-Speller, which uses the event related signal in the time-domain as classification features. The study tested five classification conditions, whereas the (online) generic model is the most interesting. The model was calibrated from ten participants' data and tested on the 11th participant. In the online condition, the generic model was used in conjunction with an online training strategy, where the generic model is adapted to each individual subject after a small number of training trials. Twelve electrodes placed at the frontal, central and parietal brain areas by 35 points in time served as (12 x 36) features for classification. Experiments showed that 7 of the 11 participants were able to use the generic model during online training, but the remaining four could not. By using the generic model with and adapting to each individual subject with an online training strategy, the average classification accuracy over seven subjects reached 82.1 %. Even when the generic model with the online adaptation shortened the calibration process compared to a normal within-task classification, an online training phase for each subject is necessary.
An additional study with the aim of avoiding the calibration phase of a classifier is from Krauledat et al. [114]. Krauledat and colleagues used a movement imagination controlled BCI, where calibration and testing data of multiple sessions from six BCI-experienced subjects were employed to identify subject-dependent prototypical spatial filters. For each subject, data from a number of past sessions has to be available, where EEG data is recorded from each subject while fulfilling the same movement imagination task. Filtering the EEG-data from all past sessions of the individual subject resulted in a 12-dimensional feature space. For classification, a least square regression is accomplished. Applying this method, a feedback accuracy between 63 % up to 89 % could be achieved. On average, the classification accuracy reached 90 %. Even though Krauledat and colleagues name there method zero training, they report that ten trials per class from the same day are required for each subject to update the bias for the classification scenario.
With the reported methods, the amount of training data for classifier calibration is reduced, but primary data or calibration data of each subject is still necessary to train or adapt a classifier model for each subject individually. The tasks, which should be solved, has to be the same, during recording training as well as testing data. Furthermore, merely the first study used transfer learning for workload detection, whereas the other studies used transfer learning in an active BCI setting.

3.2.2.2 Cross-task classification for workload detection

Since data from realistic learning tasks should not be used for classifier training, the normal within-classification method cannot be transferred to realistic learning environments. Therefore, the aim of cross-task classification is to calibrate the classifier on a set of training tasks and to test it on an independent set of testing tasks.
Heger et al. [32] compared a resting state situation to situations were subjects were conducting workload imposing Flanker tasks (pressing the appropriately oriented arrow key

according the the central arrow of five displayed arrows like: >><>>) or switching tasks (pressing keys to decide whether a number is greater/smaller than 5 or whether a number is even/odd). They classified the neuronal signature of resting versus high workload, applying a within-task classification based on Artificial Neural Networks (ANN) with over 90 % accuracy. Additionally, they applied the classifier trained on this manipulation to realistic computer-based tasks of low (reading) versus high workload (writing). Motor activity was clearly a strong confound with cognitive load in this study (resting versus key pressing and reading versus typing). As beta- and gamma-frequencies were actually used for classification, this study mainly demonstrates a classification of motor versus no-motor activity and not of different levels of workload.
Another study reporting good within-task classification results for three types of working memory tasks (two levels of difficutly each) is the study by Baldwin and Penaranda [110]. They used a reading span task, a visuospatial n-back task and a Sternberg task for inducing different levels of cognitive workload. The classifiers used to distinguish the levels of workload relied on 50 features, namely the EEG power of five frequency bands (delta, theta, alpha, beta and gamma) obtained in ten channels (three frontal, three central, three parietal and one occipital). To classify the two levels of workload for all three tasks, an ANN was applied. The within-task classification resulted in high classification accuracies (on average 80 %). As in the studies reported before, these results might go back to some perceptual-motor confounds, as many difficulty conditions differed with regard to their motor activity. The classification accuracy dropped strongly when applying cross-task classification. Therefore, EEG data during solving a working memory task from one subject is used for classifier training and EEG data from the same subject is used for classifier testing during accomplishing another working memory task. For this procedure merely a classification accuracy around chance level was reached on average (smaller 52 %).
Gevins and colleagues [111] also used ANNs. They were trained on EEG data obtained during solving two types of simple working-memory tasks (verbal and spatial n-back). Each task was presented in three levels of difficulty (1-back, 2-back and 3-back tasks). While using 27 electrodes and four frequency bands (theta, alpha, beta and gamma), they were able to successfully discriminate between the three levels of workload for each of these tasks. Additionally, they used cross-task classification to distinguish between low and high difficulty levels across the two tasks and reported classification performances of 94 % on average. Although, these results are promising at first sight, it has to be noted that subjects basically had to solve the same task (n-back) in a verbal and a spatial version. Using two versions of the same task for cross-task classification is much simpler than the goal of using rather diverse tasks like a working-memory task and a complex learning task. In this thesis, the focus lies on cross-task classification across diverse tasks and not in classifying different versions of the same task.
The results of the stated studies are widespread. The time for classifier training can be reduced by using EEG data recorded during short working memory tasks for all studies. Only one study did cross-task classification by using a complex task for classifier testing, the others merely did cross-task classification on diverse working memory tasks or even for the same task in different versions.

3.2.2.3 Cross-subject classification for workload detection

For cross-subject classification, no data of the individual subjects are necessary. The subject can be a novice in using an EEG-system because the model is calibrated based on data from independent subjects and does not have to be calibrated or adapted to each individual. For many real-world scenarios (e.g., for adaptive learning environments) this means a great advantage, since it is not possible to collect a large amount of training data.
The previous introduced workload classifiers are subject-specific. For each subject, a classifier training or adaptation has to be done. In this section, three studies will be introduced, using cross-subject classification. Compared to transfer learning (see section 3.2.2.1) where auxiliary data (= labeled data from other users) is used for optimizing classifier parameters, the cross-subject classifier is trained with data samples from a group of subjects to generate a classifier model and tested on a new unknown subject.
In the study of Jin et al. [113], a generic classification model was used, without online training by adapting the model to user specific data (see section 3.2.2.1). The generic model was calibrated from ten participants' data and tested on the 11th participant. Despite section 3.2.2.1 no additional online adaptation was conducted. As stated in section 3.2.2.1 7 of 11 participants were able to use the generic model including online training with an average classification accuracy of 82.1 %. For these seven subjects an average cross-subject classification accuracy of 67.1 % was achieved. The accuracy obtained from the generic model without additional calibration phase for each individual subject was significantly lower than accuracies from the generic model with additional user specific online adaptation.
Gevins et al. [111] also applied cross-subject classification to differentiate between the three levels of difficulty induced by a 1-back, 2-back and 3-back task (verbal and spatial). As features for classification, the signal of 27 electrodes in the frequency range of theta, alpha, beta and gamma were used. Applying an ANN by training the network on a group of seven subjects and testing it on data from a new individual lead to classification accuracies of 83 % on average.
In the study of Wang et al. [112], eight subjects had to perform a multi-attribute task battery across three levels of difficulty. This task battery incorporates tasks analogous to activities that pilots perform in flight including a tracking task, monitoring gauges and warning lights, air traffic control communications and resource allocation tasks. The participants performed the task battery on five separate sessions spread over the course of a month. The models were trained and tested using a 5-fold cross-validation. Data was sampled across workload blocks and for the models including multiple subject data, evenly across subjects. By using a hierarchical Naive Bayes Model with 133 features calculated from 17 electrodes in the frequency range of $2\,\text{Hz} - 57\,\text{Hz}$, cross-subject classification accuracies up to 84 % on average were reached.
The method of cross-subject classification seems to be a stable and promising method for an efficient classification method, since a classifier which was trained once, could handle multiple subjects without further calibration data. Just one study did cross-subject classification in a passive BCI setting [111], but none of them used real learning material or adapted the presented material online, based on the measured EEG data.

3.3 Tutoring systems based on EEG data

The studies reported in the previous sections mainly deal with cognitive workload classification in well controlled working memory tasks (e.g., n-back, Sternberg task), simple mental addition tasks or active BCI settings. Only a few studies exist where EEG data was recorded while using intelligent tutoring systems (e.g., the reading tutor from the Carnegie Mellon University [115]). In this section, diverse tutoring systems will be introduced, where EEG data was recorded during solving the tasks.

3.3.1 Offline analysis of EEG data

Galán and Beal [116] were able to measure EEG signals from 16 participants while solving eight multiple-choice math problems, four easy and four difficult ones. The items were presented within an online tutoring system that recorded the response time, as well as the correctness of a given answer. The aim of this study was to predict the correct or incorrect outcome of a math problem, using EEG data and a SVM for classification. The average classification accuracies based on the workload index performed better on the easy problems with 83 %, compared to the difficult tasks with 67 %. A reason for the low predictions in difficult problems may be that subjects needed more time to solve these. Therefore, signals were longer and possibly more noisy. Thus, it cannot be ruled out that signals recorded during the difficult condition include more artifacts, e.g., eye-movements, than in the easy condition (the authors do not report about artifact removal from the EEG data).
In an additional study, Mostow and colleagues [117, 118] used EEG signals from adults and children during reading texts and isolated words implemented in the Project "LISTEN's Reading Tutor" [119]. The subjects had to read out loud three easy and three difficult passages in alternating order. Subsequently, a silent reading condition followed, with three easy and three difficult passages, using different texts as in the previous condition. Mostow et al. trained and tested a binary logistic regression to estimate the probability that a given sentence was easy or difficult and to distinguish among easy words, difficult words, pseudo-words and unpronounceable strings, based on EEG data. For brain signal recording, only a single sensor from the NeuroSky "MindSet", placed on the forehead, was used. While using one electrode and alpha-frequencies at 8 Hz a within- as well as a cross-subject regression was accomplished. Average accuracies from about 43 % to 69 % for the within-subject classification and 41 % to 65 % for the cross-subject classification were reached for separating easy from difficult texts. Further, they calculated the rank accuracy in classifying words in easy, difficult, pseudo or illegal and reached accuracies from about 45 % to 58 % for within-subject classification and about 39 % to 59 % for cross-subject classification. The results were significantly better than chance in predicting the difficulty of the text. A single electrode from a portable recording device can differentiate between reading easy and difficult sentences across populations (children and adults) and modalities (oral and silent reading). The low classification results can be caused due to bad signal quality of the NeuroSky headset. To conclude, the results seem promising, but were not applied in an online setting so far.

3.3.2 Online analysis of EEG data

Berka and colleagues [120] calculated the amount of workload for five subjects in real time by using the Aegis simulation environment. The aim of this study was to assess the feasibility of accurately detecting and quantifying an EEG indicator of cognitive workload while solving a highly complex, cognitively challenging task. They used a battery-powered wireless EEG sensor headset to record six channels placed on frontal, central and parietal brain areas. During the task execution, a workload level ("high vigilance", "low vigilance", "relaxed wakefulness" and "sleep onset") was calculated for each second of EEG data, nearly in real time, by using LDA with the alpha-frequency band as feature. The classification model was trained prior to the study, on data from each subject acquired in sleep deprivation studies. Difficult cognitive load tasks resulting in a high EEG-workload level were detected with an efficiency of up to 100 %. These results are promising for an intelligent closed-loop system using EEG data to adapt tasks and streamlining a user's cognitive workload. By ensuring an operator remains uninterrupted during extreme workload periods, such an approach could result in an increased productivity of the user and reduced errors [120].
Furthermore, Stevens and colleagues [121] were able to measure cognitive workload changes in real time, while twelve subjects solved biological problems in a learning environment. Therefore, they recorded six channels placed on frontal, central and parietal brain areas using the same wireless EEG sensor headset as Berka et al. [120]. Despite from Berka and colleagues [120], Stevens et al. [121] used competitive, self-organizing ANNs to detect workload changes, but an online adaptation based on the classifier output was not conducted.

3.3.3 Online adaptive learning environment based on workload detection

The studies presented so far demonstrated experiments, where workload could be detected offline, as well as online, while using intelligent tutoring systems. Adapting an intelligent tutoring system based on the outcome of the classifier has not yet been implemented. There are a few studies, adapting video games online based on diverse workload levels or affective states, identified in EEG data [122, 123]. Furthermore, only some studies are dealing with real time adaptive tutoring systems reacting on affective states, detected in EEG data [124, 125, 126] to support students in there learning process. To the best of my knowledge, there are no studies dealing with online workload detection based on EEG data and complex learning material adaptation.

4 Basic ideas, hypotheses and objectives of this thesis

The aim of this thesis is to develop a non-obtrusive, objective, online workload detection method, which can be used to adapt instructional material in a learning environment in real-time. The motivation for this work is to support people optimally in their learning process. Mainly, learners with special needs, e.g., test anxiety, learning disability or attention deficit disorder can benefit from such an individualist system. Furthermore, different cognitive workload capacity ranges can be considered while adapting the learning material.
In this chapter, the motivation for developing an adaptive learning environment, based on EEG data, will be pointed out. The main challenges and objectives, that are important prerequisites for developing efficient adaptable computer based learning environments, will be reported. Furthermore, the hypothesis leading to each of the following studies will be outlined. Finally, the connection of the applied studies will schematically be shown and described.
The main research questions are:

1. How does suitable learning material for an adaptive learning environment look like?
2. What kind of features lead to a precise workload detection in EEG signals?
3. Can generalizable classifiers be developed for an accurate workload prediction?
4. Is a precise workload prediction across subjects possible in real-time?

4.1 Objectives of this thesis

Suitable learning material for an adaptive learning environment

During this thesis, various types of learning materials will be used, to find out which influences these differences have on the brain signals. The design ranges from solving complex text tasks, to reading easy comic-strips. The presentation varies from black and white to colorful images. The learning context ranges from well defined working-memory tasks, to well structured addition tasks, to complex mathematical algebra assignments. In the following studies it will be researched, which learning material is most suitable, to manipulate workload states without additional confounding factors. Furthermore, it will be evaluated which learning material is most appropriate to be used in an online learning environment, which will be adapted based on EEG signals.

Features for workload detection in EEG data

To find appropriate features representing different cognitive workload states in EEG signals across tasks and subjects, a variety of pre-processing and feature extraction methods will be applied. Thus, e.g., connectivity or ICA will be utilized to evaluate how suitable these different features are for workload detection and classification.

Generalizable classifiers for workload detection

As already stated in section 3.2.2 common within-task classification is not feasible in a real-world learning environment. The calibrated classifier should be able to detect different workload states across various tasks. The collection of training data used for classifier calibration in combination with a learning environment is challenging. A high number of training trials is needed for classifier training. For that reason, the EEG data recording is time consuming for participants. A long time period for classifier calibration can cause frustration or loss of motivation, which are crucial factors for solving learning tasks successfully. Furthermore, complex learning material cannot be used as training data for classifier calibration. Thus, subjects will develop solving strategies in the calibration phase before solving the actual learning phase. Due to the increasing knowledge during the classifier calibration phase, the amount of required workload for the same degree of difficulty will decrease over time. Therefore, EEG patterns used for classifier training are not reproducible for classifier testing. To counteract these challenges, generalizable classifiers are desirable for adaptive learning environments. Thus, two modified classification methods will be introduced and investigated in this thesis: cross-task classification and cross-subject regression.

In the first method, cross-task classification, data from the same subjects but different tasks are used to calibrate and test the classifier. Thus, using workload states induced by easy working memory tasks can enable the prediction of workload during more complex learning tasks. Furthermore, the influence of subjective cognitive workload labeling and the task order effect during cross-task classification will be evaluated.

In the second method, cross-subject regression, the regression models will be trained using data from the same tasks but different subjects. By utilizing a leave-one-subject-out validation, a precise workload prediction across subjects can be reached.

Online workload prediction across subjects

All findings and discoveries of the previous studies will be combined in the last part of this thesis. Suitable learning material and generalizable classifiers will be applied in real-time, to determine if a precise online workload prediction across subjects is possible. Furthermore, it will be analyzed if learning material can successfully be adapted online, based on the predicted workload state, to keep each learner in his/her optimal workload capacity range. Such an adaptive learning system will support learners in their learning process successfully. Finally, the newly developed EEG-based adaptive learning environment will be compared with a state of the art tutoring system to assess the performance and learning support.

4.2 Interrelation of the performed studies

In Figure 4.1 the underlying research question, as well as the dependencies of the performed studies, reported in the following chapters, is shown.
Designing the optimal learning material for an EEG-based learning system is the starting point for each study. Therefore, research question 1: *"How does suitable learning material for an adaptive learning environment look like?"* is treated in several chapters (5, 7, 10 and 11), while the obtained knowledge of each study influences the task design for the subsequent studies.
"What kind of features lead to a precise workload detection in EEG signals?" is research question 2. Since suitable neurophysiological features are key prerequisites for successful classification results, different feature selection and extraction methods are comparatively shown in chapter 6. These findings were considered for data analysis in the following studies, reported in chapter 7ff.
An additional essential research question for developing an efficient and realizable learning environment is question 3: *"Can generalizable classifiers be developed for an accurate workload prediction?"* This research question is extensively examined in chapter 7, 8, 9 and 10, where workload prediction is realized by using cross-task classification or cross-subject regression. Further analysis as subjective cognitive workload labeling and task order effects are analyzed while using cross-task classification.
"Is a precise workload prediction across subjects possible in real-time?" is research question 4 and the goal of this thesis. Thus, the findings from all prior studies regarding task design, suitable neurophysiological features, as well as workload prediction methods were considered for the last study, reported in chapter 11.

The innovation and novelty of this thesis is: using an objective, non-obtrusive adaptation measurement, which can ensure that a learner remains in their optimal workload capacity range. To the best of my knowledge, there are no studies dealing with online workload detection based on EEG data and complex learning material adaptation, keeping a learner in his/her optimal workload capacity range. The results from previous studies are promising for using EEG data to adapt tasks. Thus, it seems advisable to develop an EEG-based adaptive learning environment, providing individual technological support for learners. Therefore, this learning environment adapts learning materials to learners' levels of expertise and workload capacity, to support them in their learning process successfully.

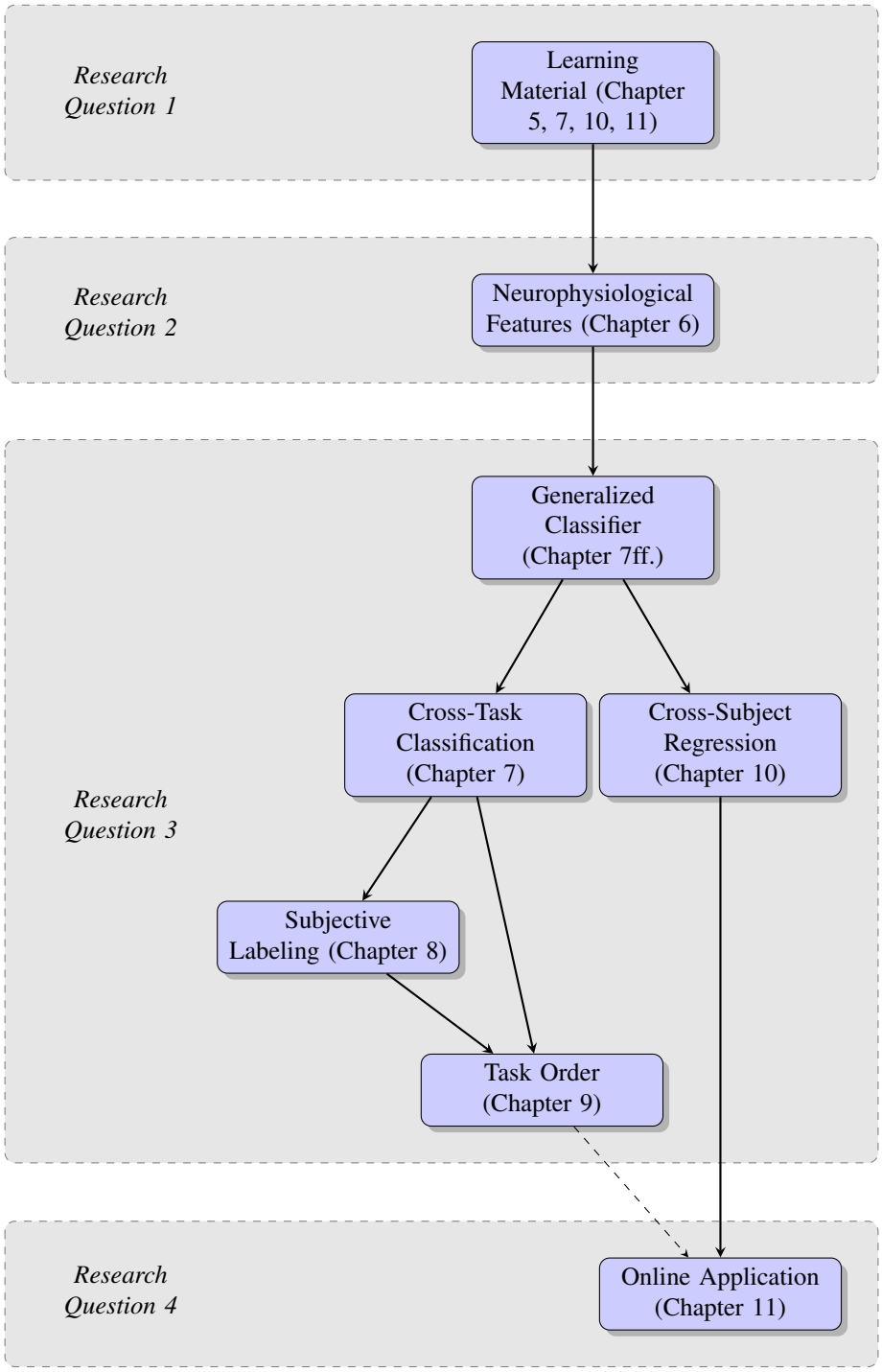


Figure 4.1: Interrelation of the performed studies in this thesis.

5 Workload classification using complex learning material

In the following sections, it will be explored, if complex learning material imposes different levels of cognitive workload which could be measured by using EEG data. This chapter deals with classifying different cognitive workload levels during learning and non-learning tasks with the help of EEG data from students. The research question was, if single-subject single-trial EEG data allows for classification and which features are used. The results presented in this chapter were partially published in [127].

5.1 Study design

An introduction in the participants data, the EEG recordings, the task design as well as data analysis will be given in the following sections.

5.1.1 Participants and EEG recordings

Subjects were ten German students (12 – 15 years old, mean age: 13.6, 6 female, 4 male) without prior knowledge of angular geometry who participated voluntarily in the experiment. None of the subjects had participated in an EEG study before. A set of 16 passive electrodes, attached to the scalp, were used to record EEG signals. The electrodes were placed according to the International Electrode (10-20) Placement System. An electrooculogram (EOG) was recorded through three additional electrodes: two placed horizontally at the outer canthi of both eyes and one placed in the middle of the forehead between the eyes. EEG and EOG signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp Generation 3.0). The sampling rate was 256 Hz and the impedance of each electrode was $< 10\,\mathrm{k}\Omega$. The EEG data was high-pass filtered at 0.1 Hz and low-pass filtered at 100 Hz during the recording. Furthermore, a notch-filter was utilized between $48\,\mathrm{Hz} - 52\,\mathrm{Hz}$ to filter the noise of the power line.

5.1.2 Paradigm

The experiment is a within-subject design and comprises of four phases. First, the subjects had to solve a pre-test to assure they had no prior knowledge of angular geometry. The pre-test is composed of four main tasks with 30 sub-tasks in total. The participants had to solve these tasks with no time constraints. Phase 2 consisted of three learning cycles, 11 minutes each. In each cycle, the subjects were asked to study five angle theorems and watch five comic-strips (see Figure 5.1). Each theorem and comic-strip was presented for

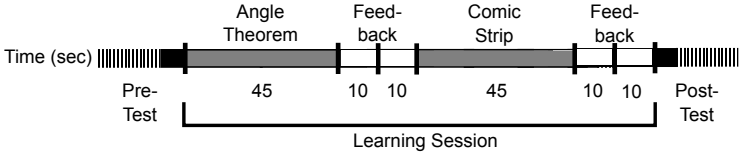


Figure 5.1: Chronological stimulus presentation during the learning session.

45 sec in an alternating order. By studying difficult material like angle theorems, high levels of workloads were induced. On the other hand, studying easy materials as reading comic-strips caused low levels of workload. Furthermore, the comic-strips were used, since the stimuli despite from the learning stimuli should also induce continuous eye-movements and require cognitive processes. All participants studied 15 angle theorems (3×5) and 15 comic-strips. In phase 3, students applied the theorems to geometrical exercises with a German version of the Carnegie Learning's Cognitive Tutor provided by Schwonke et al. [128]. Finally, a post-test had to be accomplished, where participants had to solve the same problems as in the pre-test, to achieve a direct comparison of their knowledge before and after the learning cycles. The whole experiment took approximately three hours for each subject.

5.2 Data pre-processing

EEG data was collected during the two types of stimuli, while solving angle theorems and reading comic-strips. The aim was to detect and extract features in the EEG frequency bands, with which the EEG data recorded during geometry tasks (inducing high workload) could mostly be distinguished from EEG data recorded during reading comic-strips (inducing low workload). Because of the low number of 30 trials (15 per type), all study windows were segmented in consecutive epochs of 15 sec for each subject. Accordingly, this resulted in 45 trials for each subject in each type of study window (3 epochs per study window $\times 5$ study windows per learning session $\times 3$ learning sessions), which were used for classification. This segmentation was possible because no significant variation was detected in the signal over a full trial. For spectral analysis, an ARM was calculated with the Burg-Algorithm (see section 2.4.1). Further, the EEG data was scaled with the *z-score* function from Statistics Toolbox in MATLAB.

5.3 Data analysis and classification

Before classifying, a visual analysis of the EEG data was made to determine whether differences in EEG data during learning and non-learning stimuli are distinguishable.

5.3.1 EEG data analysis

Data was visually analyzed using BCI2000 Offline Analysis, an open source MATLAB toolbox developed for neurophysiological data analysis. This toolbox is suitable to detect

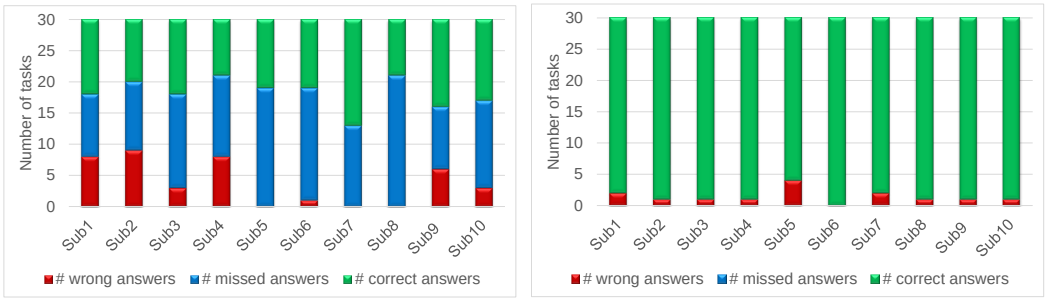
features in EEG data which can be used for classification. For the visual analysis, a power spectral density plot was generated and r^2-values of the two conditions (high workload, low workload) were calculated.

5.3.2 Machine learning algorithms

Machine learning algorithms are essential tools for an online classification of low and high cognitive workload during solving a task, which is the aim of this thesis. In this study, a SVM with RBF-kernel was used to classify mental states in EEG data during solving angle theorem tasks and reading comic-strips. As postulated in Lotte et al. [87], the RBF-kernel is a promising kernel for analyzing EEG data. Therefore, the toolbox LIBSVM [89] was used. The kernel parameter $\gamma = 0.5$ had to be specified by the user before training. A 10-fold cross validation was accomplished to verify the accuracy of the trained SVM. The selected features were focused on frontal and parietal electrodes with regard to the spectral power within the alpha ($8\,\text{Hz} - 13\,\text{Hz}$) and theta ($4\,\text{Hz} - 7\,\text{Hz}$) frequency band. The accuracy served as a quality criterion for the frequency interval and the specified electrodes.

5.4 Learning outcome

Each individual subject had to perform a pre-test before the learning phase started. This was used to assess the prior knowledge of the user. Furthermore, a post-test was accomplished after fulfilling the learning phase, to measure the learning effect. In Figure 5.2 the statistics of the pre-test, as well as of the post-test are shown for each subject. The students solved on average 11.8 tasks from 30 pre-test assignments correctly, which corresponds to an average accuracy of 39.33 %. In the pre-test, the best subject solved 17 tasks, whereas merely nine assignments were correctly answered by the worst subject.
After performing the learning session, the post-test was performed. Compared to the pre-test, an increase of correctly solved tasks is detectable for each subject. On average 28.6 from in total 30 post-test tasks were solved correctly, which corresponds to an average accuracy of 95.33 %. The best subject solved 100 % of the post-test tasks correctly, while the worst subject reached an accuracy of 86.67 %. On average, a total of 55 % more tasks were correctly answered in the post-test compared to the pre-test.
The learning effect of each individual subject after completing the learning phase was significant for each subject ($p < 0.01$, ANOVA). Furthermore, the number of skipped tasks was significantly higher in the pre-test compared to the post-test ($p < 0.05$, F-test), where no tasks were untreated. Because of the significant results, it can be assumed that all subjects have learned how to use angle theorems and attended conscientiously in the experiment.



(a) Pre-Test

(b) Post-Test

Figure 5.2: Performance data of each subject while solving the **a)** pre-test and **b)** post-test for analyzing the learning success. On the *x*-axis are four pillars for each subject, representing the total number of tasks, as well as the amount of wrong, correct or missed answers.

5.5 Neurophysiological features

To detect the frequency band and the electrodes which had to be used for classification, a visual analysis of the frequency power was done first. In the topography plot, the squared correlation coefficient r^2 between the power at 12 Hz and the workload of the corresponding trials were calculated offline. The topography plots help to estimate at which electrodes the EEG is most affected by difficulty and thus suitable for classification. Analyzing Figure 5.3, it is remarkable that task difficulty (i.e., studying angle theorems and watching comic-strips) and thereby the individual's workload capacity is reflected in the brain signals. A difficulty related effect over the parietal, occipital electrodes (Oz) can be recognized at 12 Hz, averaged over all subjects. The r^2-value is highest for this brain area in the alpha-frequency band, with $r^2 = 0.025$. This observation is in line with Kramer et al. [25]. They hypothesized brain activity in the parietal, occipital brain areas are based on cognitive workload needed for math problem solving. The effects measured over the right central brain areas (electrode C4) might be caused due to motor artifacts, as the difficulty effect seems to be located over the motor cortex. Furthermore, Figure 5.3 shows the spectral power averaged over all subjects. The continuous blue line represents the energy as a function of frequency during stimuli inducing high workload (angle theorems) and the dashed red line during stimuli inducing low workload (comic-strips). From the acquired results, it can be inferred that the alpha-band activity (10 Hz – 12 Hz) decreases during angle theorem tasks which require higher cognitive workload than comic reading tasks, as Klimesch [17] postulated. Based on the visual analysis, both stimuli types and thus both workload levels could be separated. These are essential prerequisites for a successful classification.

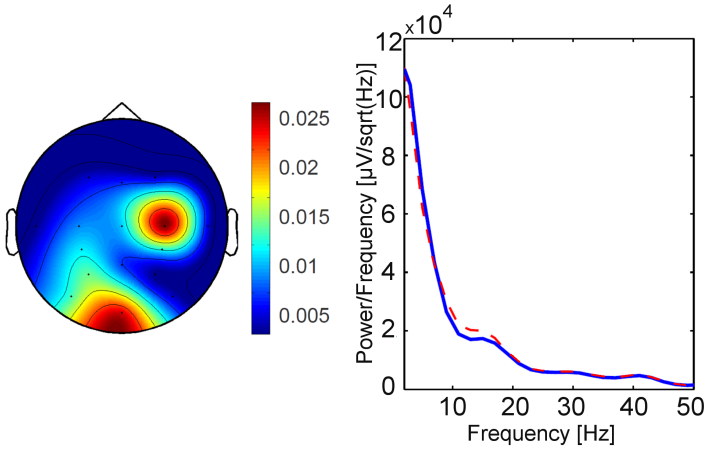


Figure 5.3: Left: Topography plot at 12 Hz averaged over all subjects with the highest $r^2 = 0.025$ at right central and occipital electrodes. **Right:** Power spectral density plots showing a desynchronization in the alpha-frequency band (8 Hz – 13 Hz) for angle theorems causing high workload (blue line) compared to comic-strips inducing low workload (red dashed line).

5.6 Offline classification results

The research question for this study was, if single-subject single-trial EEG data is applicable for classification and which features are used. Regarding the features, first assumptions could be made based on the visual analysis. Since the boundaries of frequency bands are not exactly definable between subjects, a classification over all subjects was utilized to define the most promising frequency range for individual classification. The maximal classification accuracy, averaged over all subjects reached 74.55 % classification accuracy for the frequency interval 6 Hz – 12 Hz and 4 Hz – 10 Hz, as shown in Table 5.1.
For individual classifications, the frequency interval was defined for each subject individually, since the boundaries of frequency bands are not exactly definable between different subjects. Because geometrical processing is located in the right hemisphere [129], electrodes were reduced on the left to reduce the feature space, thus 10 electrodes (Fz, Cz, C4, Cp4, P3, Pz, P4, PO7, PO8, Oz) were used as features, to differentiate two types of workload levels. In Table 5.2 the classification results of each individual subject are reported.

Table 5.1: Classification accuracy in % over all channels (column: starting frequency; row: ending frequency) averaged over all subjects.

	4 Hz	6 Hz	8 Hz	10 Hz
6 Hz	67.27	–	–	–
8 Hz	69.09	65.45	–	–
10 Hz	74.55	73.64	66.82	–
12 Hz	73.64	74.55	67.73	62.73
14 Hz	71.82	73.64	70.91	61.36

Table 5.2: Classification accuracy of differentiating high workload induced by angle theorems and low workload induced by reading comics in %. Results are shown for each subject, using individual frequency bands and a fixed number of electrodes (Fz, Cz, C4, Cp4, P3, Pz, P4, PO7, PO8, Oz) as features.

	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub9	Sub 10
Freq [Hz]	6-14	6-12	6-12	6-10	8-14	6-12	6-12	6-14	8-14	6-8
Acc [%]	90.00	77.78	81.11	91.67	70.00	84.44	82.22	73.33	63.33	84.44

The best subject reached a classification accuracy of 91.67 % for the frequency range of 6 Hz – 10 Hz, whereas the worst candidate merely solved 63.33 % for the frequency range of 8 Hz – 14 Hz.
These classification results lead to the assumption that it is indeed possible to differentiate the brain activity of both conditions inducing low or high cognitive workload on a single-subject basis by using a SVM.

5.7 Discussion

The aim of this study was to detect diverse workload levels during solving several tasks and to classify these phases utilizing EEG data. During the high workload condition subjects had to study angle theorems, while during the low workload condition they had to watch comic-strips. A desynchronisation of the alpha-band activity during learning geometrical theorems in comparison to comic-strips was detected in the right parietal, occipital brain area. This is consistent with the literature [17, 129], which shows that geometry is processed in the right hemisphere. These differentiations were successfully classified by using a SVM with RBF-kernel. For 6 out of 10 subjects, the mental states during high workload and low workload tasks were classified with an accuracy of at least 80.00 %. For the classification over all subjects using 16 electrodes and a frequency band from 4 Hz – 14 Hz a loss of accuracy was observed, reaching 74.55 %. Imprecise frequency band boundaries and individual cognitive treatments of each subject can cause insignificant differences between cognitive workload states over all subjects. By analyzing the learning effect factor, it can be seen, that each subject learned how to use the five angle theorems. Further, it can be assumed, that they attended conscientious in the learning phase. For this setting it is advisable to train each classifier for each subject individually, to get the best classification accuracy to guarantee an optimal workload prediction for each subject individually.

5.7.1 The difficulty of workload detection in EEG data

Since the literature about workload detection in EEG data is ambivalent, this first study was conducted. The results of this study are according to the hypothesis, that a difference in the right parietal occipital brain area is measurable between both tasks [25] and a decrease in the alpha-frequency band is detectable during high workload phases [17]. Some research groups postulated a decrease in the alpha-frequency band [17, 130], while others reported an increase in the alpha-frequency band [131], again others [51, 111] found the

beta-frequency band being important for workload detecting in the EEG data. A reason for this widespread findings could be individual variations in the brain activity, as well as diverse strategies which are processed in distinct brain areas, to solve such complex learning tasks. If learning tasks are too complex, it is not controllable which strategies learners are using to handle the assignments. Diverse strategies lead to significant differences in active brain regions, as well as varying patterns in the frequency bands. Thus, the more complex the task, the more important is an individual analysis for each subject, with customized feature extraction, as well as individual classifier training.

5.7.2 Confounds in EEG data induced by complex material

The results show, it is indeed possible to classify whether learners study realistic instructional materials or if they are reading comic-strips, based on single-subject EEG data. Although the results are practically and methodologically interesting, it cannot be ensured that the features used for classification are not confounded with perceptual-motor artifacts. The angle theorems and comic-strips were perceptually not identical, potentially leading to differences in semantic processing or eye-movements. Thus, even if there are no obvious motor confounds of task difficulty in this study, there might be nevertheless perceptual confounds, which might be picked up by the SVM. Furthermore, it is uncertain whether workload classifiers trained on realistic learning tasks really represent a measure of workload.

5.8 Conclusion

The classification of two mental states in EEG data during learning angle theorems (high workload) and reading comic-strips (low workload) using machine learning algorithms has been investigated. Although the results are promising it cannot be assumed, that workload classifiers trained on realistic tasks really represent a measure of workload. Diverse strategies, perceptual-motor confounds and semantic processing can lead to significant differences in the EEG data across subjects, as well as across the two types of presented material. An individual feature selection and classifier training is advisable, which is not feasible in real-world settings. Furthermore, the presented stimuli have to be revised for the next studies, so that the induced EEG data is comparable across stimuli and subjects.

6 Feature selection for workload classification

The first study showed that EEG signals can be used for classifying cognitive workload states during two complete diverse stimuli. Although the results are promising, it cannot be assumed, that workload classifiers trained on realistic tasks really represent a measure of workload. Therefore, a second study will be introduced in this chapter, where the detection of characteristics in EEG signals with high consistency over all subjects will be focused. The presented stimuli were revised, so that the induced EEG data is comparable across stimuli and subjects. The characteristics in the EEG signals should validly and reliably represent cognitive workload states during solving math-tasks in varying difficulty levels. Furthermore, various signal pre-processing steps as well as diverse feature extraction methods were implemented and calculated, to get the optimal preferences for further studies and analyses. The results presented in this chapter are exemplary for individual subjects and were partially presented and published in [132, 133].

6.1 Study design

In this section, a short introduction to the participants data, the task design, as well as the procedure will be given.

6.1.1 Participants and EEG recordings

In the experiment, seven subjects (4 female, 3 male) in the age of 22 – 28 years with a mean age of 25.3, participated voluntarily. A set of 32 active electrodes (actiCap BrainProducts GmbH) attached to the scalp, were used to record EEG signals. 29 electrodes were placed according to the extended International Electrode (10-20) Placement System. EOG was recorded through placement of the three remaining electrodes: two placed horizontally at the outer canthi of both eyes and one placed in the middle of the forehead between the eyes. EEG and EOG signals were amplified by two 16-channel biosignal amplifier systems (g.USB Generation 3.0, gTec). The sampling rate was 512 Hz and the impedance of each electrode was $< 5\,\mathrm{k}\Omega$. EEG data was high-pass filtered at 0.1 Hz and low-pass filtered at 100 Hz during the recording. Furthermore, a notch-filter was disposed between 48 Hz – 52 Hz to filter the noise of the electrical circuit.

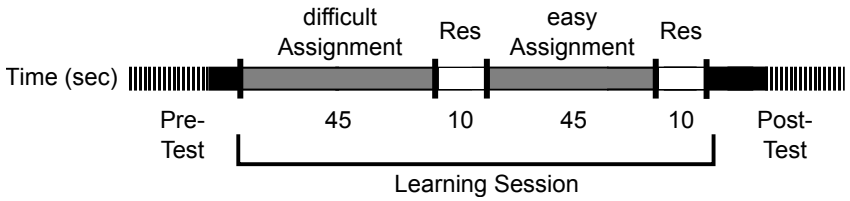


Figure 6.1: Schematic display of one EEG-trial during a learning session (Res = response).

6.1.2 Paradigm

The experiment has a within-subject design, which aimed at differentiating between high and low workload by means of EEG data. Participants were asked to learn angle theorems and to read information of varied complexity from diagrams. The experiment comprised three phases. In phase 1, the subjects had to solve a pre-test to survey their knowledge of angular geometry and diagrams prior to the study. Phase 2 consisted of four learning cycles, 17 min each. In cycle 1 and 3, the participants were asked to study four different angle theorems provided by Schwonke et al. [128]. Each angle theorem was alternatingly presented in an easy and a more complex way. The easy presentation supported the learning process of each learner with color-coding the angles and not using split attention effects. In the difficult presentation mode, the angles were black and white with no color-coding. Furthermore, a split attention effect was integrated. In cylce 2 and 4, the participants were asked to work with information of four different types of diagrams, in an easy and a more complex way (based on Schuh et al. [134]). In the easy condition, subjects only had to read out a number. During the difficult presentation mode they had to compare two relations. Each angle theorem and diagram was presented for 45 sec. Each cycle started with the more complex stimulus, followed by the same stimulus in an easier way. Subsequently, the next difficult stimulus appeared. The temporal sequence of one EEG trial is depicted in Figure 6.1. All in all, each participant studied 8 difficult and 8 easy angle theorems and worked with 8 complex as well as 8 effortless diagram information. Finally, a post-test had to be accomplished in phase 3. Participants had to solve the same assignments as in the pre-test, to achieve a direct comparison of their knowledge before and after the learning cycles. The whole experiment took approximately two hours for each subject.

6.2 Data pre-processing

For further steps, EEG data during the four learning and application cycles of each subject were concatenated. This resulted in two datasets. Session 1 and 3 were concatenated to one long “angle-theorem” session. Session 2 and 4 yield one long “diagram” session. The data was down sampled from 512 Hz to 300 Hz to speed up the data analysis. Linear drifts of the baseline were removed in the EEG signals, with the application of a detrend function. Furthermore, a Butterworth-Filter was used as high-pass filter at 0.5 Hz as well as low-pass filter at 48 Hz. Additionally a Common Average Reference Filter (CAR) was applied. CAR

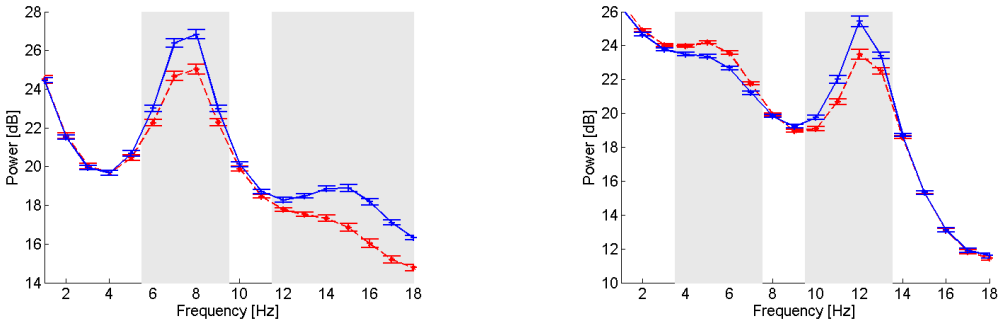
is the most suited re-reference method for EEG data [135, 136]. The EOG signals were used to detect and remove blinks, as well as eye-movements. The other accessory artifacts were also manually removed using ICA with the runica-algorithm [137]. With the aid of an experiment protocol, it was ensured that the noise in the data came from head movement or other artifacts. Neuronal activities were unaffected by filtering. Because of the low number of trials, the 45 sec long artifact free recordings were segmented in 15 consecutive epochs of 3 sec for each subject and each stimulus. For each subject and each condition (easy/difficult angle theorem; easy/difficult diagram), 120 ($= 8 \times 15$) trials occurred, which were used for classification, after pre-processing and visual analysis. This segmentation was possible because no significant variations were detected in the signal over a full trial. The variations over time were deconstructed by calculating a spectrogram and analyzing the power values over time during each trial for each channel.

6.3 Data analysis

EEG data was visually analyzed to detect features, that can be used for classification. FieldTrip [138], an open source MATLAB-toolbox, was used for analyzing the data. First, the power spectrum was generated by using autoregressive coefficients, whereby the ARM with model order 25 was calculated with the Burg-Algorithm [139] for each type of assignment (angle theorem; diagram). By calculating and plotting power spectra, the maximal differences in frequency bands between the two difficulty conditions were prospected. To detect statistically significant differences in the power spectrum of the easy versus the difficult condition, the p-value was calculated using the dependent samples t-statistic. Further, r^2-values were calculated for the easy and difficult angle theorems, as well as for the easy and difficult diagram tasks. The feature extraction method was utilized to detect the electrode positions, where the most disparities between easy and difficult stimuli occurred. As this study focused the detection of characteristics for cognitive workload in EEG signals, the common repertory of feature extraction methods were extended with an ICA and dipole fitting to identify individual neural sources underlying the scalp EEG. Furthermore, connectivity measures as coherence and Granger causality were utilized, to assess underlying networks in the brain. In the following sections, merely the results of participants 4 and 6 are exemplarily presented, as they illustrate the functionality of each method best.

6.3.1 Power spectrum for workload detection

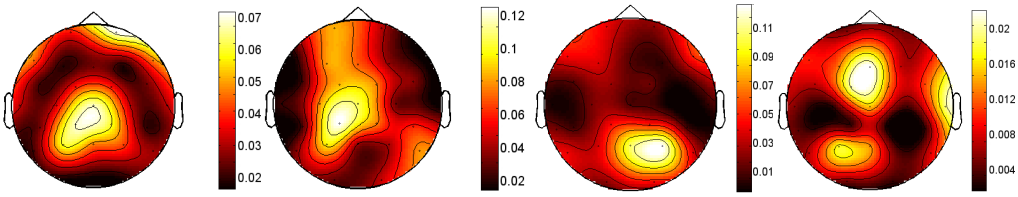
The power spectrum can be used to detect precise changes in the frequency bands. Figure 6.2 (a) shows the power spectrum plot of subject 4 during calculation of angle theorems and Figure 6.2 (b) displays the energy as a function of frequency of subject 6 while solving diagram tasks.
For subject 4, calculating angle theorems (Figure 6.2 (a)) lead to a significant ($p < 0.05$) decrease of the alpha-band activity (5.5 Hz – 9.5 Hz) during solving difficult material compared to easy material. For the theta-band activity, no significant differentiations could be detected. Consistent over the majority of subjects, a desynchronization of the alpha-



(a) Subject 4 while solving angle theorem tasks.

(b) Subject 6 while solving diagram tasks.

Figure 6.2: Power spectra plots show the energy as a function of frequency for subject 4 and 6, using 10 electrodes (F3, Fz, F4, FC9, FC10, CP1, CP2, P3, Pz, P4); Power spectrum during solving easy material inducing low workload (blue line), as well as the power spectrum while accomplishing difficult material inducing high workload (red dashed line). The grayed area in each chart indicates significant differentiations ($p < 0.05$) between the two conditions of difficulty.



(a) $5-7\,Hz$ (b) $12-14\,Hz$ (c) $4-6\,Hz$ (d) $8-13\,Hz$

Figure 6.3: r^2-values visualized as topography plots. **(a)** Theta-frequency band and **(b)** alpha-frequency band from subject 4 while calculating angle theorem tasks. **(c)** Theta-frequency band and **(d)** alpha-frequency band from subject 6 while solving diagram tasks.

frequency band could be detected while solving difficult angle theorems, compared to the easy ones. Although, the power spectrum trend is more widespread. Subject 6 is exemplarily shown as the best subject for solving diagrams (Figure 6.2 (b)). Solving diagram tasks leads to a significant ($p < 0.05$) increase of theta-band activity ($3.5\,\text{Hz} - 7\,\text{Hz}$) as well as a significant ($p < 0.05$) decrease of alpha-band activity ($9.5\,\text{Hz} - 13.5\,\text{Hz}$) during difficult material in comparison to the signal during easy material. Consistent with the literature, in 5 out of 7 subjects, the theta-band activity increases and the alpha-band activity decreases while solving diagram tasks, which require high cognitive workload.

6.3.2 Localities specified with r^2-values

Calculating r^2-values enables the detection of electrodes, which are qualified as features for the ongoing classification. Figure 6.3 exemplarily shows the r^2-plots for theta- and alpha-frequencies of subject 4 while calculating easy and difficult angle theorem tasks and subject 6 while solving easy and difficult diagram tasks. These two subjects are exemplary

shown since their EEG data indicate clear differences between the two levels of difficulty in the particular task. The r^2-values for subject 4 while solving angle theorem tasks is highest in the theta-frequency band in the central parietal brain ares, with $r^2 = 0.0625$. In the alpha-frequency band, the highest r^2-value ($r^2 = 0.12$) was detected in the more left central parietal brain area (see Figure 6.3 (a), (b)). For subject 6, the highest r^2-value ($r^2 = 0.11$) was measured in the theta-frequency band in the right parietal, occipital brain area, while solving diagram tasks. In the alpha-frequency band, the highest r^2-value could mainly be detected in the frontal, central brain area, with $r^2 = 0.02$ (see Figure 6.3 (c), (d)). Based on the presented results, signals from electrodes F3, Fz, F4, CP1, Cz, CP2, P3, Pz, P4 are advisable to use for classification.

6.3.3 Connectivity as additional workload indicator

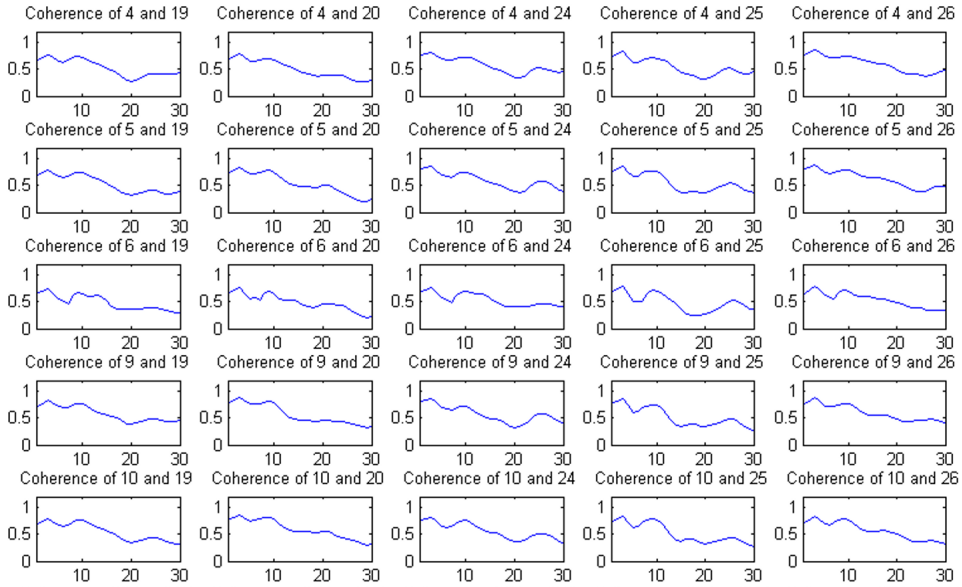
As connectivity measures can be used to extract information about the interaction of two signals, it was analyzed if connectivity measurements are more robust indicators for various cognitive workload states compared to normal power spectrum information. In this thesis two methods of connectivity measurements were comparatively accomplished.

Coherence

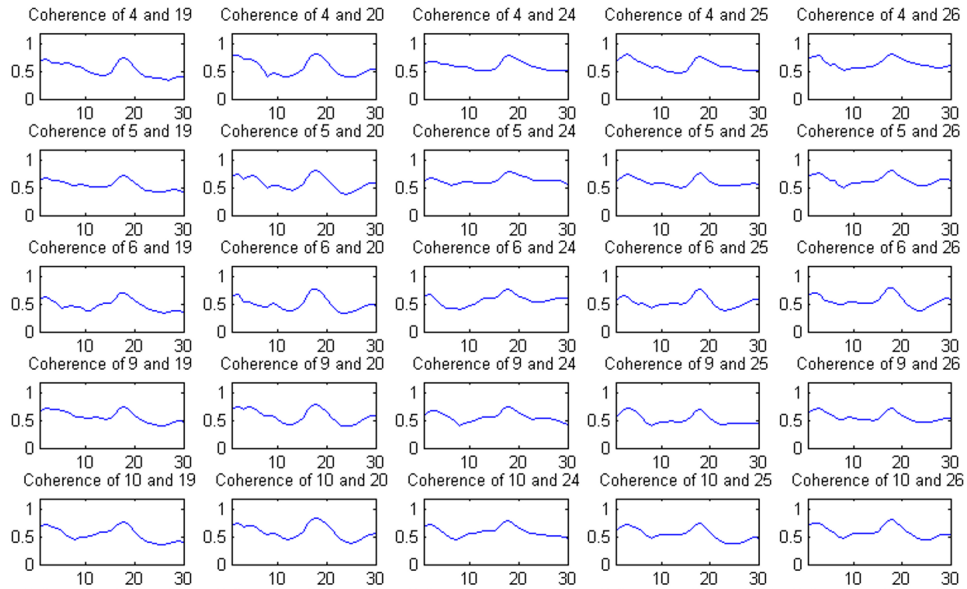
To measure the functional connectivity, coherence was evaluated for each combination of the frontal electrodes (F3, Fz, F4, FC1, FC2) and the parietal electrodes (CP1, CP2, P3, Pz, P4). High coherence between two EEG signals assumes high synchronization between underlying brain regions within a certain frequency band. As can be seen in Figure 6.4 (a) a high coherence between signals from frontal and parietal brain regions exists within the theta-, as well as alpha-frequency bands for subject 4 during calculating angle theorems. Compared to these results, a high coherence within the theta- and beta-frequencies can be measured while subject 6 is solving diagram tasks (Figure 6.4 (b)). The connectivity patterns for both tasks correspond to regions in which workload specific activation is expected, but merely the frequency range measured during angle theorem tasks equates to known literature. The beta-frequency band is most prominent during solving diagram tasks. Although the beta-frequency band is not commonly used for workload detection, some researchers postulated a correlation between the beta-band activity and the cognitive workload states [111, 51]. As the frequency ranges are not consistent over these two diverse task types and subjects, coherence will not be used as a feature in further studies.

Granger causality

Granger causality was applied as an additional measurement for feature extraction in EEG data. The effective connectivity was calculated by using the workload specific electrodes F3, Fz, F4, FC1, FC2, C3, Cz, C4, CP1, Cp2, P3, Pz, P4. The Granger causalities for subject 4 and 6 are plotted in Figure 6.5 and Figure 6.6. The red lines represent the effective connectivity for the difficult tasks, whereas the blue lines indicate the Granger causality for the easy tasks. While solving difficult angle theorem tasks, the theta-frequency power in the occipital brain area causes theta-band activity in the frontal parts of the brain (see Figure 6.5 (b)). For the easier angle theorems, the alpha-frequencies have their origin in



(a) Subject 4 - angle theorems



(b) Subject 6 - diagram tasks

Figure 6.4: Coherence for selected channels (4:F3, 5:Fz, 6:F4, 9:FC1, 10:FC2, 19:CP1, 20:CP2, 24:P3, 25:Pz, 26:P4), while solving angle theorems or diagram tasks for the frequency range of 0 Hz – 30 Hz.

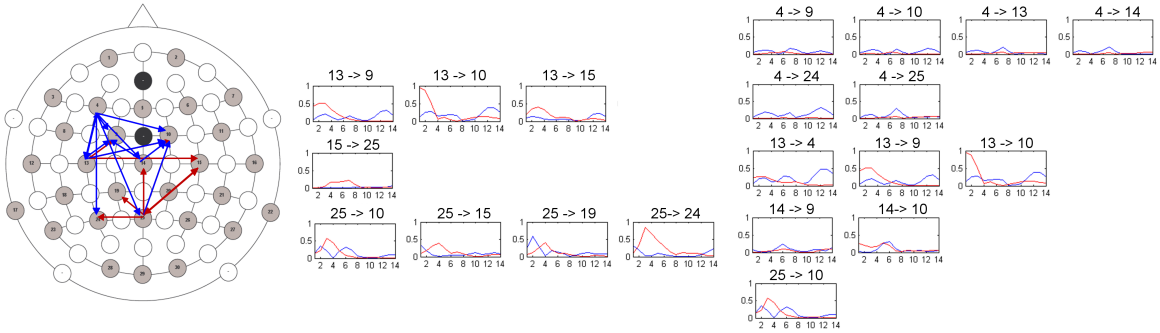


Figure 6.5: Granger causality for subject 4 while calculating angle theorems for selected channels (4:F3, 9:FC1, 10:FC2, 13:C3 14:Cz 15:C4 19:CP1, 24:P3, 25:Pz). Red lines indicate the connectivity between two electrodes in a specific frequency range while solving difficult tasks, whereas blue lines indicate the Granger causality during conducting easy tasks. **Left:** Overall map of Granger causality. **Center:** Granger causality in theta-frequency. **Right:** Granger causality in alpha-frequency.

the frontal brain areas and spread to the occipital brain areas over time (see Figure 6.5 (c)). These patterns, can likewise be seen in Figure 6.5 (a), the overall map of the effective connectivity. Compared to these results, the theta-frequencies for subject 6 during solving difficult diagram tasks have their origin in the frontal brain regions and spread to the occipital brain area over time. In contrast, alpha-band activity in the occipital brain area causes alpha-frequency power in the frontal brain regions, mainly while solving the easier assignments (see Figure 6.6). As the frequency gradients were not consistent over these two diverse task types and subjects, Granger causality is as feature not as robust as the normal power spectrum and thus will not be used in further studies.

6.3.4 Identification of workload based on neural source level

As shown in section 6.3.1 it is possible to find features at the surface channel level like an increasing theta-band activity and decreasing alpha-frequency power with rising workload. Since EEG suffers from the inverse problem and thus the localization of neuronal sources is weak, a source reconstruction method was applied to the data. Using ICA, the activity of selected independent components (IC) can be used for classification which allows a smaller dimensional feature space, as using features based on surface channels. After applying ICA, the activity of different sources (muscular artifacts, eye blinks, neural activity) is separated into independent signals which dramatically improve the signal to noise ratio (SNR). For reasons of better SNR and to avoid the curse of dimensionality, it was hypothesized, that it is favorable to use a few ICs for classification instead of using features based on surface channels. To prove this hypothesis, the ICA was calculated by using the logistic infomax-algorithm [140]. Subsequently, the component weights were mapped back to the original but filtered and re-referenced dataset, to retain as much information as possible. Furthermore, each component was additionally localized by utilizing dipole fit-

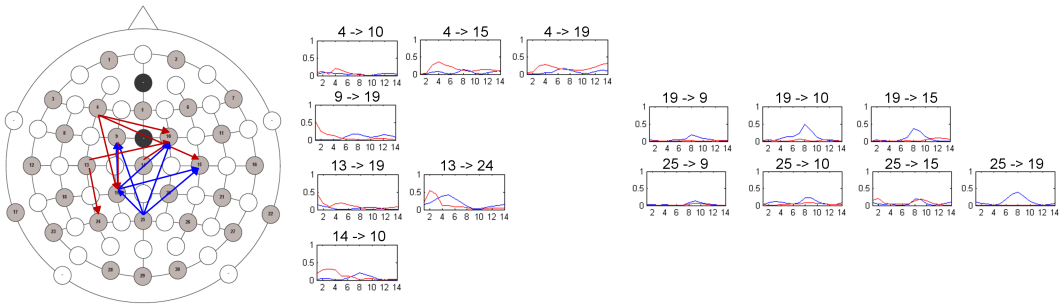
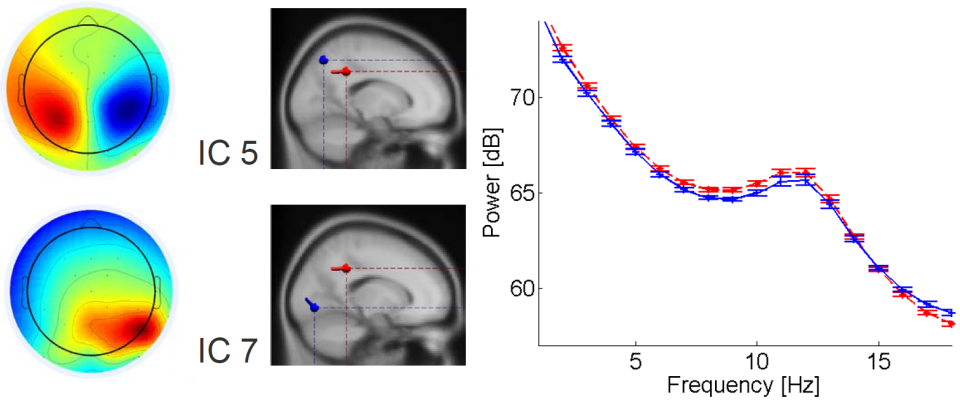
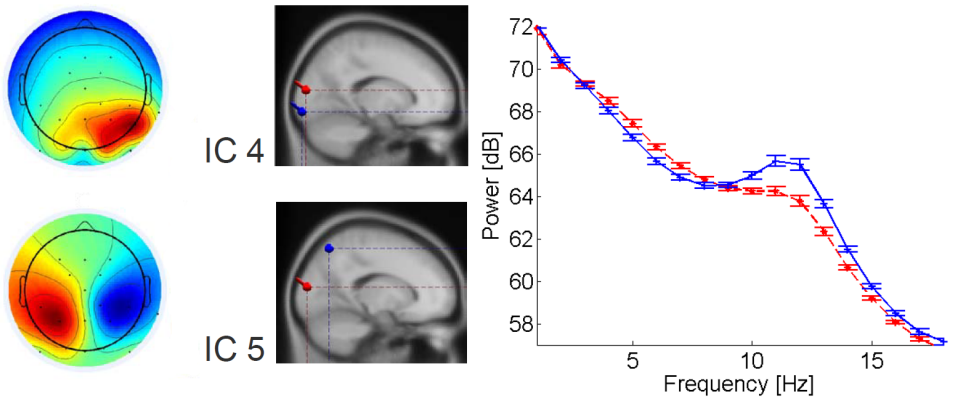


Figure 6.6: Granger causality for subject 6 while solving diagram tasks for selected channels (4:F3, 9:FC1, 10:FC2, 13:C3 14:Cz 15:C4 19:CP1, 24:P3, 25:Pz). Red lines indicate the connectivity between two electrodes in a specific frequency range while solving difficult tasks, whereas blue lines indicate the Granger causality during conducting easy tasks. **Left:** Overall map of Granger causality. **Center:** Granger causality in theta-frequency. **Right:** Granger causality in alpha-frequency.

ting. All locations of ICs were fitted using the "three shell boundary element head model" [141, 142]. This step minimizes the remaining residual variance in the EEG, which cannot be explained by the dipole's location. Afterwards, ICs with less than 15 % residual variance were selected for each participant. Dipoles outside the brain are automatically excluded. This allows the mapping of the brain signal to individual cortical areas and thereby a functional interpretation of each dipole. Due to the functional interpretation, neural components of interest can be selected. Furthermore, features which are distinctive between high and low workload states can be extracted. The selection of individual relevant ICs, as well as the improved SNR leads to a reduced dimensionality of the feature space.
The power changes were compared for each IC across the two difficulty conditions within each task. Figure 6.7 shows the scalp maps from ICs with significant differences in power, measured by r^2-values. The locations of dipoles after fitting, as well as the spectral power for the selected ICs are presented in Figure 6.7, as well. As the stated ICs have dipole-like scalp maps, as well as spectral peaks in the alpha-frequency band for both task conditions, these components can be assumed to be brain related. For both tasks, IC 5 is located at the right intra-parietal sulcus (IPS), IC 7 in the angle theorem task is fitted between the temporal and occipital right brain areas, whereas IC 4 for the diagram task is positioned at the right occipital brain areas. Especially IC 5 is neurophysiolgically meaningful, as the IPS is thought to be part of the fronto-parietal working-memory network. Further, an increased theta-frequency power in theorem and diagram tasks can be observed. In the diagram tasks, a reduced alpha-band activity was recognized, while the level of difficulty increased. However, in the theorem tasks, an increased alpha-band activity could be measured during rising level of task difficulty, which is not consistent with common literature. As the SNR is improved, it was argued, using ICs for classification leads to better classification results than using features based on surface channels. In the next section, this hypothesis will be analyzed. Hence, it will be examined, if ICs are the more robust features for workload detection and classification, compared to features based on channel level.



(a) Source level analysis for subject 4 (IC 5 and IC 7) while calculating angle theorems.



(b) Source level analysis for subject 6 (IC 4 and IC 5) during solving diagram tasks.

Figure 6.7: Left: Scalp maps from the ICA inverse weight matrix for selected independent components (IC). **Center:** Dipole fitting based on chosen ICs. **Right:** Spectral power plot. The red line indicate the power while solving difficult tasks, whereas the blue line indicate the power during conducting easy tasks.

6.4 Classification results

In this section, classification accuracies will be reported as quality criterion for differentiating workload states based on channel level, as well as on source level. The intention of using source level features is the improved SNR and thus the improved feature space, which should achieve better classification results. As reported in section 6.2, the EEG data was segmented in 3 sec time windows. Each time window was classified by using a SVM with RBF-kernel. The predicted class label $y_i = \{-1, 1\}$ of each segment i was saved in an array. Furthermore, classification probabilities [88, 143] of each 3 sec segments were estimated. After each classification step, the labels were weighted by the probabilities. Afterwards, the weighted label array was summed over all 3 sec time windows composing the 45 sec trial. After classifying all 15 of the 3 sec time windows of one 45 sec trial, the mean of the label array was calculated. If a low probability was estimated, the classification is unreliable and the value is weighted less. Whereas high probability indicates a reliable classification of the 3 sec segment, which is why the segment is heavily weighted. If the result is negative, the complete trial will be labeled as an easy stimulus. When the result is positive or zero, the trial will be labeled as difficult stimulus. For calculating the modified overall accuracy, the following formula was used:

$$modAcc = \frac{\sum_{n=1}^{N} l_d \cdot p_d + \sum_{m=1}^{M} l_e \cdot p_e}{x} = \frac{x_{wd} + x_{we}}{x} \tag{6.1}$$

where, l_d and l_e represent the correctly predicted label of each 3 sec window being a difficult or an easy stimulus, whereas p_d and p_e describe the probability values of each correct classified segment being difficult or easy. N and M are the total number of correctly classified trials, as easy or difficult. x_{wd} and x_{we} are the number of correctly labeled weighted trials, for a difficult or an easy task. x is the total number of segments used for classification. The ultimate goal of this method is to increase the accuracy in order to reduce false rates.

6.4.1 Classification on channel level

In previous analyses, most differences between two workload states during easy and difficult tasks appear in the theta-frequency band in the central frontal brain area [50], as well as in the alpha-frequency band in the right parietal, occipital brain area [17, 129]. The boundaries of frequency bands are not exactly definable between subjects, thus the frequency interval was individually chosen for each subject. Each frequency range with significant differentiations between the two conditions of difficulty were used for classification (see Figure 6.2). Furthermore, solely electrodes with high r^2-values, consistent with results from Klein et al. [18] were used for classification to enhance computing time and to improve the SNR. The modified classification accuracies based on channel level are reported for each subject in Table 6.1.

Table 6.1: Modified classification accuracies of channel and source level features. The classifier was applied to data recorded while calculating angle theorems (left) or solving diagram tasks (right).

	Angle Theorem		Diagram	
Level	Channel (EEG)	Source (IC)	Channel (EEG)	Source (IC)
Sub 1	81.25 %	81.25 %	56.25 %	68.75 %
Sub 2	81.25 %	87.50 %	87.50 %	81.25 %
Sub 3	87.50 %	93.75 %	87.50 %	75.00 %
Sub 4	93.75 %	93.75 %	56.25 %	43.75 %
Sub 5	87.50 %	75.00 %	75.00 %	75.00 %
Sub 6	68.75 %	62.50 %	100.00 %	81.25 %
Sub 7	68.75 %	75.00 %	75.00 %	68.75 %
Mean	81.25 %	81.25 %	76.79 %	70.54 %

The weighted classification of two varying workload states, while calculating angle theorem tasks, lead to high classification accuracy ($> 81\,\%$) in 5 out of 7 subjects. Subject 4 reached highest accuracy ($modAcc = 93.75\,\%$) during calculating angle theorem tasks. The individual features used for classification were electrodes over central and parietal brain areas (FC9, FC10, CP1, CP2, P3, Pz, P4) in the frequency range of $5.5\,\text{Hz} - 9.5\,\text{Hz}$. The modified classification of two workload states while solving diagram tasks led to good classification accuracy ($\geq 75\,\%$) in 5 out of 7 subjects (see Table 6.1). The highest accuracy rate with $modAcc = 100\,\%$ was reached for subject 6 while solving diagram tasks, by using the electrodes over frontal, central and parietal brain areas (F3, Fz, F4, FC9, FC10, CP1, CP2, P3, Pz, P4) in the frequency range of $3.5\,\text{Hz} - 7.5\,\text{Hz}$ and $9.5\,\text{Hz} - 13.5\,\text{Hz}$.

6.4.2 Classification on source level

As stated in the previous section 6.3.4, it was possible to detect independent sources (i.e., ICs) which could be clustered according to different types of underlying workload. By using features on source level, the SNR should be improved, to achieve better classification accuracies. High classification results were reached for the classification of two different workload states based on ICs, while calculating angle theorem tasks ($\geq 75\,\%$) in 6 out of 7 subjects (see Table 6.1). For the angle theorem tasks, subject 3 and 4 achieved the maximal classification accuracy, with $modAcc = 93.75\,\%$. The classification based on source level features (i.e., ICs) for diagram tasks lead to good classification accuracies ($\geq 75\,\%$) in 5 out of 7 subjects (Table 6.1). Subject 2 and 6 reached the highest accuracy rate with $modAcc = 81.25\,\%$ for the diagram task.

6.5 Discussion

Using a within subject study design, it was possible to classify two types of workload induced by realistic learning materials presented in an easy and difficult way. A variety of feature extraction methods were comparatively applied. Furthermore, the classifications of channel as well as source level features were successfully conducted.

6.5.1 Various neurophysiological features

In this pilot study, three different neurophysiological features were compared, to get the optimal preferences for further analysis. First, a standardized power spectrum was calculated. The type of diagram is the same in both representations, but the questions to be answered differed in complexity, which induce different workload states. Therefore, it was possible to detect the hypothesized characteristics in the power values in 5 out of 7 subjects. In comparison, the difference levels of angle theorem tasks are not that clearly defined. The presentation of the learning material differs in description complexity and color coding, though the assignment which had to be solved were the same for each theorem, independent of their complexity. Depending on prior knowledge of each user and the way of presenting the learning material, it is possible that additional cognitive processes were activated. Although, the power spectrum trend was more widespread during solving angle theorem tasks, for the majority of all subjects an increased magnitude of an alpha-desynchronization was detected. This is a result of increasing task complexity or attention, because the alpha-frequency is known as a general attention index [144]. The theta-frequency band desynchronization for more complex tasks could be induced by retrieval processes [145]. As postulated in previous studies, these EEG patterns concerning electrode location and frequency bands are indeed advisable as features for further classification.
Second, it was analyzed, if connectivity measurements are more robust features compared to the power spectrum information. The connectivity patterns for both tasks correspond to regions in which workload specific activation is expected, but merely the frequency range measured during angle theorem tasks equates to known literature. As the frequency ranges are not consistent over the two diverse task types, coherence will not be used as feature in further studies. A successful connectivity measurement was Granger causality, which showed opposing frequency gradients. Granger causality is as feature not as robust as the standardized power spectrum, since the frequency gradients are not consistent over diverse task types or subjects and thus will not be used in further studies.
To summarize, characteristics in the EEG signals from frontal, parietal and occipital positioned electrodes, as well as frequency bands in alpha- and theta-frequency ranges come along with findings in the literature [17, 50, 129]. These findings are robust over diverse task types and can be calculated in real time, so that they can be used as features for the upcoming studies. Furthermore, the localizations are consistent with results of the fMRI study postulated in [18]. Klein et al. [18] reported activations in the left IPS while dealing with easy material, whereas treating more complex material causes a bilateral activation of the IPS.

6.5.2 Channel versus source level features

A successful classification of different levels of workload was possible by using channel, as well as source level features. In some cases the classification accuracy did not exceed the 70 %. No suitable features, resulting from a low number of significant differentiations in the power spectrum, can be the reason for these results. The source based features contain precise information about the brain areas responsible for the activations. Furthermore, using ICs as features should improve the feature space and thus improve the classification accuracies. Contrary as hypothesized, the classification accuracies do not differ significantly, or the classification results based on ICs are worthless. The calculation of ICs is time consuming and needs a high number of electrodes to be reliable. This is not practical in an online learning setting. The cost-benefit factor of using ICs as features for classification is low. Therefore, the source level classification will not be used in the following studies, but the standardized power spectrum as classification feature seems to be promising.

6.5.3 Issues using complex learning material

The various levels of difficulty for angle theorems and diagram tasks were perceptually not identical, potentially leading to differences in processing. This can cause different levels of workload onto working-memory. Even if there are no obvious motor confounds between task difficulty in this study, there might be nevertheless perceptual confounds. More difficult tasks might have more complex visual displays and might need more text information. More complex displays might result in different eye-movements that can be picked up by a classifier. Even if classification is possible for realistic tasks imposing different levels of workload, it might still not be very helpful theoretically. As in chapter 5, it cannot be assumed, whether workload classifiers trained on realistic tasks really represent a measure of workload.

6.6 Conclusion

The results show that it is possible to classify based on single-subject single-trial EEG data whether learners study realistic instructional materials of low or high workload levels. Although the results are practically and methodologically interesting, the same conceptual critique with regard to perceptual-motor confounds as in the previous study 5 applies to this study.
For further studies, the standard features on frequency level are most promising. Compared to ICA, the standard feature selection methods can be applied in real time, so that they can be used in online settings. Furthermore, the presented stimuli used for classifier training should be not as complex and well controlled. Moreover, for a learning environment in a school context, a generalized classifier is preferred. The calibration of the classifier should be performed with a minimal effort of the user. Additionally, the classifier should be able to classify various workload states, regardless of the task.

To solve these challenges, the method of cross-task classification will be introduced in the following chapter. By using cross-task classification, it should be possible to identify neural signatures of workload in complex working-memory tasks, as well as in simple learning tasks.

7 Cross-task workload prediction

In the previous chapter 6, a classification of EEG characteristics induced by various cognitive workload states was possible. Although these are important findings, this approach cannot be used in an online adaptive learning environment. The calibrated classifiers are only trained on task specific features and the calibration phase is too time consuming for complex learning material, because a high number of training trials is needed. Therefore, this chapter will focus on generalizing classification methods. It will be explored, if a generalized classifier trained on simple working-memory tasks is able to predict various cognitive workload states in complex learning tasks. The results reported in the following sections were partially presented or published in [8, 99, 146].

7.1 Study design

In the following, the participants data will be introduced and the methods for EEG recording, the task design, as well as the procedure will be described.

7.1.1 Participants and EEG recordings

A total of 21 subjects, 10 men and 11 women aged between 20 and 68 (median age 25), voluntarily participated in the EEG experiment. All of them were native or fluent in German. All participants signed a written informed consent. A set of 30 active electrodes (actiCap, BrainProducts GmbH), attached to the scalp and placed according to the extended International Electrode 10 - 20 Placement System [36], were used to record EEG signals. Two additional electrodes record an EOG; one placed horizontally at the outer canthus of the right eye to measure horizontal eye movements and one placed in the middle of the forehead, between the eyes, to measure vertical eye movements. The reference electrode was placed on the left mastoid, the ground electrode at AFz. EOG and EEG signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp, g.tec). The sampling rate was 512 Hz and the impedance of each electrode was less than 5 kΩ. EEG data was high-pass filtered at 0.1 Hz and low-pass filtered at 100 Hz during the recording. Furthermore, a notch-filter was applied between $48\,\text{Hz} - 52\,\text{Hz}$ to filter power line noise.

7.1.2 Paradigm

The experiment had a within-subject design and comprised three working-memory tasks as well as arithmetic and algebra problems, all presented in three levels of difficulty. It was ensured that subjects received identical visual input in all levels of task difficulty to avoid perceptual confounds. A pre-test at the beginning and a post-test at the end of the

experiment were accomplished to achieve a direct comparison of a subjects' knowledge before and after solving learning assignments. At the beginning and the end of each task, a fixation cross appeared for 5 sec. Furthermore, after each level of difficulty, subjects were asked for a subjective workload rating. The cognitive workload scale ranged from "1" being too easy, up to "7" being too difficult. The whole experiment took approximately three hours for each subject.
For modeling a realistic learning scenario, the mathematical word problems were presented in a fixed order, at increasing levels of difficulty, for each learner. For reasons of comparison, the same increasing order of difficulty was employed for each working-memory task, to generate the data used for classifier training. To rule out potential negative effects of such a fixed task order (e.g., slow drifts in the EEG signal that can influence the classifier) the $\%ERS/ERD$ ratios were calculated, using a baseline for comparison immediately following or preceding the window of analysis. The $\%ERS/ERD$ ratios are based on power values using two time intervals for each task. One interval strongly imposes the task-specific workload (activation interval, I_A) and one interval imposes no or only a low level of workload (resting interval, I_R). It was ensured that each trial of a task included both intervals to avoid systematic effects of drifts in the EEG signal (see Figure 7.1).

7.1.2.1 Working-memory tasks - Training tasks

Three working-memory tasks (go/no-go, n-back and reading span) were chosen to measure different workload levels, which is necessary for successful performance and prediction.

Go/no-go task
The go/no-go task [147] was used to measure inhibition control with respect to workload. To modify workload, the task was presented in three difficulty levels. The easiest condition presented two letters, N as Go and X as No-Go stimuli. The medium difficulty had 15 consonants as Go (B, C, D, F, G, H, J, K, L, M, N, P, R, S, T) and X as No-Go stimuli. The most difficult condition had the same 15 Go stimuli, but X and Y as No-Go stimuli. Each letter was presented for 300 ms followed by a black screen of 700 ms (see Figure 7.1(a)). The subjects had to solve 40 trials per condition.
In the go/no-go task, every 1000 ms a new letter appeared at the screen. Subjects reacted by pressing a button or inhibit the reaction, depending on the presented letter. The data from stimulus onset until 125 ms before keypress served as action interval I_A. Inhibition was required during that time interval. The time period from 125 ms after keypress until the next stimulus was shown, was used for the resting period I_R. In this interval no cognitive processing was required.

N-back task
The n-back task, requires an updating of items together with a replacement (inhibition) of previous set members interleaved with an identity-matching task. This task is commonly used to manipulate cognitive workload. In this task, single-digit numbers had to be memorized (except the number seven, the only two-syllable number). Subjects had to decide upon number presentation, whether or not it was the same as the n-th number before. The

easiest condition was a 1-back, the medium condition was a 2-back and the most difficult condition was a 3-back task. All numerals were displayed for 500 ms each, followed by a fixation cross for 1500 ms (see Figure 7.1 (b)). For each level of difficulty, 36 trials were conducted within a block design.
In the n-back task, every 2000 ms a new digit was presented for identity matching. Subjects could react by pressing "yes" or "no" anytime. The time interval from stimulus onset until 125 ms before keypress was used to define I_A. In this interval, identity matching and updating was required. The time period from 125 ms after keypress until the next stimulus appeared was used to define I_R. In this interval mainly storage was required.

Reading span task

A reading span task requires shifting between semantic processing and simple additive updating of items. This task is based on memorizing letters (B, F, H, J, L, M, Q, R, X) and on verifying 60 short German sentences. In the easiest level, subjects had to type in two remembered consonants after every second sentence, in the second level of difficulty four letters after every fourth sentence and in the most difficult condition six letters after every sixth sentence. The subjects had a maximum of 10 sec to type in the remembered consonants. Each sentence was presented for a maximum of 5 sec, followed by a letter displayed for 1 sec (see Figure 7.1(c)). Because of time constrains, subjects had to solve only 20 trials per condition.
In the reading-span task, a list of sentences was presented for verification (e.g., oranges are blue). After each sentence, subjects could react by pressing "yes" or "no" anytime. After keypress, a fixation cross was presented for 500 ms before a letter (that had to be remembered) appeared for 1000 ms. The time interval in which the letter appeared at the screen, was used to define I_A. In this interval a shifting between the semantic processing task and an updating of the set of letters to be remembered was required. The period from 125 ms after keypress (sentence verification) until the next letter appears was used as I_R. In this interval mainly storage was required.

Combining these three working-memory tasks differing with regard to specific executive functions should prevent the classifier from picking up task-specific features.

7.1.2.2 Realistic learning tasks - Testing tasks

Subsequent to the working-memory tasks, which serve as training sets for the classifier, participants worked on two types of mathematical word problems (arithmetic and algebra). For both types of tasks, again three levels of task difficulty were implemented that strongly differed with regard to the level of workload they induced. To avoid perceptual confounds of task difficulty, the word problems were matched for the amount of words. Furthermore, exactly four numbers per problem and task difficulty occurred, either numbers or fractions. These conditions were made to ensure, more difficult word problems do not result in more text information, leading to differences in processing, which may show up in the EEG.

First, subjects read a page with facts as long as they wanted. After keypress, a problem statement appeared that had to be solved. The participants could react by pressing a next-button anytime. Subsequently the calculation phase appeared, during which subjects could use the calculator for max. 10 sec and then select one out of four multiple-choice options in max. 5 sec to provide their solution. Subsequently, a fixation cross was presented for 500 ms. Since solving complex learning tasks is very time consuming and induces learning effects during the performance of the task, only 10 trials per condition were conducted for the arithmetic, as well as for the algebra task.

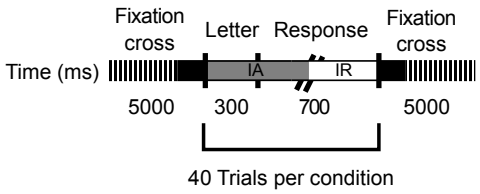
Arithmetic tasks

In order to solve the arithmetic tasks, a variable x had to be calculated by selecting and integrating appropriate numbers. For the first level of task difficulty, the subjects merely had to select one out of four numbers ($x = a$). The second level required them to subtract two relevant numbers ($x = a - b$). The third level, asked for the subtraction of two differences, thus involving all four given numbers ($x = (a - b) - (c - d)$). This manipulation of task difficulty was based on a taxonomy by Schnotz et al. [148] and can be assumed to induce strong differences in workload. For solving the assignment, subjects had as much time as they desired.

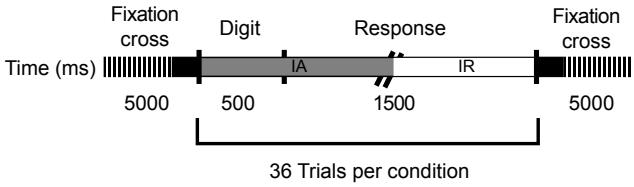
Algebra tasks

The algebra word problems required subjects to select and instantiate algebraic expressions containing fractions and multiplications in order to determine the value of a variable x. For the first level of task difficulty, the appropriate expression contained one fraction which had to be multiplied with a number ($x = c \cdot \frac{a}{b}$). In the second difficulty level the expression contained the same fraction two times, which had to be multiplied by two numbers and the results had to be summed up ($x = c \cdot \frac{a}{b} + d \cdot \frac{a}{b}$). The third level contained two different fractions, which had to be multiplied with two numbers ($x = c \cdot \frac{a}{b} + d \cdot \frac{e}{f}$). This manipulation of task difficult due to increased workload was based on a taxonomy by Scheiter et al. [149]. For solving the task, subjects had as much time as they needed.

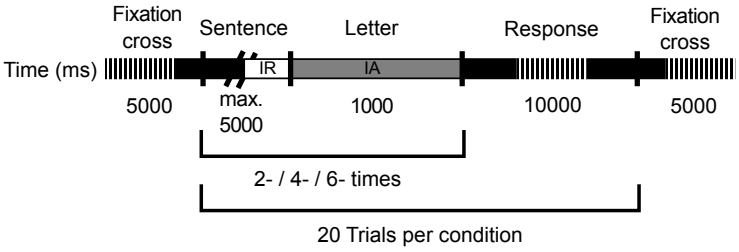
In both realistic learning tasks (i.e., arithmetic and algebra), a series of word problems was presented to the subjects. The time interval from the onset of the problem statement until 125 ms before keypress was used as I_A. In this interval the necessary facts had to be remembered and inferences had to be drawn for problem solution. For I_R, the time period from 125 ms after keypress (providing problem solution) until the next page of facts appeared, was used. During this interval, no cognitive processing was required.
In all tasks, subjects had to react with key presses to provide their answers (either with "yes" or "no" in the working-memory tasks or with one out of four multiple choice options for the learning tasks used for classifier testing). The motor reaction was exactly the same for both working-memory tasks and both math word problems. Furthermore, there were no differences in the motor reaction between different levels of task difficulty. No feedback was given to subjects in order to avoid confounding task difficulty and ratio of negative feedback.



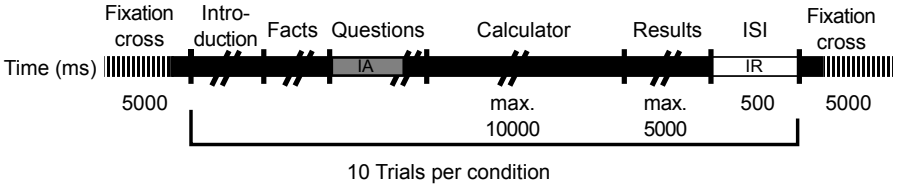
(a) Go/no-go Task



(b) N-Back Task



(c) Reading Span Task



(d) Math Word Problems

Figure 7.1: Schematic flow of each task type. The grey line indicates the activation interval (I_A), whereas the white line represents the reference interval (I_R) used for the $\%ERS/ERD$ calculation. The two slashes represent the keypress and the dashed line stands for a shortened representation of the visualized time range.

7.2 Data pre-processing

With regard to the windows used for analyzing EEG data, it was ensured, they did not contain any motor artifacts or perceptual confounds that could be picked up by a classifier. For each task too noisy EEG data was removed for further analysis after visual inspection. Furthermore, the windows of EEG analysis always ended at least 125 ms before any keypress, to exclude EEG signals based on motor planning [145]. Additionally, a time interval of 125 ms after any keypress was excluded from the analysis due to potential motor artifacts. Compared to the studies reported in chapter 5 and 6, an additional data pre-processing step was conducted. An artifact reduction method described by Schlögl et al. [150] was used, to reduce EOG artifacts. It is important to note that the predefined length of each trial, both in I_A as well as in I_R, varied between individuals and across trials, since the intervals were dependent on subjects' response time. This analysis procedure has been frequently used in preliminary EEG studies analyzing cognitive processes [145, 151]. The defined trial interval covers the entire time period of problem solving independent of differences between individuals or task conditions. Therefore, each predefined variable trial length was individually adjusted for each task and user to the shortest occurring trial length by using shifting windows with a minimum length of 65 samples (127 ms) up to a maximum length of 171 samples (345 ms) and 90 % overlap. The power spectrum was calculated for each overlapping window by using ARM based on Burg's maximum entropy method [139], using a model order of 64, which was determined by means of the MATLAB toolbox ARMASA [152]. Subsequently all sub-power spectra per trial were averaged, resulting in one power spectrum for each trial, with a frequency resolution of 1 Hz bins. Using fixed trial intervals for analyzing the data would not capture the whole part of the cognitive processes related to solving the given problem (interval length < response latency) or additionally include task-unrelated cognitive processes (interval length > response latency) [145].

7.3 Methods for classification

In the present study, the easiest and most difficult condition of the working-memory tasks, as well as of the complex learning tasks, were used for classification. LIBSVM [89] was employed for classifying EEG data. In both classification methods, SVMs with a RBF kernel were utilized to classify workload states in EEG data during easy and difficult tasks. As postulated in Lotte et al. [87], this is a promising kernel for analyzing EEG data. The RBF kernel parameters $C = 1$ and $\gamma = (\frac{1}{\#features}) = 0.0025$, were used for each subject to maintain comparability over subjects.
As features $\%ERS/ERD$ values of the channels F3, Fz, F4, FC1, FC2, CP1, CP2, P3, Pz, P4 [153] for the frequency range $1\,\text{Hz} - 40\,\text{Hz}$ [51] were used. By calculating r^2-values for all frequencies and each subject, no clear frequency band, e.g., $3\,\text{Hz} - 6\,\text{Hz}$ or $8\,\text{Hz} - 13\,\text{Hz}$ consistent over all subjects, could be identified. Therefore, the wide frequency range of $1\,\text{Hz} - 40\,\text{Hz}$ was used, resulting in 40 features for each of the 10 electrodes. These strict features were used for the sake of consistency across subjects and tasks.

For the remaining differences in the difficulty levels across the tasks, the scaling method of the EEG power values was improved by scaling the data over trials and by adjusting the range of the diverse datasets. First, the training data were *z-score* normalized resulting in a centered, scaled version of the input data. Subsequently, the means and standard deviations of the power values of the training set were calculated. Finally, the testing data was normalized with regard to these means and standard deviations calculated from the training data. To determine the significance of classification performance, bootstrapping was utilized for each train-test combination with a significance level of 0.05.

7.4 Generalizable classification methods

The goal of the present study was to develop generalizable classification methods which are able to differentiate levels of workload across different tasks. Therefore, a within-task classification as well as a cross-task classification were comparatively applied for each subject individually. For each participant the same number of datasets used for classification were created: go/no-go (“A”), n-back (“B”), reading span (“C”), arithmetic (“D”) and algebra (“E”).

7.4.1 Within-task classification

In order to show that it is possible to indicate differences of workload in the EEG data a within-task classification was accomplished. Therefore, training and testing of a classifier were performed on EEG data recorded during the same task. This method was applied for the datasets A, B, C, D and E, to be sure each task was classifiable in general. Good differentiability of workload levels within each task was required to apply cross-task classification. For within-task classification a 10-fold cross validation was accomplished to verify the separability of the independent dataset, to verify the separability.

7.4.2 Cross-task classification

To investigate how well workload levels can be predicted across tasks, cross-task classification was accomplished. Cross-task classification uses EEG data recorded by solving well-defined and short working-memory tasks (go/no-go, n-back or reading span) for SVM training, whereas the calibrated SVM is tested on EEG data recorded while processing a different complex learning task (arithmetic or algebra). For cross-task classification all trials of the training data were used to calibrate the classifier. Afterwards, each trial of the test data served as new independent input data, to verify the generalizability of the independent test data according to the pre-trained classifier model.
Good generalizability of the classifier as well as working-memory tasks inducing nearly the same workload states and types of neural processing as in the later used learning tasks are required to achieve high classification performances.

7.5 Linear mixed models for non-linear data

In this study, subjects had to rate tasks with predefined difficulty levels on a cognitive workload rating scale. As subjects have diverse workload capacities, as well as different prior knowledge, the experience for difficulty might be different between subjects. Hence, the multiple ratings from the same subject cannot be regarded as independent from each other. Therefore, a different "baseline" difficulty value for each subject was assumed, to resolve the non-independence. To analyze if these dependent individual differences influence the workload prediction, linear mixed models (LMM) were generated.
The LMMs were calculated for each frequency band, averaged over all subjects, by computing the model with R as follows:

```
power ~ workload rating; random =~ (workload rating | subject)
```

meaning, the cognitive workload ratings are assumed to be the fixed effect, as each subject has the same numerical range ("1 - 7"), to rate each trial in its difficulty level. Furthermore, the random effect is `(cognitive workload rating | subject)`, assuming an interception for the cognitive workload rating that is different for each subject. There are multiple responses per subject and these responses depend on each subjects baseline for difficulty experiences.

7.6 Behavioral results

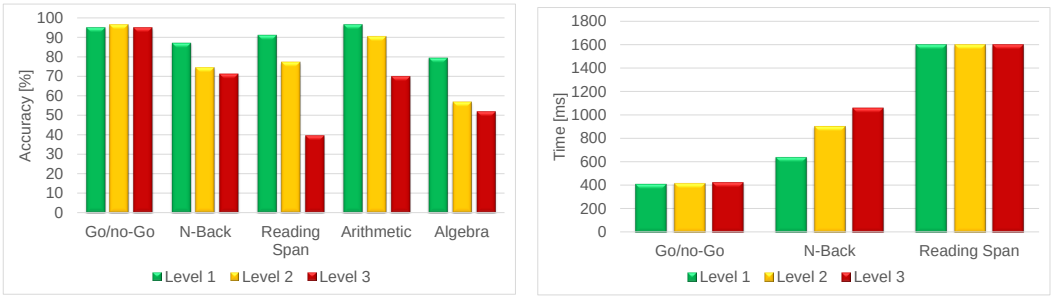
In this section, the behavioral data for each task will be reported. The performance results can be seen in Figure 7.2.

Go/no-go task

In the go/no-go task, the mean reaction time was 414 ms for the easy, 420 ms for the medium and 428 ms for the difficult condition. Participants correctly answered 95.11 % of trials in level 1, 96.54 % in level 2 and 94.88 % in level 3. Response accuracies of level 1 and level 3 differed significantly ($p = 0.02$ paired, two-sided Wilcoxon test). There was no significant difference in reaction times and response accuracies between level 1 and level 2 as well as between level 2 and level 3 ($p > 0.05$, paired, two-sided Wilcoxon test).

N-back task

During the n-back task, the mean reaction time was 641 ms for the 1-back, 905 ms for the 2-back and 1061 ms for the 3-back task. The response times for 1-back vs. 2-back, as well as 1-back vs. 3-back differed significantly (1-back vs. 2-back: $p = 0.0033$ paired, two-sided Wilcoxon test; 1-back vs. 3-back: $p = 0.0001$ paired, two-sided Wilcoxon test). There was no significant difference in reaction times for 2-back vs. 3-back ($p > 0.05$, paired, two-sided Wilcoxon test). Participants answered on average 85.12 % in 1-back, 74.61 % in 2-back and 71.19 % in 3-back task correctly. For 1-back vs. 3-back and 2-back vs. 3-back a significant difference in the response accuracy could be noticed (1-back vs. 3-back: $p = 0.0099$, paired, two-sided Wilcoxon test; 2-back vs. 3-back: $p = 0.01$, paired,



(a) Response Accucary

(b) Response Time

Figure 7.2: Behavioral results for each task and difficulty level, averaged over all subjects.

two-sided Wilcoxon test). The difficulty levels 1-back vs. 2-back showed no significant difference in the response accuracy ($p > 0.05$, paired, two-sided Wilcoxon test).

Reading span task

In the reading span task, the response time and response accuracy of typing in the remembered letters is most interesting for performance measurement. The mean response times (1.6 sec for all three conditions) did not differ significantly across the three difficulty levels, but the response accuracy (level 1 = 91.19 %, level 2 = 77.38 % and level 3 = 39.76 %) differed highly significant (for all three combinations $p < 0.0001$, paired, two-sided Wilcoxon test).

Arithmetic task

During arithmetic tasks participants answered 96.67 % in level 1, 90.47 % in level 2 and 69.05 % in level 3 correctly on average (level 1 vs. level 2: $p = 0.0181$, paired, two-sided Wilcoxon test; level 1 vs. level 3: $p < 0.0001$, paired, two-sided Wilcoxon test; level 2 vs. level 3: $p = 0.0023$, paired, two-sided Wilcoxon test). 20 out of 21 subjects solved the algebra task, because one subject terminated the experiment prematurely.

Algebra task

In the algebra task participants achieved 79.5 % correct responses in the easy, 57 % correct responses in the medium and 52 % correct responses in the difficult condition (level 1 vs. level 2: $p < 0.0001$, paired, two-sided Wilcoxon test; level 1 vs. level 3: $p < 0.001$, paired, two-sided Wilcoxon test; level 2 vs. level 3: $p > 0.05$, paired, two-sided Wilcoxon test). Reaction times were not measured in the learning tasks, because speed was not relevant for our experiment in these tasks.

7.7 Classification results

The within-task classification results are reported in section 7.7.1, Table 7.1 and the cross-task classification results are shown in section 7.7.2, Table7.2.

Table 7.1: Within-task classification accuracies in % for a two-class problem for all subjects. The bold values represent the significant accuracies ($p < 0.05$, permutation test). Row 2 exhibits the datasets used for classifier training. Row 3 represents the datasets used for classifier testing. A = go/no-go, B = n-back, C = reading span, D = arithmetic, E = algebra.

	Within-task Classification				
Training	A	B	C	D	E
Test	A	B	C	D	E
Sub 1	**96.67**	**88.89**	**95**	**75**	**40**
Sub 2	**98.48**	**95.16**	**93.75**	**75**	**80**
Sub 3	**98.48**	**95.45**	**95**	**80**	70
Sub 4	**95.31**	**98.57**	**95**	**90**	**85**
Sub 5	**98.48**	**85.71**	**95**	**90**	65
Sub 6	**98.33**	**97.14**	**96.25**	**70**	50
Sub 7	**100**	**95.71**	**95**	65	60
Sub 8	**95.45**	**94.28**	**92.85**	65	60
Sub 9	**84.37**	**95.58**	**95**	**90**	70
Sub 10	**100**	**93.75**	**100**	**80**	**85**
Sub 11	**100**	**100**	**95**	**75**	50
Sub 12	**98.43**	**100**	**95**	**90**	50
Sub 13	**97.05**	**95.83**	**95**	–	65
Sub 14	**98.43**	**100**	**96.25**	40	**75**
Sub 15	**96.87**	**95.31**	**95**	55	**75**
Sub 16	**98.48**	**96.87**	**95**	**80**	**80**
Sub 17	**95.45**	**83.33**	**95**	60	–
Sub 18	**100**	**100**	**95**	70	**75**
Sub 19	**98.48**	**97.05**	**95**	**75**	**80**
Sub 20	**95.31**	**100**	**95**	65	60
Sub 21	**100**	**92.85**	**95**	**75**	70
Mean	**97.35**	**95.31**	**95.19**	**73.25**	67.25

7.7.1 Within-task classification

For the within-task classification, the classifier was trained and tested using EEG data recorded during the same tasks (datasets A, B, C, D and E). Furthermore a 10-fold cross validation was used to estimate the accuracy of the trained SVM and to guarantee that training and test data do not overlap. The result for within-task classification of $\%ERS/ERD$ are shown in Table 7.1.

The go/no-go task had the highest accuracy over all subjects with 97.35 % prediction accuracy over all subjects. The n-back task performance over all subjects reached an accuracy of 95.31 %, whereas the reading span task reached a classifier performance over all subjects of 95.19 %. The complex learning tasks reached lower performance accuracies. The arithmetic tasks were classified with an average accuracy of 73.25 % and the algebra tasks with 67.25 %. In almost all cases within-task classification worked well and reached significant classification accuracies ($p < 0.05$, permutation test) on average, except the algebra task ($p > 0.05$, permutation test). Since EEG data of subject 13 was not recorded properly while solving the arithmetic task, no within-task classification results can be reported for this task. Subject 17 quit the experiment after solving the arithmetic task, therefore no within-task classification results can be reported for this subject during the algebra task, neither.

7.7.2 Cross-task classification

For cross-task classification, data from working-memory tasks were used to train a SVM. Additional working-memory tasks or complex learning tasks were used for classifier testing. Thus, the SVM was trained by using dataset A, B, C, ABC, or BC and then tested on EEG data from remaining workload tasks or from complex learning tasks (dataset D and E). The results for cross-task classification of the $\%ERS/ERD$ values are shown in Table 7.2. Predicting cognitive workload during complex learning tasks based on classifiers trained with working-memory tasks seems to be challenging.
Mean classification performances for all train-test combinations, were not significant ($p > 0.05$, permutation test). N-back reached best results for workload classification prediction in complex arithmetic tasks (B-D), with 53.3 % on average. For this cross-task classification 3 out of 20 subjects reached a significant classification performance ($p < 0.05$, permutation test) on individual basis. For n-back and algebra data as train-test combination (B-E) a mean accuracy of 52.4 % was achieved. Again 3 out of 20 subjects reached significant classification accuracies ($p < 0.05$, permutation test) on individual basis for this cross-task classification. The cross-task classification using go/no-go tasks for training and arithmetic respectively algebra tasks for testing (A-D, resp., A-E) resulted in 52.8 % and 51 % classification accuracies. For both classifications only 2 out of 20 subjects reached significant classification accuracies ($p < 0.05$, permutation test) on individual basis. Reading span and the combination of all three working-memory tasks were worst suited for workload prediction in complex learning tasks (on average: C-D = 50.8 %; C-E = 49.7 %; ABC-D = 50.5 % and ABC-E = 50.3 %). For the C-D and C-E cross-task classification, no significant performance ($p > 0.05$, permutation test) could be reached on individual basis. For the train-test combination ABC-D 2 out of 20 subjects, whereas for ABC-E merely 1 out of 20 subjects achieved a significant classifier performance ($p < 0.05$, permutation test). Based on the cognitive rating scales it can be assumed, that the go/no-go task is much easier compared to the other working-memory tasks. Therefore, the train-test combination B-D, as well as, BC-E was additionally conducted. For the train-test combination BC-D, 9 out of 11 subjects reached classification accuracies $\geq 80\,\%$ ($p < 0.05$, permutation test), whereas for the train-test combination BC-E 8 out of 14 subjects achieved a significant

Table 7.2: Cross-task classification accuracies in % for a two-class problem for all subjects. The bold accuracies represent the significant accuracies ($p < 0.05$, permutation test). Row 2 exhibits the datasets used for classifier training. Row 3 represents the datasets used for classifier testing. A = go/no-go, B = n-back, C = reading span, D = arithmetic, E = algebra, ABC = combination of 3 working-memory tasks and BC = combination of 2 working-memory tasks. In each cross-task classification combination, level 1 vs. level 3 were classified.

	Cross-task Classification - Objective Labeling															
Train	A				B				C				ABC		BC	
Test	B	C	D	E	A	C	D	E	A	B	D	E	D	E	D	E
Sub 1	68	68	50	50	67	60	30	45	50	49	45	50	35	50	–	**93**
Sub 2	39	49	55	70	32	68	55	45	50	50	50	50	55	55	–	57
Sub 3	35	45	55	35	42	19	50	50	50	50	50	50	40	45	–	**70**
Sub 4	59	50	60	45	54	21	55	50	50	50	50	50	60	50	**93**	**90**
Sub 5	32	21	25	50	49	21	**65**	**70**	50	50	45	55	**75**	45	**90**	55
Sub 6	36	36	50	50	22	**83**	35	55	53	54	55	50	55	45	13	63
Sub 7	59	28	50	50	52	45	50	50	50	50	50	45	45	35	–	–
Sub 8	49	48	45	55	47	57	50	45	50	50	50	50	60	45	**93**	**87**
Sub 9	40	48	**65**	**75**	**67**	78	70	65	50	53	60	55	60	**75**	–	–
Sub 10	49	15	60	25	30	71	50	**70**	50	50	50	50	50	50	–	–
Sub 11	57	53	45	45	57	61	**65**	50	50	50	50	50	45	45	**87**	47
Sub 12	51	9	55	**65**	47	10	45	**60**	48	49	50	50	55	60	**97**	**90**
Sub 13	57	78	–	45	47	**83**	–	40	51	50	–	45	–	50	–	**90**
Sub 14	57	58	55	45	67	38	**70**	40	48	50	50	50	50	45	–	–
Sub 15	52	34	50	50	61	69	50	55	50	50	50	50	40	50	**93**	**83**
Sub 16	50	40	50	50	50	56	55	45	50	50	50	50	50	50	**93**	**77**
Sub 17	50	24	45	–	50	45	40	–	47	50	50	–	25	–	**83**	–
Sub 18	47	50	55	55	52	60	50	55	48	50	45	45	65	50	**93**	63
Sub 19	63	48	**80**	55	48	**85**	75	55	50	50	55	50	70	55	30	–
Sub 20	46	51	50	60	**66**	49	45	55	50	50	50	50	45	50	–	–
Sub 21	55	46	55	45	42	64	65	50	50	46	60	50	**70**	55	–	27
Mean	49.7	42.7	52.8	51	50	54.3	53.3	52.4	49.9	50	50.8	49.7	50.5	50.3	–	–

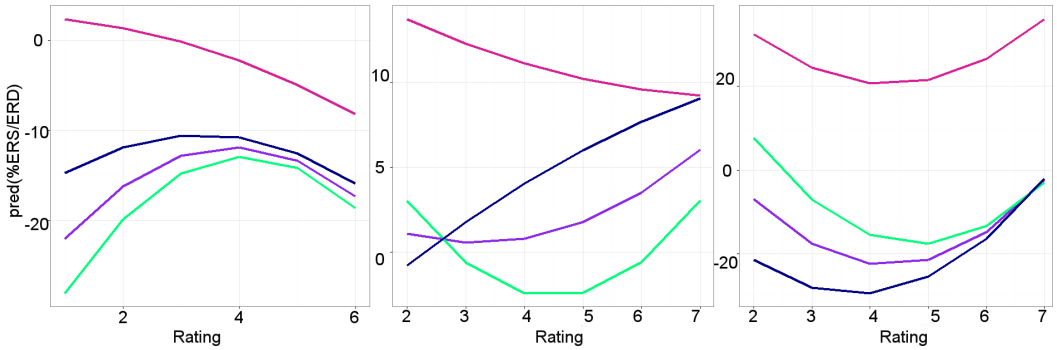


Figure 7.3: The predicted $\%ERS/ERD$ values for each frequency band (alpha = violet, lower alpha = green, upper alpha = blue, theta = red) against the cognitive workload rating scale, averaged over all subjects. **Left:** Go/no-go task, **Center:** n-back task, **Right:** Reading span task.

classification accuracy $\geq 70\,\%$ ($p < 0.05$, permutation test) on individual basis. Since no EEG data from subject 13 was available while working on the arithmetic task and from subject 17 while solving the algebra task, no cross-task classifications based on that data could be performed for these subjects.

7.8 Non-linearity effect in the EEG data

The LMMs were calculated for each frequency band, averaged over all subjects, to analyze if dependent individual differences influence the workload prediction.
In Figure 7.3 and Figure 7.4 the predicted power values are plotted against the cognitive workload rating scale for each task, averaged over all subjects. For the working-memory tasks, Figure 7.3 shows a synchronization of the theta-power compared to the alpha-power for each task. The scale on the y-axis differs for all three working-memory tasks. This is an indicator for the different amount of workload needed for solving the diverse tasks. Furthermore, a non-linear effect depending on the subjective workload rating can be noticed for each frequency band independent on the task. Especially for the lower alpha-band power the prediction is not linearly increasing with the increase of the cognitive rating scale. The predicted power values have an U-shaped pattern. For small rating values the predicted power is the same as for large rating values. The U-shaped pattern of each frequency band in the go/no-go task is the other way around, compared to the n-back and reading span task.
Figure 7.4 reports the predicted power values for the complex learning tasks. For the arithmetic task, no clear synchronization of the theta-frequency can be recognized compared to the alpha-frequency bands. Moreover, the patterns for both word math problems differ from the prediction patterns for the working-memory tasks. Merely for the lower alpha-frequency band a similar course as in the go/no-go task can be recognized. Furthermore, the proceeding of the upper alpha- and theta-frequency band during the arithmetic task is similar to the prediction pattern for the reading span task. The predicted power values for

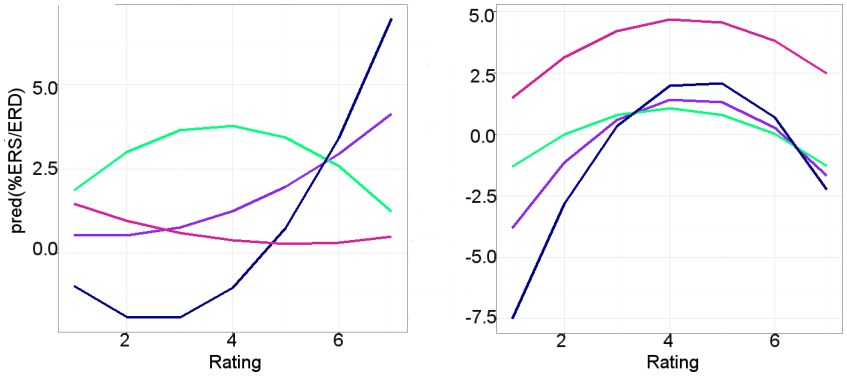


Figure 7.4: The predicted $\%ERS/ERD$ values for each frequency band (alpha = violet, lower alpha = green, upper alpha = blue, theta = red) against the cognitive workload rating scale, averaged over all subjects. **Left:** Arithmetic task, **Right:** Algebra task

the upper alpha- and theta-frequency band in the algebra task is the opposite and resembles the prediction patterns for the go/no-go task. Although no significant results could be achieved by this analysis, two important observations could be made. First, the power of the lower and upper alpha-frequency bands operate in a contrary way. Second, for nearly all frequency bands the predicted power averaged over all subjects resulted in an U-shaped curve. This leads to the assumption, that there are non-linear effects in the EEG data, depending on the difficulty level of each task.

7.9 Discussion

This chapter examines how feasible it is to develop EEG-based classifiers based on student's workload in working-memory tasks and to use those classifiers in more complex learning tasks. Applying this method should enable to classify neural signatures of single executive functions in working-memory tasks as well as in learning tasks. The innovation of using working-memory tasks for classifier training is, that it is known which executive functions are loaded without perceptual confounds. This makes working-memory tasks ideally suited for cross-task classification. The SVM performs well for the within-task classification of workload levels during solving working-memory tasks. The results for the within-task classification of arithmetic and algebra tasks were not very successful, which could be due to overwhelmed subjects. Furthermore, the accuracy drops off greatly when cross-task classification is used. Moreover it was recognized, using a combination of at least two working-memory tasks for classifier calibration led to better classification accuracies.

7.9.1 Absolute difficulty varies across tasks

Workload manipulation is relative within a task, this is a substantial issue. For each task, the problems were divided into low, medium and high levels of difficulty. A hypothesis regarding the relatively poor cross-task classification results is, that the absolute difficulty varies across tasks, but also within a task for one subject, which would significantly influence the performance. Therefore, subjective cognitive workload ratings were collected after each trial. No correlation was found between the predetermined levels and the subjective workload ratings. Hence, classes of learning tasks should not necessarily be formed based on objective complexity, but should represent that the same task could elicit different levels of cognitive workload for different learners based on levels of expertise and working-memory capacity. This leads to the assumption, that labels for the trained tasks were not optimal and classes based on subjective ratings of cognitive workload seem to be more appropriate for classifier training and classifier testing.
In future studies, working-memory tasks, as well as complex learning tasks, have to be modified in their complexity to invoke similar patterns in EEG.

7.9.2 Solving strategies varies across tasks

Solving strategies can vary across the tasks and across degrees of difficulty. Workload independent EEG activity between working-memory and complex learning tasks, as well as between diverse levels of difficulty within one task, varies to much for a precise cross-task classification. Therefore, the classifier training should not be based on exclusively one working-memory task. Instead, combinations of at least two working-memory tasks should be used, preventing the classifier from picking up task-specific features [8]. These tasks should differ with regard to the executive functions they require (updating, shifting, inhibition) and with regard to the representational codes (numbers, letters, words etc.) they involve in order to train a generic classifier sensible to general changes in the requirements. This can be ensured by selecting working-memory tasks which only have a weak correlation with each other [8]. These assumptions were confirmed in the studies postulated in this chapter. A mixture of different working-memory tasks for classier training led to the best classification results, compared to cross-task classifications where only a single task was used for calibrating the classifier. Especially a combination of n-back and reading span tasks for classifier training was most promising to differentiate diverse levels of difficulty, for arithmetic and algebra tasks.

7.9.3 Presentation order of tasks

The main limitation is that diverse tasks were presented in the same order and that three levels of difficulty were presented at advancing levels of difficulty for each task. There may be a strong order effect in the data, which may bias the analysis due to multiple possible confounding factors (e.g., drying gel, user's fatigue, learning effect, slow drifts in the EEG). Indeed, the difficult level was always measured after the easy level, so EEG signals between these two difficulty levels may change irrespectively of the actual workload level. Furthermore, workload is a highly transient learner characteristic that changes

during learning not only due to variations in instructional materials and task requirements but also due to a learner's increasing knowledge which allows the subjects to understand more and more complex contents over time with the same number of activated knowledge structures. It is not clear whether the classification performances are due to workload estimation, the order effect or both. An additional study with the same learning material should be accomplished in future, to control for the confounding factor.

7.9.4 Non-linearity effects in the EEG data

By applying LMMs, the predicted power of the lower and upper alpha-frequency bands operate in a contrary way, this was also postulated by Klimesch [17]. While the lower alpha-frequency band reflects different types of attentional demands, the upper alpha-frequency band responds selectively to semantic long term memory demands. Furthermore, the workload prediction based on the subjective rating cause U-shaped patterns for nearly all frequency bands and tasks. This leads to the assumption of having non-linearity effects in the EEG data. Although these effects are not significant, it can be hypothesized that the difficult tasks were overwhelming for subjects, resulting in disengagement and thus in the respective neural signatures of low cognitive workload. Chanel et al. [154] postulated a further increase in workload might lead to disengagement, which is detectable by a reversed EEG pattern (i.e., theta-desynchronization and alpha-synchronization). This observation is in line with the findings for the LMM frequency band predictions.

7.9.5 Comparison with the state of the art

The cross-task classification is a new innovative method for a generalized classifier generation. Merely a few studies are dealing with cross-task classification [110, 111]. Baldwin et al. [110] used well defined workload tasks for cross-task classification and reached classification accuracies around chance level for a small subject size of 5. In contrast, Gevins et al. [111] reached classification accuracies $> 90\,\%$, but did cross-task classification by using a verbal n-back task for training and a spatial n-back task for testing and vice versa. In this thesis, this would not be assumed as cross-task classification, because for solving the tasks correctly, the same executive functions have to be used. Moreover, Gevins and colleagues [111] solely used 8 subjects for the cross-task classification. Hence, the study reported in this chapter is the first cross-task classification study during which classifiers were trained on well defined working-memory tasks and subsequently used to predict workload levels in complex learning tasks, for a larger group of 20 subjects. Although the initial performance for the cross-task classification was low, it could be observed that using at least two working-memory tasks for classifier calibration improved the cross-task classification accuracies significantly for individual subjects. These modifications lead to higher prediction accuracies as they were postulated in [110].

7.10 Conclusion

In conclusion, an important precondition for the aim of developing workload-adaptive instructional environments seems to be fulfilled. The availability of a continuous and non-intrusive assessment of workload during solving realistic tasks is practicable. But the results have shown, a cross-task classification is possible but prone to confoundations. The controlled working-memory tasks (used for classifier training) induce a different level of workload, compared to the realistic learning tasks (used for classifier testing). This is a natural implication of using different tasks for classifier training than for classifier testing. Furthermore, the linear task order can lead to non-stationarities in the data. Thus, additional studies should be conducted to examine the influence on diverse individual workload capacity ranges and task order effects for cross-task classification.

8 Subjective cognitive workload labeling during cross-task classification

As supposed in chapter 7, the same tasks could elicit different levels of cognitive workload for different learners based on levels of expertise and working-memory capacity. For instance, in a block of 20 learning tasks of similar complexity, the first ten tasks might impose rather high levels of workload onto learners due to novelty, whereas the last ten tasks might impose lower levels of cognitive workload onto learners due to learned solving strategies. Furthermore, varying difficulties can be perceived across tasks during the same difficulty level, e.g., go/no-go tasks seem relatively easy, compared to n-back tasks. This leads to the assumption that objective labels as used in chapter 7 were not optimal for differentiating the diverse workload states. Labeling classes based on subjective cognitive workload ratings are hypothesized to be more appropriate for classifier training and testing. Thus, in the following sections it will be explored, if the modified class labeling method that utilizes subjective cognitive workload ratings can improve the cross-task classification results reported in chapter 7. The results reported in the following sections were partially presented or published in [8, 155].

8.1 EEG data and paradigm

To evaluate if the subjective cognitive workload labeling can enhance the performance of the cross-task classification, the data recorded in chapter 7 were reanalyzed. Thus, the study design, the participants, as well as the data pre-processing steps are similar to the previous chapter 7.

8.2 Labeling

In supervised classification methods, labeled data consists of multidimensional feature vectors and for each vector, a label corresponding to a category exists [156]. In the present study, the subjective cognitive workload labeling method is applied.

8.2.1 Objective labeling

In the previous study, all tasks were presented in three types of difficulty. Using the objective labeling method, each trial got the level of their difficulty as class label. This method is independent of subjective sensations.

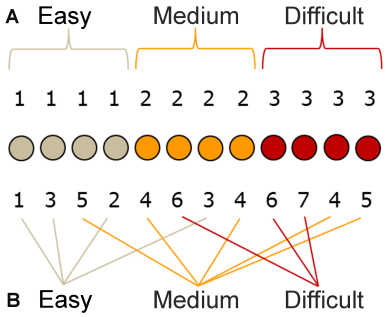


Figure 8.1: Twelve trials are divided into three classes of difficulty by using **A:** objective class labeling. The objective difficulty scale ranges from "1" being easy and "3" being difficulty. **B:** subjective cognitive workload labeling. The cognitive workload scale ranges from "1" being too easy up to "7" being too hard.

8.2.2 Subjective cognitive workload labeling

In order to define classes for classifier training and testing, the relation between difficulty levels of diverse tasks and the respective cognitive workload ratings were analyzed. Classes of learning tasks should not necessarily be formed based on objective complexity, but should represent that the same task could elicit different levels of workload for various learners. As a consequence, subjective cognitive workload ratings were used for defining classes (see Figure 8.1).
For this purpose, the subjectively experienced cognitive workload of all working-memory tasks was used to calculate a mean cognitive workload value for each subject. This value was used to define all tasks into two classes, low versus high level of workload according to the cognitive workload ratings. All trials with a subjective cognitive workload labeling smaller than the mean cognitive workload value for each subject were assigned to "class easy". Further, all trials with a cognitive workload labeling equally or greater than the mean cognitive workload value for each subject were assigned to "class difficult".

8.3 Methods for classification

For classifying, a SVM was applied on data, labeled with subjective cognitive workload ratings. The general classification parameters were similar to those in section 7.3. Furthermore, within-task classification and cross-task classification as described in section 7.4 were comparatively applied for each subject individually.

8.4 Results of subjective cognitive workload rating

In this section, the subjective cognitive workload ratings for each task will be reported. The average ratings are stated in Figure 8.2.

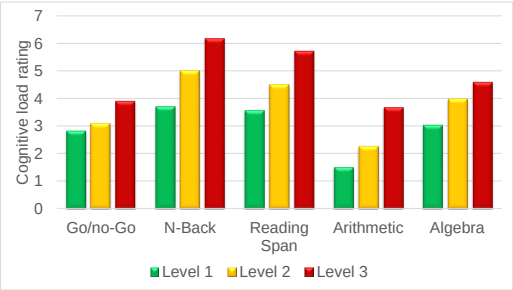


Figure 8.2: Subjective cognitive workload rating for each task and difficulty level, averaged over all subjects.

Since the working-memory tasks were conducted as block design the subjects were asked for a cognitive workload rating ("1" being too easy up to "7" being too hard) after each level of difficulty. For the math word problems, the participants had to rate their cognitive workload after each trial.

Go/no-go task

For the go/no-go task, the mean subjective workload rating value was 2.8 for the easy, 3.09 for the medium and 3.9 for the difficult condition. The subjective rating values of level 1 vs. level 3 and level 2 vs. level 3 differed significantly (level 1 vs. level 3: $p < 0.05$, paired, two-sided Wilcoxon test; level 2 vs. level 3: $p < 0.05$, paired, two-sided Wilcoxon test). There was no significant difference in subjective rating values between level 1 and level 2 ($p > 0.05$, paired, two-sided Wilcoxon test).

N-back task

The average subjective workload rating values for the n-back tasks differed highly significant (for all three combinations: $p < 0.0001$, paired, two-sided Wilcoxon test), with 3.68 for the 1-back, 5 for the 2-back and 6.18 for the 3-back task.

Reading span task

For the reading span task, the mean subjective workload rating value was 3.54 for the easy, 4.5 for the medium and 5.73 for the difficult condition. The subjective workload rating values of level 1 vs. level 3 as well as for level 2 vs. level 3 differed highly significant (for both combinations: $p < 0.0001$, paired, two-sided Wilcoxon test), whereby level 1 vs. 2 differs significantly ($p < 0.05$, paired, two-sided Wilcoxon test).

Arithmetic task

During arithmetic tasks participants rated level 1 with 1.5, level 2 with 2.26 and level 3 with 3.68 on average. All three difficulty levels differed highly significant in the subjective workload ratings ($p < 0.0001$, paired, two-sided Wilcoxon test).

Algebra task

The average subjective workload rating in the algebra task was 3.02 for level 1, 3.83 for level 2 and 4.59 for level 3. Subjective rating values of level 2 vs. level 3 differed significantly ($p < 0.05$, paired, two-sided Wilcoxon test), however level 1 vs. 2 as well as level 1 vs. 3 differed highly significant (for both combinations: $p < 0.0001$, paired, two-sided Wilcoxon test).

8.5 Classification results

The within-task classification results are shown in section 8.5.1, Table 8.1 and the cross-task classification results are stated in section 8.5.2, Table 8.2.

8.5.1 Within-task classification

As in section 7.7.1 the classifier was trained and tested using EEG data recorded during the same tasks (datasets A, B, C, D and E). Furthermore a 10-fold cross validation was used to estimate the accuracy of the trained SVM to guarantee that training and test data do not overlap. This section will present the within-task results utilizing subjective cognitive workload labeling. The result for within-task classification of $\%ERS/ERD$ are shown in Table 8.1.
Using the subjective cognitive workload ratings for feature labeling, a classification for all subjects and all tasks was possible, instead of subject 7 since this subject rated each task with value 4. The rearrangement of the trials in two classes was not possible for this subject. The working-memory tasks reached nearly the same within-task classification accuracy rates as using the objective labeling in the previous chapter (see Table 7.1). With 96.55 % prediction accuracy the go/no-go task had again the highest accuracy over all subjects. The n-back task performance reached 94.86 % accuracy over all subjects, whereas the reading span task reached a classifier performance of 93.63 % over all subjects. The complex learning tasks reached lower performance accuracies. The arithmetic tasks were classified with an average accuracy of 68 % and the algebra tasks with 67.83 % on average. In all cases within-task classification worked well and reached significant classification accuracies ($p < 0.05$, permutation test) for the working-memory tasks on average. The EEG data of subject 13 was not recorded properly while solving the arithmetic task, therefore, no within-task classification results can be reported for this task. Subject 17 quit the experiment after solving the arithmetic task, therefore no within-task classification results can be reported for this subject during the algebra task, neither.

8.5.2 Cross-task classification

The cross-task classification was accomplished similar to section 7.7.2. The SVM was trained by using dataset A, B, C, ABC, or BC and then tested on EEG data from remaining workload tasks or from complex learning tasks (dataset D and E). The results for cross-task classification of the $\%ERS/ERD$ values are shown in Table 8.2. Average values will not be reported, as mean accuracies were based on different numbers of subjects for each

Table 8.1: Within-task classification accuracies in % for a two-class problem for all subjects, with subjective cognitive workload labeling. The bold values represent the significant accuracies ($p < 0.05$, permutation test). Row 2 exhibits the datasets used for classifier training. Row 3 represents the datasets used for classifier testing. A = go/no-go, B = n-back, C = reading span, D = arithmetic, E = algebra. "–" indicates when a rearrangement of trials in two classes was not possible.

	Within-task Classification				
Training	A	B	C	D	E
Test	A	B	C	D	E
Sub 1	**95.96**	**93.52**	**92.5**	**66.67**	**73.33**
Sub 2	**-**	**91.4**	**96.67**	46.47	66.67
Sub 3	**97.98**	**91.92**	**91.67**	63.33	66.67
Sub 4	**98.96**	**91.42**	**85.83**	**80**	56.67
Sub 5	**96.97**	**96.43**	**96.67**	53.33	60
Sub 6	**100**	**96.19**	**92.5**	66.67	63.33
Sub 7	–	–	–	–	–
Sub 8	**97.98**	**95.24**	**95.23**	56.67	**73.33**
Sub 9	**93.75**	**92.16**	**96.67**	**80**	66.67
Sub 10	**97.98**	**96.87**	**96.67**	**76.67**	63.33
Sub 11	**99.02**	**96.08**	**96.67**	40	53.33
Sub 12	–	**94.44**	**96.67**	60	63.33
Sub 13	**92.16**	**95.37**	**90**	–	**90**
Sub 14	**92.16**	**98.04**	**88.33**	**66.67**	**73.33**
Sub 15	**93.75**	**93.75**	**89.17**	**73.33**	**76.67**
Sub 16	**98.98**	**91.67**	**89.17**	**73.33**	**70**
Sub 17	**81.82**	–	**96.67**	66.67	–
Sub 18	**96.87**	**94.44**	**96**	**90**	50
Sub 19	**100**	**100**	**96.67**	**76.67**	50
Sub 20	**94.79**	**96.19**	**89.17**	63.33	**76.67**
Sub 21	**97.98**	**86.9**	**94.17**	60	63.33
Mean	**96.55**	**94.86**	**93.63**	68	67.83

task, which cannot be compared. In general, predicting cognitive workload during complex learning tasks based on classifiers trained with working-memory tasks seems to be challenging.

The diverse tasks differed substantially with regard to the level of workload. Therefore, the difficulty levels of each trial in the training and testing tasks were adjusted, based on the subjective cognitive workload rating. For some subjects, the rearrangement of trials in two classes (easy vs. difficult) lead to an unbalanced set of data, so that no efficient classifier training was possible and no significant results could be reached. Subject 7 rated

Table 8.2: Cross-task classification accuracies in % for a two-class problem for all subjects, with subjective cognitive workload labeling. The bold accuracies represent the significant accuracies ($p < 0.05$, permutation test). Row 2 represents the datasets used for classifier training, whereas row 3 exhibits the datasets used for classifier testing. A = go/no-go, B = n-back, C = reading span, D = arithmetic, E = algebra, ABC = combination of 3 working-memory tasks and BC = combination of 2 working-memory tasks. "–" indicates when a rearrangement of trials in two classes was not possible. In each cross-task classification combination, level 1 vs. level 3 was classified except in the last two columns.

	Cross-task Classification - Subjective Cognitive Workload Labeling															
Train	A				B				C				ABC		B (1-2) C (1-3)	
Test	B	C	D	E	A	C	D	E	A	B	D	E	D	E	D (2-3)	E (1-2)
Sub 1	36	37	**93**	70	–	67	3	10	–	56	43	20	33	43	**95**	**95**
Sub 2	–	–	–	–	–	44	–	57	–	–	**93**	23	**90**	27	–	40
Sub 3	–	**51**	**50**	37	–	33	–	37	–	60	33	53	**83**	43	**85**	40
Sub 4	–	–	**50**	–	–	50	**60**	**57**	–	51	33	**50**	63	50	**90**	60
Sub 5	–	–	–	–	–	–	30	15	–	50	90	55	45	20	–	–
Sub 6	66	64	20	**67**	**67**	**73**	13	**63**	59	**69**	27	**70**	30	63	20	50
Sub 7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Sub 8	–	–	**67**	27	–	–	–	**87**	–	33	**93**	**87**	**77**	**73**	**90**	**90**
Sub 9	–	–	37	33	–	**69**	73	73	–	69	**90**	**80**	**83**	47	**85**	**85**
Sub 10	–	–	20	30	–	71	–	75	–	–	**75**	40	70	55	–	–
Sub 11	–	–	30	**70**	32	45	27	**63**	**67**	33	**77**	40	47	27	70	50
Sub 12	–	–	–	–	–	35	7	**50**	–	**68**	**97**	**83**	**97**	**83**	**95**	**80**
Sub 13	47	63	–	77	42	**52**	–	17	–	–	–	73	–	77	–	**95**
Sub 14	32	31	**87**	**87**	–	**68**	**63**	**60**	–	**77**	43	30	**87**	**87**	**80**	–
Sub 15	61	–	50	**77**	–	66	17	70	–	48	37	67	67	37	**85**	35
Sub 16	–	–	30	3	–	33	10	30	–	33	26	57	**77**	30	20	**70**
Sub 17	–	–	70	–	–	–	–	–	–	–	**80**	–	**73**	–	**85**	–
Sub 18	28	26	**73**	33	–	**73**	13	43	–	**67**	**57**	30	**63**	50	**90**	60
Sub 19	–	48	35	20	–	**85**	**70**	60	–	50	**65**	**60**	**65**	**65**	–	–
Sub 20	–	–	47	33	–	–	–	**93**	–	43	–	–	–	**97**	–	–
Sub 21	32	**68**	–	37	–	–	17	**57**	**67**	33	–	40	–	53	–	**85**

each task with the same difficulty value. Thus, a rearrangement of trials in two classes was not possible for these subjects. By using a SVM with RBF-kernel, two levels of workload of the learning tasks could be successfully classified on a single-subject single-trial basis. Compared to the objective labeling method in the previous chapter 7 (see Table 7.2), the majority of combinations reached a significant classifier performance ($p < 0.05$, permutation test) (see Table 8.2). On individual basis the train-test combination ABC-D, ABC-E and C-D reached best classification accuracies, with 97 %. Regarding the within-task classification, cross-task classification using go/no-go tasks for training or testing data only, is possible for just a few subjects (for training: max. 15 out of 20 subjects; for testing: 3 out of 20 subjects). As the subjective rating threshold is calculated as mean of the rating values from the training data, the boundary can be too high or too low for the go/no-go task. Since the cognitive workload ratings for the three levels in the go/no-go tasks do not differ significantly, the test-data will only be assigned to one class, based on its subjective rating values. Therefore, the train-test combination BC-D, as well as, BC-E were additionally conducted. For SVM training the difficulty levels 1 and 2 from the n-back task, as well as level 1 and 3 for the reading span task were chosen. The classifier was tested based on arithmetic tasks of difficulty levels 2 and 3, as well as for the difficulty levels 1 and 2 of the algebra task. For the arithmetic cross-task condition (BC-D), 10 out of 20 subjects reached significant classification accuracies $\geq 80\,\%$ ($p < 0.05$, permutation test). Furthermore, 7 out of 20 subjects reached significant classification accuracies ≥ 70 % for the train-test combination BC-E ($p < 0.05$, permutation test).
Since no EEG data from subject 13 was available while working on the arithmetic task and from subject 17 while solving the algebra task, no cross-task classifications based on that data could be performed.

Area under the curve for performance prediction

Using the cognitive workload rating to rearrange the trials in new difficulty levels leads to unbalanced classes. A well-known performance measurement for unbalanced classes is the area under the curve (*AUC*). The *AUC* was additionally computed for the most promising combinations (BC-D and BC-E) to analyze, if the good accuracy levels are solely based on the classes being unbalanced or because of the modified labeling method. Unlike the accuracy results, the *AUC* were not significant for all subjects ($AUC < 0.6, \text{BC-D}: \varnothing = 0.43, \text{BC-E}: \varnothing = 0.41$).

8.6 Discussion

This chapter explored, if using subjective cognitive workload ratings for class labeling can improve the classification results reported in chapter 7. As in the previous chapter, high classification accuracies were reached for the within-task classification for EEG data recorded during working-memory tasks, whereas the results for the within-task classification of arithmetic and algebra tasks were not highly successful. This might be due to overwhelmed subjects. Again, the accuracy drops off when cross-task classification is used. But compared to the results of section 7.7.2, an improvement in the classification accuracy by using the subjective cognitive workload rating can be noticed.

The combination of n-back and reading span task for classifier calibration led to best classification accuracies for workload prediction in complex learning tasks.

Unbalanced classes for classification

Workload manipulation is relative within a task, this is a substantial issue. To counteract these phenomena, the method of labeling features based on subjective cognitive workload ratings was developed, which indeed causes an improvement of classification performance. The major drawback of rearranging trials based on subject ratings is: classes are not balanced anymore. Since classification accuracy is only appropriate for a balanced number of trials per class [157], it is not reliable as performance measurement for this modified method. With an unbalanced number of trials per class, changes in classification accuracy can simply be due to the classifier being biased towards one class, rather than to any actual performance change. In case of the stated complex learning tasks where merely ten trials per class are available, it can have the following consequences: if one class has only a single trial more (or less) compared to the other class, it can already change the accuracy by 5 % [158]. Therefore, the *AUC* was additionally used, as this is a stable performance measurement if classes are unbalanced. As calculating the *AUC*, did not lead to significant results it can be assumed, that the accuracy improvement of the modified labeling method is actually based on the unbalanced number of trials. Although the initial performance for cross-task classification could be improved by applying the cognitive workload labeling method, it could not be ruled out that it is just based on unbalanced data.

8.7 Conclusion

To conclude, using subjective cognitive workload ratings is a new and innovative idea, which has potential to improve cross-task classifications. But the number of trials should be big enough, so rearranging the difficulty levels should not lead to unbalanced classes and still enough data for classification. To my knowledge, this is the first study modifying the labeling process for SVMs by using subjective cognitive workload ratings.

9 Task order effect in cross-task classification

The main limitation in the task design of chapter 7, is diverse tasks were presented in the same order and the three levels of difficulty were presented at advancing levels of difficulty for each task. There might be a strong order effect in the data, which may bias the analysis due to multiple possible confounding factors (e.g., drying gel, user's fatigue, learning effect, slow drifts in the EEG). Furthermore, workload changes during learning not only due to variations in instructional materials and task requirements, but also due to a learner's increasing knowledge. This allows the subjects to understand more and more complex contents over time with the same amount of workload. Indeed, the difficult level was always performed after the easy level, so EEG signals between these two difficulty levels may change irrespectively of the actual workload level. As it is not clear whether the classification performances in chapter 7 and 8 are due to workload estimation, the order effect, or both, an additional study was conducted in this chapter, with presenting the learning material in a randomized order.

9.1 Study design

In the following, the participants data will be introduced and the methods for EEG recording, the task design, as well as the procedure, will be described.

9.1.1 Participants and EEG recordings

A total of 12 subjects, 7 male and 5 female aged between 19 and 29 (median age 23), voluntarily participated in the EEG experiment. All participants signed a written informed consent. All of them were native or fluent in German. For comparison reasons, the EEG recording settings were similar to those of the previous study in chapter 7. 30 active electrodes (actiCap, BrainProducts GmbH), attached to the scalp, placed according to the extended International Electrode 10 - 20 Placement System [36], were used to record EEG signals. Two additional electrodes record an EOG; one placed horizontally at the outer canthus of the right eye to measure horizontal eye movements and one placed in the middle of the forehead between the eyes to measure vertical eye movements. The reference electrode was placed on the left mastoid, the ground electrode at AFz. EOG and EEG signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp, g.tec). The sampling rate was 512 Hz and the impedance of each electrode was smaller than 5 kΩ. EEG data was high-pass filtered at 0.1 Hz and low-pass filtered at 100 Hz during the recording.

Furthermore, a notch-filter was applied between 48 Hz – 52 Hz to filter power line noise. The EEG data pre-processing steps are similar to section 7.2.

9.1.2 Paradigm

The used tasks, as well as the data pre-processing steps, were similar to the previous chapter 7. As the current study was conducted to analyze the effect of non-stationarities in EEG data, all tasks were presented in a pseudo-randomized order of difficulty. The go/no-go tasks did not lead to meaningful results in the previous studies, thus it was not recorded in this study anymore. Resulting in a paradigm with four different tasks, i.e., n-back, reading span, arithmetic and algebra tasks.

9.2 Classification using a randomized task order

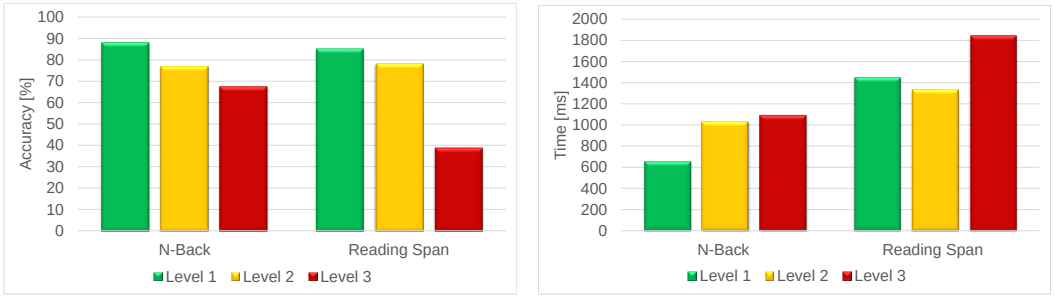
The general classification parameters were similar to those in section 7.3. Furthermore, within-task classification and cross-task classification as described in section 7.4 were comparatively applied for each subject individually. Additionally, the effect of subjective workload labeling compared to objective labeling (see chapter 8) was analyzed.

9.3 Behavioral results

In this section, the performance results and the subjective cognitive workload rating for each task will be reported.

9.3.1 Performance results

The response accuracies and response times of each working memory task can be seen in Figure 9.1.



(a) Response Accucary

(b) Response Time

Figure 9.1: Performance results for each task and difficulty level, averaged over all subjects.

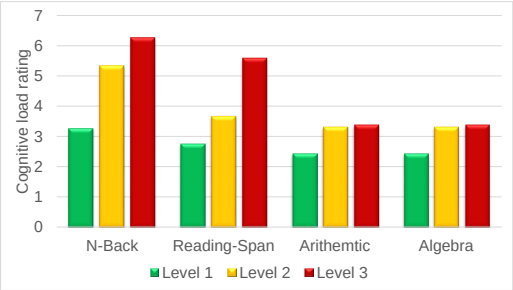


Figure 9.2: Subjective cognitive workload rating for each task and difficulty level, averaged over all subjects.

N-back task

In the n-back, the mean reaction time was 657 ms for the easy, 1030 ms for the medium and 1094 ms for the difficult condition. Participants answered 87.84 % of trials in the 1-back, 76.75 % in the 2-back and 67.31 % in the 3-back correctly. The mean response times, as well as the mean response accuracies differed significantly (for all combinations $p < 0.05$, paired, t-test).

Reading span task

During the reading span task the mean reaction time was 1446 ms for level 1, 1333 ms for level 2 and 1844 ms for level 3 (level 1 vs. level 3, as well as level 2 vs. level 3 differed highly significant: $p < 0.05$ paired, t-test). The response accuracy of remembered consonants was 85 % in level 1, 78 % in level 2 and finally dropped to 39 % in level 3. The response accuracies differed highly significant for all combinations ($p < 0.01$, paired, t-test).

Realistic learning tasks

Neither in the arithmetic nor in the algebra tasks reaction times were measured because speed was not relevant for our experiment in these tasks. Due to technical problems no reaction accuracies can be reported for the word math problems.

9.3.2 Subjective cognitive workload rating

Since the working-memory tasks were conducted as block design the subjects were asked for a cognitive workload rating (“1” being too easy up to “7” being too hard) after each level of difficulty. For the math word problems, the participants had to rate their cognitive workload after each trial. The average ratings are stated in Figure 9.2.

N-back task

For the n-back task, the mean cognitive workload rating increases from 3.25 for the easy, to 5.33 for the medium and to 6.25 for the difficult condition. All combination of ratings differed significantly ($p < 0.05$, paired, t-test).

Reading span task

The average rating values for the reading-span tasks differed significantly (for all three combinations: $p < 0.05$, paired, t-test), with 2.75 for level 1, 3.67 for level 2 and 5.58 for level 3.

Realistic learning tasks

For the arithmetic task, the mean cognitive workload rating was 2.43 for the easy, 3.23 for the medium and 3.39 for the difficult condition. During algebra tasks participants rated level 1 with 2.43, level 2 with 3.23 and level 3 with 3.39. For both word math problems, level 1 vs. level 2, as well as level 1 vs. level 3 differed significantly ($p < 0.05$, paired, t-test).

9.4 Classification results

The within-task classification results based on objective labeling and subjective cognitive workload labeling are stated in section 9.4.1, Table 9.1. The cross-task classification results using objective labeling and subjective cognitive workload labeling are shown in section 9.4.2, Table 9.2 and Table 9.3.

9.4.1 Within-task classification

As in section 7.7.1 the classifier was trained and tested using EEG data recorded during the same tasks (datasets B, C, D and E). Furthermore a 10-fold cross validation was used to estimate the accuracy of the trained SVM to guarantee that training and test data do not overlap. This section will firstly present the within-task results with objective labeling, followed by the within-task results using subjective cognitive workload labeling (see Table 9.1).

Objective labeling

In the case of objective labeling, the arithmetic task reached the highest accuracy over all subjects with 76.79 %. As shown in Table 9.1 the reading span task had an accuracy of 73.35 % over all subjects, whereas the n-back task performance over all subjects reached 72.69 %. The algebra task reached lower performance accuracies, the average accuracy was 63.62 %. The within-task classification worked well in almost all cases and reached significant classification accuracies ($p < 0.05$, permutation test) in average, except the algebra task ($p > 0.05$, permutation test). Remarkable are the classification results from subject 5, since they are smaller than 50 % for all tasks, except the arithmetic task. Removing the channels Fz and Pz the classification accuracies increased significantly for this subject (on average 67.43 %). Thus, it can be assumed, that the poor results are based on artifacts, measured at these electrodes.

Table 9.1: Within-task classification accuracies in % for a two-class problem for all subjects, with objective as well as subjective cognitive workload labeling. The bold values represent the significant accuracies ($p < 0.05$, permutation test). Row 2 exhibits the datasets used for classifier training, whereas row 3 represents the datasets for classifier testing. B = n-back, C = reading span, D = arithmetic, E = algebra. "–" indicates when a rearrangement of trials in two classes was not possible.

	Within-task Classification							
	Objective Labeling				**Subjective Workload Labeling**			
Training	B	C	D	E	B	C	D	E
Test	B	C	D	E	B	C	D	E
Sub 1	**73.24**	**60.26**	**90.00**	**75**	–	–	–	–
Sub 2	**80.88**	**79.73**	**73.33**	**85.71**	–	–	–	**72.73**
Sub 3	**79.17**	**91.03**	**89.47**	**50**	–	–	–	60
Sub 4	**74.65**	**67.95**	**76.47**	**73.68**	–	–	58.62	48.28
Sub 5	**47.06**	**47.95**	**78.95**	**44.44**	–	–	64.29	57.14
Sub 6	**62.5**	**73.08**	**84.21**	**47.37**	–	–	–	–
Sub 7	**94.44**	**53.95**	**52.94**	72.22	–	–	60	61.54
Sub 8	**79.17**	**82.43**	**80.00**	55.00	–	–	–	41.38
Sub 9	**62.9**	**78.57**	**89.47**	**53.33**	–	–	68.97	70.83
Sub 10	**76.39**	**87.18**	**85.00**	**75.00**	–	–	–	73.33
Sub 11	**66.2**	**85.71**	**90.00**	**66.67**	–	–	73.33	60.71
Sub 12	**75.71**	**72.37**	**31.58**	**65**	–	–	65.52	–
Mean	**72.69**	**73.35**	**76.79**	**63.62**	–	–	–	–

Subjective cognitive workload labeling

Compared to chapter 8, the subjective cognitive workload ratings for class labeling were not so promising in the current study. In some cases the rearrangement of trials in two classes (easy vs. difficult) led to an unbalanced set of data, no efficient classifier training was possible and no significant results could be reached. In the arithmetic tasks subject 2, 3, 6, 8 and 9 were affected by this problem and in the algebra task subject 12. Furthermore, subject 1 and 6 rated each task with the same difficulty value. Thus, a rearrangement of the trials in two classes was not possible for these subjects (see Table 9.1). As in the objective labeling method the arithmetic task reached higher classification accuracies over all subjects (65.12 %) compared to the algebra task (60.66 %). In nearly all applied cases within-task classification worked well, but a significant classification accuracy ($p < 0.05$, permutation test) could only be reached for one subject in the algebra task.

Table 9.2: Cross-task classification accuracies in % for a two-class problem for all subjects, with objective feature labeling. The bold accuracies represent the significant accuracies ($p < 0.05$, permutation test). Row 2 states the datasets used for classifier training, whereas row 3 presents the datasets used for classifier testing. B = n-back, C = reading span, D = arithmetic, E = algebra, BC = combination of 2 working-memory tasks and DE = combination of 2 complex learning tasks. In each cross-task classification combination, level 1 vs. level 3 was classified.

	Cross-task Classification - Objective Labeling						
Train	B		C		BC		
Test	D	E	D	E	D	E	DE
Sub 1	25	35	45	30	35	40	43
Sub 2	**60**	**64**	**60**	36	53	36	48
Sub 3	**84**	40	47	45	**79**	40	49
Sub 4	**94**	53	35	32	**65**	37	53
Sub 5	53	50	53	56	53	56	54
Sub 6	**74**	58	32	**74**	**68**	**63**	58
Sub 7	59	**67**	59	50	**65**	33	43
Sub 8	**75**	45	50	40	**60**	55	**60**
Sub 9	**89**	33	**79**	33	**95**	27	**65**
Sub 10	55	**70**	**70**	60	**65**	**70**	**65**
Sub 11	**60**	33	**65**	39	**75**	39	45
Sub 12	53	**60**	42	40	53	35	46
Mean	65.1	50.7	53.1	44.5	63.8	44.2	52.3

9.4.2 Cross-task classification

Similar to the previous study (see section 7.7.1) a cross-task classification was accomplished. A SVM was trained by using dataset B, C, or BC and then tested on EEG data from independent complex learning tasks (dataset D, E and DE). This section will firstly present the within-task results with objective labeling (see Table 9.2), followed by the within-task results using subjective cognitive workload labeling (see Table 9.3). In general, predicting cognitive workload during complex learning tasks based on classifiers trained with working-memory tasks seems to be challenging.

Objective labeling

As it is known from previous chapters 7 and 8 cross-task classification of diverse working-memory tasks is possible. Thus, the combination of working-memory tasks for classifier training and complex learning tasks for classifier testing is the most interesting combination for an adaptive learning environment. Therefore, data recorded during solving a n-back task or a reading span task (B, C and BC) was utilized for classifier training, whereas data recorded during complex learning tasks (D, E and DE) were used for classifier testing. The cross-task classification results based on objective labeling are shown in Table 9.2.

As in the results part of section 7.7.2, n-back reached best results for workload prediction in complex arithmetic tasks (B-D), with 65 %. For this train-test combination, as well as for BC-D best classification accuracies could be achieved on individual basis, with 94 % and 95 %. Further, 7 out of 12 subjects, as well as 8 out of 12 reached a significant classification performance ($p < 0.05$, permutation test) on individual basis for these train-test combinations. For n-back and algebra data as train-test combination (B-E) a mean accuracy of 50.7 % was achieved. 4 out of 20 subjects reached significant classification accuracies ($p < 0.05$, permutation test) on individual basis for this cross-task classification. The cross-task classification using reading-span for training and arithmetic respectively algebra tasks for testing (C-D, resp., C-E) resulted in 53.1 % and 44.5 % classification accuracies. For C-D only 4 out of 12 subjects, whereas for C-E merely 1 out of 12 subjects achieved a significant classifier performance ($p < 0.05$, permutation test) on individual basis. The combination of both working-memory tasks resulted in lower performance rates than training with a single working-memory task (on average: BC-D =63.8 %, BC-E =44.2 % and BC-DE =52.3 %). For the train-test combination BC-E 2 out of 12 subjects, whereas for BC-DE 3 out of 12 subjects achieved a significant classifier performance ($p < 0.05$, permutation test). None of the classification accuracies averaged over all subjects were significant ($p > 0.05$, permutation test).

Subjective cognitive workload labeling

Using the subjective cognitive load labeling, the classifier was trained on dataset B, C or BC and evaluated on dataset D, E, or DE. The diverse tasks differed substantially with regard to the level of workload they induced. Based on the method stated in section 8.2.2, the trials were rearranged in two classes (easy vs. difficult). For some subjects, this reordering led to an unbalanced set of data, so that no efficient classifier training was possible and no significant results could be reached. This was the case for subjects 2, 3, 8, 10 and 12 in the arithmetic task. Furthermore, subject 1 and 6 rated each task with the same difficulty value, whereas, a rearrangement of trials in two classes was not possible. The results for cross-task classification of the $\%ERS/ERD$ values are stated in Table 9.3. As in chapter 8 no average values will be reported, as mean accuracies were based on different numbers of subjects for each task, which cannot be compared. The majority of classification combinations led to no significant classifications with accuracies around chance level ($p > 0.05$, permutation test). Overall it can be recognized, that the classifier trained on n-back data reached higher classification accuracies, than SVMs trained on data recorded during solving a reading span task. The train-test combination B-D reached stable classification accuracies on individual basis, with a maximum of 72 %. Further, 4 out of 5 subjects reached a significant classification performance ($p < 0.05$, permutation test) on individual basis for these train-test combination. The cross-task classification using n-back for classifier training and algebra tasks for testing (B-E) resulted in accuracies of 70 % on individual basis. Over all subjects lower performance rates were reached, merely 3 out of 10 subjects achieved a significant classifier performance ($p < 0.05$, permutation test). For the train-test combination BC-D and BC-E, 2 out of 12 subjects achieved a significant classifier performance ($p < 0.05$, permutation test). On individual basis a workload prediction with an accuracy up to

Table 9.3: Cross-task classification accuracies in % for a two-class problem for all subjects, with subjective cognitive workload labeling. The bold accuracies represent the significant accuracies ($p < 0.05$, permutation test). Row 2 reports the datasets used for classifier training. Row 3 exhibit the datasets used for classifier testing. B = n-back, C = reading span, D = arithmetic, E = algebra, BC = combination of 2 working-memory tasks and DE = combination of 2 complex learning tasks. "–" indicates when a rearrangement of trials in two classes was not possible. In each cross-task classification combination, level 1 vs. level 3 was classified.

	Cross-task Classification - Subjective Cognitive Workload Labeling						
Train	B		C		BC		
Test	D	E	D	E	D	E	DE
Sub 1	–	–	–	–	–	–	–
Sub 2	–	**64**	–	55	–	**68**	50
Sub 3	–	47	–	43	–	50	61
Sub 4	**72**	48	55	55	**69**	52	**66**
Sub 5	61	**71**	54	57	**64**	**71**	61
Sub 6	–	–	–	–	–	–	–
Sub 7	**68**	62	56	54	56	50	47
Sub 8	–	56	–	52	–	59	53
Sub 9	**66**	33	59	42	59	29	58
Sub 10	–	43	–	57	–	57	53
Sub 11	**67**	54	53	50	57	54	55
Sub 12	–	**79**	–	48	–	59	55

69 % and 71 % was possible. The cross-task classification using the combination of n-back and reading span data for classifier training and the dataset from both complex learning tasks for classifier testing (BC-DE) resulted in low performance rates. Only 1 subject out of 10 reached a significant classifier performance ($p < 0.05$, permutation test) with 66 % classification accuracy. Reading span was worst suited for workload prediction in complex learning tasks. For the C-D and C-E cross-task classification, no significant performance ($p > 0.05$, permutation test) could be reached on individual basis. Further, the classifier performance was around chance level ($< 60\,\%$) for all subjects.

Rearranging the trials into two classes based on subjective cognitive workload ratings led to unbalanced classes. As in the previous chapter 8, the *AUC* was calculated for the most important combinations (BC-D and BC-E). The *AUC* is a more stable and reliable performance measurement for unbalanced data. According to the findings in chapter 8 the *AUC* results were not significant for all subjects ($AUC < 0.65$, BC-D : $\varnothing = 0.48$, BC-E : $\varnothing = 0.44$). The classification accuracies and the *AUC*-values are not significant for the results current study. This might indicate that cross-task classification using these type of working-memory tasks, as well as these complex learning tasks, is challenging.

9.5 Discussion

The study examines, if the task order induces an effect for cross-task classification. A pseudo-randomized order for task presentation was used to ensure, classification performances are based on changes in workload only and not due to order effects (see chapter 7). As in chapter 7, the accuracy using cross-task classification is low. The combination of the n-back and the reading span task for classifier calibration led to best classification accuracies for workload prediction in complex learning tasks on individual basis. But compared to the results of chapter 8, a decrease of the classification accuracy can be noticed, by using the subjective cognitive workload rating. These results led to the assumption, that there might be a strong order effect in the data, which might bias the classification. Thus, the precise classification results are mainly based on non-stationarities over time.

Presentation order of tasks

The main problem in a real learning setting is handling learning effects. Two materials with identical degree of difficulty will impose a different amount of cognitive workload at different times during learning. Additionally, certain materials will be too complex to be understood at the beginning of a learning phase, while they might be quite easy at a later point in the learning phase. This implies, a randomized task order for classifier testing cannot be realized in a real learning setting, because the instructional materials need to be presented in an increasing order of difficulty. Confounding task difficulty and presentation time is inevitable for learning materials. To avoid these confounds in future applications, methods for covariate shift adaptation which alleviate non-stationarities [159, 160] can help to improve classification accuracies.

9.6 Conclusion

In conclusion, the task order of the presented learning material has an effect on the classifier performance. As the presentation of a fixed order of task-difficulty is necessary in real-world learning environments, cross-task classification seem to be too prone to confoundations. Using cross-task classification in future studies, working-memory tasks, as well as complex learning tasks, have to be modified in their complexity to invoke similar patterns in EEG. Furthermore, generalizable classification methods should be used, which are less susceptible to non-stationarities induced by time effects.

10 Cross-subject workload prediction

As the cross-task approach, presented in chapter 7, leads to classification performances around chance level a cross-subject workload prediction based on linear regression will be introduced in this chapter. The possibility to calibrate a generalized regression model independently from each subject is an essential benefit for realizing an adaptive learning environment. The results presented in the following sections are partially published in [100].

10.1 Study design

An introduction to the participants data, the EEG recordings, the task design, as well as the procedure will be given in the following sections.

10.1.1 Participants and EEG recordings

A total of 10 subjects, 4 male and 6 female aged between 17 and 32 (median age 26) were recruited for this EEG experiment. A set of 29 active electrodes (actiCap, BrainProducts GmbH), attached to the scalp and placed according to the extended International Electrode 10 - 20 Placement System [36] (FPz, AFz, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, CPz, P7, P3, Pz, P4, P8, PO7, POz, PO8, O1, Oz, O2), was used to record EEG signals. Three additional electrodes were utilized to record EOG signals; two placed horizontally at the outer canthus of the left and right eye to measure horizontal eye movements and one placed in the middle of the forehead between the eyes to measure vertical eye movements. EOG and EEG signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp, g.tec). The sampling rate was 512 Hz and the impedance of each electrode was $< 5\,\mathrm{k}\Omega$. EEG data was high-pass filtered at 0.1 Hz and low-pass filtered at 60 Hz during the recording. Furthermore, a notch-filter was applied between $48\,\mathrm{Hz} - 52\,\mathrm{Hz}$ to filter power line noise.

10.1.2 Paradigm

Each subject had to solve 240 addition tasks with an increasing level of difficulty while EEG was measured. The math problems were presented in blocks with six different degrees of difficulty. While the first block consisted of 90 easy problems, the following five blocks with increasing difficulty consisted of 30 math problems each. Depending on the level of difficulty, two numbers between one and four digits had to be added, including a different amount of carry effects. Each trial, in which an addition task should be solved, consisted of three phases (see Figure 10.1). First, the calculation phase occurred where

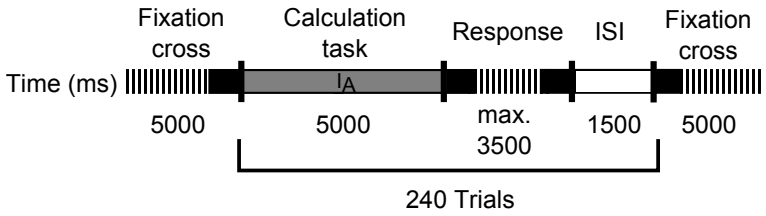


Figure 10.1: Schematic flow of the addition task. The grey line indicates the activation interval (I_A), followed by a black dashed line representing the response phase. The white line indicates the inter-stimulus interval (*ISI*). The parts with the stripped surface indicate shortened time windows for better visualization.

the problem to be solved was shown for 5 sec. Subsequently, subjects had 3.5 sec to type in their result, followed by an inter-trial interval of 1.5 sec, resulting in a total length of approximately 42 min. To avoid the classifier being based on perceptual-motor confounds, the time windows used for analyzing EEG data should not contain motor events. As typing in the answer leads to motoric artifacts, only the calculation phase was used for EEG analysis.

Measure of task difficulty

As postulated from Kantowitz [161], increasing task difficulty always increases workload, since by definition, an increase of task difficulty demands additional workload capacity. Further, an increase of workload is characterized by changes in the theta-, alpha- and beta-frequency bands [51] in the EEG. The presented math problems in this study varied in degrees of difficulty as measured by the information content (Q) of an arithmetic task, according to Thomas [162]. The paper and pencil tasks under which Q was derived were designed to reflect typical mental arithmetic performance, by individually adding the single digits and including the carry in the next single digit addition. Given an arithmetic problem $x+y=?$, with x_i being the i-th digit of x and y_i being the i-th digit of y, the difficulty of the problem $x+y$ will be described by the information content $Q(x+y)$. $Q(x+y)$ is calculated by separating the multi-digit problem into corresponding single-digit problems and calculating the sum of all single digit problems, taking occurring carries into account.

$$Q(x+y) = Q(x_1+y_1) + Q(x_2+y_2+c_2) + \ldots + Q(x_n+y_n+c_n) \tag{10.1}$$

where x_1 is the rightmost digit of number x and x_n the leftmost digit of number x. c_n indicates the carry, coming from the single digit addition before $(x_{n-1}+y_{n-1})$. For an addition problem with two single-digit numbers including no carry effect, the Q-value is generally calculated as follows:

$$Q(x_1+y_1) = log(x_1+y_1+res(x_1+y_1)) \tag{10.2}$$

$$\text{E.g.:} \quad Q(1+2) = log(1+2+3) = 0.78 \tag{10.3}$$

For an addition of two single-digit numbers including a carry effect, the Q-value is calculated by:

$$Q(x_1+y_1) = log(x_1+y_1+res(x_1+y_1)+10+res(x_1+y_1-10)) \tag{10.4}$$

$$\text{E.g.:} \quad Q(9+8) = log(9+8+17+10+7) = log(51) = 1.71 \tag{10.5}$$

If the addition problem is a bit more complex, with x and y being two digit numbers and the carry effect carries over to the next single digit problem, the Q-value is calculated as:

$$\begin{aligned} Q(x+y) &= Q(x_1+y_1)+Q(x_2+y_2+c_2) \\ &= log(x_1+y_1+res(x_1+y_1)+10+res(x_1+y_1-10)) \\ &+ log(x_2+y_2+res(x_2+y_2)+c_2+res(x_2+y_2+c_2)) \\ \text{E.g.:} \quad Q(27+49) &= log(7+9+16+10+6)+log(2+4+6+1+7) = 2.98 \end{aligned}$$

The addition problems presented in this study were ranging from $Q = 0.6$ to $Q = 7.2$. Math tasks with $Q = 0.6$ indicate easy addition problems with one digit numbers, whereas tasks rated with $Q = 7.2$ are more complex math problems where four digit numbers have to be added.

10.2 Data pre-processing and analysis

By using only the calculation phase for analyzing the EEG data, it was ensured, they did not contain any motor artifacts or perceptual confounds which could be picked up by a classifier. For each task and subject, EEG data that contained too much noise were removed for further analysis after visual inspection. To analyze the data, the power spectrum during the calculation phase of each trial was estimated using the ARM based on Burg's maximum entropy method [139] with a model order of 32, estimated by means of ARMASA [152]. As frequency bands were not consistent and varied between subjects, a wide frequency range of $4\,\text{Hz} - 30\,\text{Hz}$ [51] in $1\,\text{Hz}$ bins was used.
Before using a regression to predict the task difficulty, the squared correlation coefficients r^2 between the power at each frequency bin (for each electrode) and the information content Q of the corresponding trials were calculated. This analysis can be used to estimate which electrodes and frequencies are mainly affected by task difficulty and which EEG features are most important to predict the cognitive demands of each subject.

10.3 Predicting the amount of workload

To predict the amount of workload, the previously calculated power spectrum was used as feature. Furthermore, the data was *z-score* normalized along the channels, to correct for inter-subject variability in the subjects baseline EEG power. Meaning for each trial the mean of each frequency bin equals zero. Based on the normalized data of nine subjects,

a linear ridge regression model with a fixed regularization parameter of $\lambda = 0.001$ was trained. For further analysis the number of electrodes were reduced to 17 inner electrodes (FPz, AFz, F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz, POz), to lower the influence of possible artifacts, which are most prominent on the outer channels. The goal of the present study was to develop an efficient and generalized prediction method which is able to differentiate and predict levels of workload. Therefore, a within-subject regression as well as a cross-subject regression were comparatively applied.

10.3.1 Within-subject regression

In a within-subject regression, training and testing of a regression model were performed on EEG data recorded from the same task and subject. This method was applied for each participant individually. Good differentiability and prediction of workload levels within each subject are required to apply cross-subject regression. For within-subject regression, a 10-fold cross validation was accomplished to verify the separability of the independent dataset, in order to show that it is possible to indicate workload differences in the EEG data.

10.3.2 Cross-subject regression

As mentioned in previous studies, the collection of training data for classification and prediction methods in combination with a real-world learning environment is challenging. Therefore, a method called cross-subject regression was applied. For cross-subject regression, a leave-one-subject-out validation was conducted to verify the separability of the independent datasets, in order to predict different workload levels in the EEG data. Therefore, EEG data recorded from $n-1$ subjects are used for calibrating a regression model. Subsequently, the workload prediction is evaluated based on the remaining EEG data from subject n. Again, the ultimate goal of this approach is to overcome the challenge of collecting training data for classifier training in combination with learning environments. In this study, the regression model was trained on data of nine subjects to predict the difficulty level on a single-trial basis for the one remaining subject, resulting in a predicted Q-value for each trial. This process was repeated ten times, so the data of each subject was used once for testing. For the cross-subject regression, two different models of training were tested. Either the regression model was trained on all trials, or only trials with a $Q < 6$ were used for the regression model training.

10.3.3 Evaluating the prediction performance

Since single-trial prediction is not necessary in a learning environment, the regression output was additionally smoothed to get a more robust prediction at the expense of increased delay of the system. To smooth the data, the moving average with a window length of six trials was calculated, which still guaranteed a response time < 1 min of the system. This delay is feasible for the detection of workload since it is not recommendable to adapt an

online learning environment every 8.5 sec (i.e., single trial duration). As criterion for performance evaluation of the presented prediction method, the *CC* to observe the statistical relationship between the actual and the predicted Q-values, as well as *RMSE* to examine the difference between the actual and the predicted Q-values were used.

10.4 Behavioral results

To ensure that the participants solved the tasks conscientiously, as well as to have an additional parameter for task difficulty, the error rate of each subject was logged. An analysis of the performance data shows that the Q-value is a suitable measure for task difficulty since it correlates well ($r = 0.945$, $p < 0.0001$) with the percentage of correctly solved trials (see Figure 10.2).

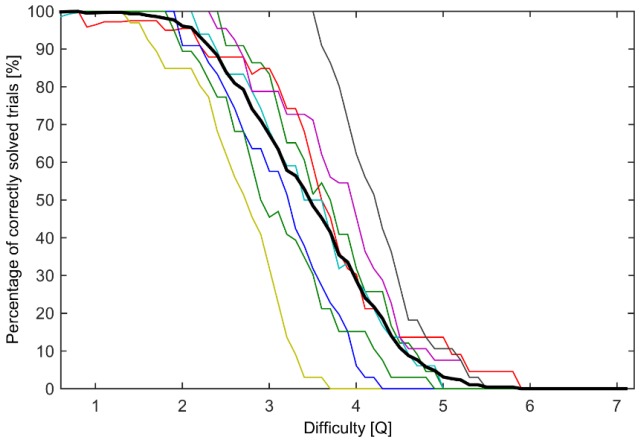


Figure 10.2: Percentage of correctly solved trials depending on the degrees of difficulty of each trial measured by Q. Each subject is shown by the thin colored lines and the average over all subjects is shown by the bold black line.

Although, there are inter-individual differences, trials with $Q < 2$ were solved correctly in nearly all cases, while trials with the difficulty of $Q = 4$ were solved correctly in about 30 % of the trials. None of the subjects were able to solve trials with a $Q \geq 6$. On average, all subjects solved 64.25 % of all 240 assignments correct. The best subject solved 72.08 % correctly, whereas the worst subject reached an accuracy of 56.25 % (see Table 10.1). Relative to the presented difficulty levels, of which 25 % were unlikely to be solved, this are realistic performance results. This leads to the assumption, that subjects participated conscientiously and tried to solve the tasks correctly.

Table 10.1: Percentage of correctly solved trials averaged over all trials for each individual subject, as well as the mean performance averaged over all subjects.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	mean
Correct [%]	66.25	66.25	65.42	64.17	67.92	56.25	72.08	61.25	60.00	62.92	64.25

10.5 Neurophysiological features

Analyzing the EEG data shows that task difficulty and thereby the individual's workload capacity is reflected in the power spectrum (see Figure 10.3). In the delta frequency band, a strong difficulty related effect over the occipital electrodes can be recognized, while the effect in the theta-band smaller. These measured effects might be caused due to artifacts. In the alpha-frequency, a difficulty related effect can be seen over the frontal electrodes, which might be caused by eye-movements. Diverse patterns are shown in the lower and higher beta-frequency band, which might be caused due to muscle artifacts.

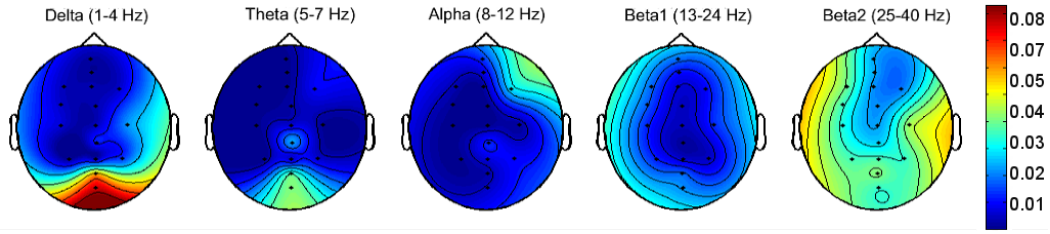


Figure 10.3: Topographic display of r^2-values averaged over all ten subjects, showing the influence of the Q-values for each electrode in different frequency bands, before EOG artifact correction.

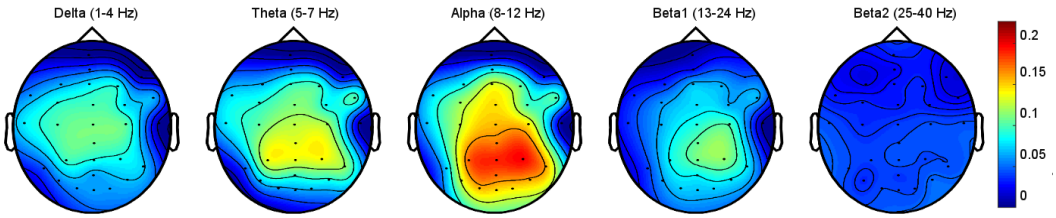


Figure 10.4: Topographic display of r^2-values averaged over all subjects, showing the influence of the Q-values for each electrode in different frequency bands, after EOG artifact correction.

Since the workload prediction should not just work because of eye-movements or muscle artifacts during various levels of difficulty, but because of changes in brain signals, further pre-processing steps were additionally conducted. First, an EOG-based regression described by Schlögl et al. [150] was applied, to remove the influence of eye movements. After calculating the power spectrum, the *z-score* normalization was replaced by a baseline correction to correct for inter-subject variability in the subjects baseline EEG power. Therefore, the first 30 trials (easy trials, with $Q < 1$) were used for normalization. The mean standard deviation for each frequency bin at each electrode was calculated over the first 30 trials and the remaining 210 trials were scaled according to these means and standard deviations. The 30 trials used for normalization were not used further in the prediction process, neither for training nor testing the model.
Analyzing the cleaned, pre-processed EEG data, it can be seen that task difficulty and thereby the individual's workload capacity is still reflected in the power spectrum (see Figure 10.4). Compared to Figure 10.3, the r^2 patterns of the EOG artifact corrected data

agree with postulated features in the literature. In the delta frequency band, a small difficulty related effect over the central electrodes is detectable, while the effect in the theta-, alpha- and beta-frequency band is more prominent over the parieto-occipital electrodes. This effect is especially strong for the alpha-frequency band. While a low beta-band still shows some effect related to task difficulty and thus to workload, they cannot be observed in the upper beta-band.

10.6 Prediction results

The performance results for the within-subject workload prediction are reported in section 10.6.1, Table 10.2 and the cross-subject prediction results are shown in section 10.6.2, Table 10.3. Furthermore, cross-subject workload prediction based on EOG corrected data and additional models were analyzed (see section 10.6.2.2, Table 10.4 and Table 10.5).

10.6.1 Within-subject

The within-subject workload prediction served as performance validation because good differentiability and prediction of workload levels within each subject is required to apply cross-subject regression. The within-subject workload prediction reached on average a correlation coefficient of $CC = 0.66$ and a $RMSE = 1.70$. Using the smoothed regression output (see section 10.3.3) for within-subject workload prediction leads to more robust and even better results with an average over all subjects of $CC = 0.88$ and $RMSE = 1.02$ (see Table 10.2). The prediction performance on individual basis reached for the best subject a CC with 0.94 and a $RMSE$ of 0.68, while for the worst subject a $CC = 0.69$ and a $RMSE = 1.39$ was achieved. The results demonstrate, that the amount of workload can be predicted using a within-subject regression method.

10.6.2 Cross-subject prediction results

Since the workload prediction was successful by using within-subject regression, it is possible to indicate workload differences in the EEG data. For an adaptive learning environment, a more efficient prediction method is preferable. Therefore, the cross-subject regression was analyzed in more detail.

Table 10.2: Performance results of the within-subject workload prediction using 10-fold cross-validation. The CC and the $RMSE$ between the actual and predicted Q are used for performance prediction. For both cases the trial-based, as well as smoothed results are reported.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	mean
CC	0.43	0.69	0.68	0.58	0.84	0.58	0.60	0.72	0.65	0.78	0.66
CC (smooth)	0.69	0.93	0.93	0.89	0.94	0.87	0.76	0.91	0.93	0.93	0.88
$RMSE$	2.20	1.61	1.71	2.09	1.11	1.95	1.63	1.60	1.82	1.30	1.70
$RMSE$ (smooth)	1.39	0.86	0.99	1.21	0.68	1.15	1.04	1.04	1.08	0.73	1.02

Table 10.3: Performance results of the cross-subject workload prediction using a leave-one-subject-out validation. The *CC* and *RMSE* between the actual and predicted Q are used for performance prediction. For both cases, the trial-based, as well as smoothed results are reported.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	mean
CC	0.67	0.50	0.65	0.63	0.63	0.59	0.39	0.62	0.52	0.57	0.58
CC (smooth)	0.89	0.91	0.88	0.91	0.89	0.86	0.74	0.84	0.71	0.79	0.84
RMSE	1.68	2.56	1.69	1.73	1.72	1.77	2.65	1.87	2.04	1.81	1.95
RMSE (smooth)	1.13	2.00	1.25	1.39	1.60	1.44	2.11	1.58	1.57	1.68	1.58

10.6.2.1 Data without EOG correction

First, the linear ridge regression was performed on non-EOG corrected EEG data. As stated in Table 10.3, the average result over all ten subjects for cross-subject workload prediction slightly decreased, compared to the within-subject prediction (see Table 10.2). Applying linear ridge regression to unsmoothed data lead on average to a $CC = 0.58$ and a $RMSE = 1.95$. Using the smoothed data (see section 10.3.3) for cross-subject workload prediction improved the prediction results to $CC = 0.84$ and $RMSE = 1.58$. For the best subject, a $CC = 0.89$ and a $RMSE = 1.13$ was achieved, whereas the worst subject merely reached a *CC* of 0.74 with a *RMSE* of 2.11. Since smoothing the data increases the prediction accuracies and is applicable in realistic learning environments, only smoothed data results will be reported in the following sections. The distribution of actual and predicted workload for cross-subjects are shown in Figure 10.5. The increasing workload is successfully predicted by using the linear regression method. It can be noticed that during the first 90 trials ($Q < 0.9$) and the last 30 trials ($Q > 6$) the actual and predicted workload deviate strongest from each other. In the first trials, it may be caused of an unfamiliar surrounding and solving new tasks, which might induce a higher amount of cognitive workload. For trials with $Q > 6$, the imprecise prediction may be due to subjects being overburdened while solving these tasks.

10.6.2.2 EOG corrected data and additional regression models

To ensure workload prediction is not only successful because of EOG artifacts, additional analysis were accomplished. After filtering EOG artifacts, baseline correction and smoothing the EEG data, again a linear ridge regression was applied to the pre-processed EEG data. The average result over all ten subjects was a correlation coefficient between the actual and the predicted Q of $CC = 0.75$ and a root mean square error of $RMSE = 1.73$, see Table 10.4. The best subject reached a $CC = 0.93$ and a $RMSE = 1.43$, while the worst subject achieved a $CC = 0.38$ and a $RMSE = 2.18$. Compared to the cross-subject prediction based on non-EOG corrected EEG data, the average workload prediction over all subjects slightly decreased, but is still stable and possible. The result of the difficulty prediction for each of the ten subjects can be seen in Figure 10.6. As in Figure 10.5, the increase in Q is generally well predicted by the regression model for each $Q < 4$, but the workload prediction is not that precise for $Q \geq 4$. For these trials the model reaches a plateau or even predicts a decline in task difficulty for about half of the subjects.

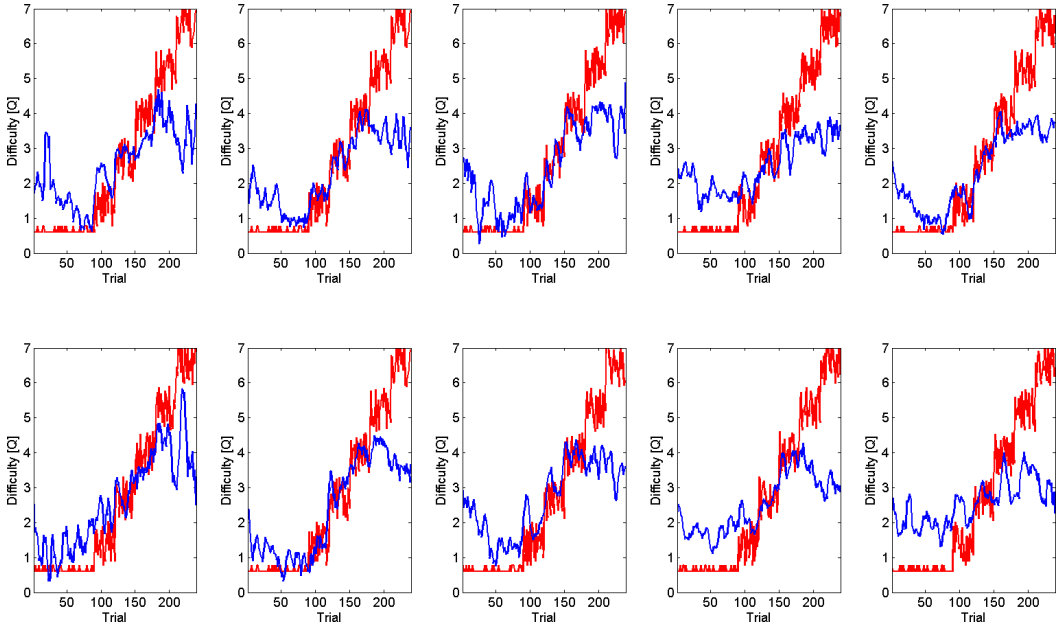


Figure 10.5: The blue line represents the difficulty prediction using a linear regression on features from 17 EEG channels (FPz, AFz, F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz, POz), while the red line indicates the actual degree of difficulty (Q-value) for all subjects without EOG correction. Subjects are enumerated from top left to bottom right line by line.

Table 10.4: Performance results of the cross-subject workload prediction with EOG corrected data using a leave-one-subject-out validation. Regression was trained on data of nine subjects and tested with data from the remaining subject. The *CC* and the root mean square error *RMSE* between the actual and predicted Q are used for performance prediction.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	mean
CC (smooth)	0.69	0.93	0.87	0.72	0.89	0.84	0.64	0.81	0.69	0.38	0.75
RMSE (smooth)	1.75	1.43	1.70	1.98	1.52	1.57	1.84	1.53	1.77	2.18	1.73

Due to this finding, two different regression models were tested: one trained on smoothed data from all trials and one trained on smoothed data from trials with $Q < 6$. While the detailed results for the model trained on data from all trials were already shown in Table 10.4, a comparison of the average prediction performance metrics for both models are shown in Table 10.5.

The results regarding *CC* and *RMSE* do not differ significantly, when using all trials for training, compared to using only $Q < 6$ trials for training ($CC = 0.76$ and $RMSE = 1.78$). When reducing the test data to trials with $Q < 6$ as well, $CC = 0.82$ and $RMSE = 1.34$ are significantly better ($p < 0.01$, t-test).

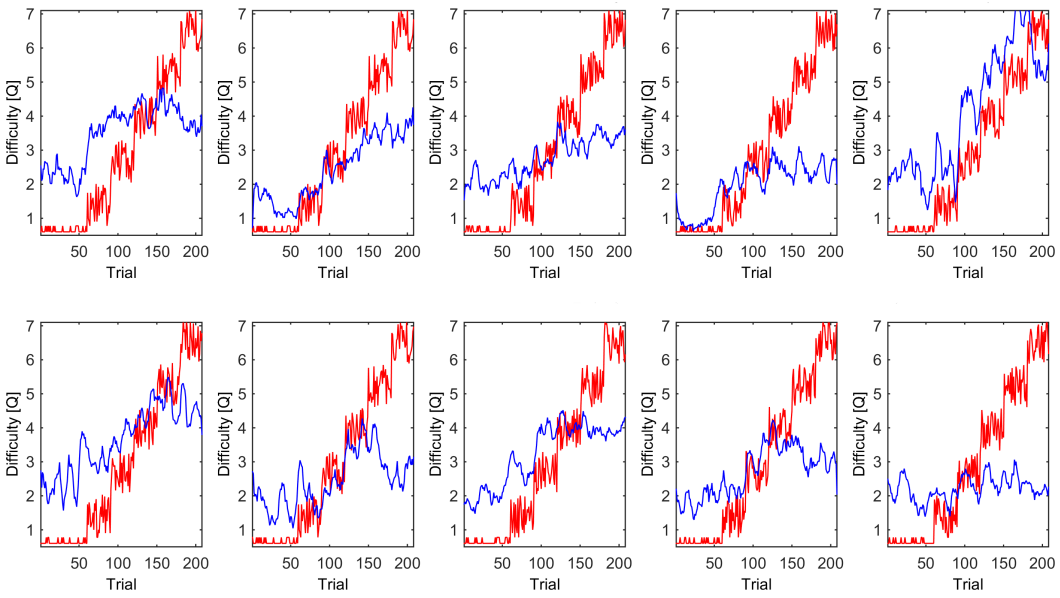


Figure 10.6: The blue line represents the difficulty prediction using a linear regression on features from 17 EEG channels (FPz, AFz, F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz, POz), while the red line indicates the actual degree of difficulty (Q-value) for all subjects after EOG artifact rejection. Subjects are enumerated from top left to bottom right line by line.

Table 10.5: Prediction performance with various training- and testing data. The three combinations of using either all, or only trials with $Q < 6$ for training and testing averaged over all subjects are stated here. For performance prediction, the *CC* and the *RMSE* were used.

Train	Test	*CC*	*RMSE*
All	All	0.75	1.73
$Q < 6$	All	0.76	1.78
$Q < 6$	$Q < 6$	0.82	1.34

10.6.3 Visual feedback in real-time

In the previous sections, it was demonstrated that the cross-subject regression method can precisely predict the amount of workload. To evaluate the online applicability of the proposed method, the regression model described in section 10.6.2.2 was used in an online simulation. Dependent on the predicted amount of workload, visual feedback for the subjects was generated. If the predicted amount of workload was smaller than $Q = 3.5$ a green circle was presented, otherwise a red circle appeared. This online simulation worked well and without timing problems.

10.7 Discussion

The results show that the amount of workload can be predicted using a cross-subject regression method. Compared to earlier studies, where cross-task classification (see chapter 7, 8 and 9) led to classification accuracies merely around chance level, cross-subject prediction seems to be more robust and recommendable.

10.7.1 Non-linear effects in the EEG data

The workload prediction led to good results for $Q < 4$, but for trials with a $Q \geq 4$, the performance drops. The workload prediction results are more precise when using only trials with $Q < 6$ for classifier training and testing. This leads to the assumption, that there are non-linear effects in the EEG data. These effects are likely induced by disengagement due to tasks being too difficult or boring. As Chanel et al. [154] postulated, further increasing workload leads to disengagement, which causes reverse EEG patterns and thus neural signatures being similar to EEG characteristic for low cognitive workload.

10.7.2 Necessity of EOG correction

The cross-subject workload prediction based on non-EOG corrected EEG data, lead to better results in terms of *CC* and *RMSE* compared to prediction results based on EOG and baseline corrected EEG data. In both cases, the increasing workload was successfully predicted by the linear regression model, but in the analysis without EOG correction, a difficulty related increase of spectral power in the frontal electrodes could be detected. These findings could be solely explained by artifacts introduced by eye-movements. Nevertheless, it could be demonstrated, that a workload prediction after EOG artifact correction is still possible. Thus, it is unlikely, that the regression model can predict the workload amount just based on artifacts. To conclude, eye-movements can be a confounding factor in EEG-based adaptation of numerical learning, but it is not the only component, workload prediction is based on.

10.7.3 Non-stationarities in the EEG signals

As the math problems were presented in a fixed order at advancing levels of difficulty, EEG signals might change due to non-stationarity over time, e.g., caused by fatigue. Actually, non-stationarity within a session should not have a great influence on the EEG data while using cross-subject methods, as non-stationarities between subjects are assumed to be larger than within a session. Since only simple methods for normalization and thereby alleviation of non-stationarities were used, more advanced methods for reduction of non-stationarities [163] might be used in future work.

10.7.4 Comparison with the state of the art

Compared to previous studies using the cross-subject classification method [50, 112, 113], a higher prediction accuracy was reached. Applying a SVM for classifying the EEG data of this study, a classifier performance of 90 % was achieved on average. Wang et al. [112] used a Bayes classifier and reached a classification accuracy of 84 % on average, whereas Gevins et al. [50] utilized ANNs and achieved an average classification accuracy of 83 %. Furthermore, the used electrodes (frontal, central, parietal, occipital) and frequency bands (4 Hz – 30 Hz) which served as features for the cross-subject regression and classification in this study were more workload specific than in the studies from Wang et al. [112] and Jin et al. [113]. Wang and colleagues [112] positioned 17 electrodes above the whole cortex and frequencies ranging from 2 Hz – 57 Hz served as features. Jin et al. [113] used features of 12 electrodes placed at the frontal, central and parietal brain areas in the time domain with 35 points in time. Thus, no previous study was able to predict workload levels across subjects as precise as it was shown in this study.

10.8 Conclusion

To conclude, the cross-subject workload prediction worked successfully. Furthermore, it was possible to generate visual feedback for the subjects in an online simulation depending on the predicted workload. Finally, it has to be noted, that no previous study was able to predict workload levels by using a cross-subject regression during solving addition tasks, as precise as it was shown in this study.
In the next chapter, an online learning environment will be introduced, to predict the user's workload and adapt the presented exercises accordingly, to successfully support students in their learning process. This learning environment is developed based on the results stated in this chapter.

11 Online workload detection in an adaptive learning environment

As it has been demonstrated in the previous chapter 10, it is indeed possible to predict the amount of workload across subjects by EEG data. In this chapter, it will be shown how EEG data can be utilized to adapt a learning environment in real-time. The developed cross-subject regression method (see section 10.3.2) is applied in an online adaptive learning system, which presents the learning material in a format that keeps the learner engaged and motivated without overwhelming his or her limited workload resources. To evaluate the learning effect of the EEG-based adaptive learning environment, the performance is compared with a control group using an error-based adaptive learning environment, which is state of the art.

11.1 Study design

In the following, the participants data of the experimental group, as well as of the control group, will be introduced and the methods for EEG recording will be described.

11.1.1 Participants

To be able to evaluate and interpret the developed EEG-based adaptive learning environment, a second learning environment was used by a control group, which adapts based on the correctness of a given answer. The participants of both groups were students of various disciplines.

Experimental group

The EEG-adaptive learning environment was used by 13 subjects, 7 men and 6 women aged between 21 and 35 (median age 30). They voluntarily participated in the EEG experiment. None of them had prior knowledge in calculating using base 8, which was verified by a pre-test.

Control group

The control group consists of 12 subjects, 8 male and 4 female aged between 22 and 27 (median age: 23). They voluntarily participated in the EEG experiment, using an error-adaptive learning environment, where the EEG data was just measured but not used as additional information for adaptation. All of them had no prior knowledge in calculating using the octal number system, which was verified by a pre-test.

11.1.2 EEG recordings

A set of 29 active electrodes (actiCap, BrainProducts GmbH) was used for recording the brain signals. They were attached to the scalp, placed according to the extended International Electrode 10 - 20 Placement System.

Experimental group

As the online workload prediction for the experimental group was based on data from the previous study described in chapter 10, the electrode positions were kept the same: FPz, AFz, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, CPz, P7, P3, Pz, P4, P8, PO7, POz, PO8, O1, Oz, O2. The number of electrodes used for further analysis and online adaptation, were reduced to 16 inner electrodes (FPz, AFz, F3, Fz, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz, POz), to lower the influence of possible artifacts which are most prominent in the outer channels.

Control group

For the control group, standard positions indicated at the actiCap from Brain Products GmbH were used: Fp1, Fp2, F7, F3, F4, F8, FC3, FCz, FC4, T7, C3, Cz, C4, T8, TP9, CP5, CP1, CP2, CP6, TP10, P7, P3, Pz, P4, P8, PO9, O1, O2, PO10. Since the EEG signal was solely utilized for offline workload estimation, not for online adaptation, these differences are not crucial. For the offline EEG analysis stated in the following sections merely nine electrodes (FC3, FCz, FC4, C3, Cz, C4, P3, Pz, P4), the same as in the experimental group, were considered for the purposes of comparability.

Three additional EOG electrodes were used to record the eye-movements, two placed horizontally at the outer canthus of the left and right eye to measure horizontal eye-movements and one placed in the middle of the forehead between the eyes to measure vertical eye-movements. The reference electrode was placed on the left and the ground electrode at the right mastoid. EOG and EEG signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp, g.tec). The sampling rate was 512 Hz and the impedance of each electrode was less than 5 kΩ. EEG data was high-pass filtered at 0.1 Hz and low-pass filtered at 100 Hz for the experimental group. For the recordings during the error-based learning environment a low-pass filter at 60 Hz was applied. The interesting frequencies for workload analysis are smaller than 30 Hz, thus the two different low-pass filters did not have any impact on the following analysis. Furthermore a notch-filter was utilized between $48\,\text{Hz} - 52\,\text{Hz}$ to filter power line noise.

11.1.3 Task design

An introduction to the presented learning material, as well as a detailed description of the general paradigm will be given in the following section.

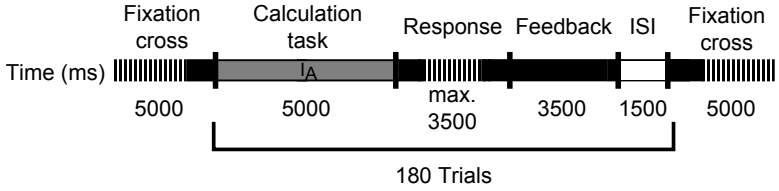


Figure 11.1: Schematic flow of the learning phase for an EEG-based, as well as an error-based adaptive learning environment. The grey line indicates the activation interval (I_A), whereas the black dashed line represents the response interval. Subsequently a feedback phase occures indicated by the black line, followed by an inter-stimulus interval (ISI) shown as white line.

11.1.3.1 Octal number system as learning content

As in the previous chapter 10, subjects had to solve 180 addition tasks with diverse levels of difficulty, while EEG was measured. Different from the previous study, subjects did not calculate in the decimal number system, but in the octal number system so they had to learn a new task.
There were three reasons for choosing the octal number system. First, a learning process should be induced while solving the tasks. The participants in the experimental group, as well as in the control group, were able to calculate addition tasks using base 10, but solving assignments using base 8 was mostly new for them, this was ensured by a pre-test. Second, the information content Q according to Thomas [162] (see section 10.1.2) turned out to be a good and objective difficulty measure for math tasks, corresponding to the amount of workload (see section 10.6). By using addition tasks in the octal number system, the utilization of this objective measure Q for defining task difficulties was possible again. Thus, the presented math problems varied in difficulty as measured by the Q-value and ranges from 0.6 (easy) to 6.6 (difficult). Third, with the addition tasks from section 10.1.2, clear EEG correlates corresponding to the amount of workload were detected. Using the developed cross-subject regression method led to successful workload prediction results.
To enable a comparison of the experimental group with the control group, the task design was as similar as possible and without additional confounds.
By using addition tasks in the octal number system, the same task design, with the same presentation scheme and the same objective difficulty measurement Q as in section 10.1.2 were used, by additionally inducing a learning effect to the participants. Furthermore, perceptual confounds in the EEG data due to complex learning materials were avoided.

11.1.3.2 Paradigm

Based on the paradigm introduced in section 10.1.2, addition tasks were supposed to be solved. Each trial consisted of four phases (see Fig 11.1). First, the calculation phase occurred, where the problem to be solved was shown for 5 sec. Subsequently, subjects had 3.5 sec to type in their result, followed by a feedback phase. Because subjects had to learn calculating using base 8 they got feedback by presenting the correct answer for 3.5 sec. Each trial ended with an inter-trial interval (ISI) of 1.5 sec. To avoid the classifier

of being based on perceptual-motor confounds, only the calculation phase was used for further EEG analysis. For each participant, the learning session started with an addition task of difficulty level $Q = 2$ and changed dependent on the adaptation method of the utilized learning environment, either adaptive on EEG data or on correctness of the given solution. A pre-test at the beginning and a post-test at the end of each experiment were accomplished with eleven assignments of linear increasing degrees of difficulty. These tests were utilized, to enable a direct comparison of subjects' knowledge before and after solving the learning phase, of how to calculate using the octal number system.

11.2 EEG data pre-processing and analysis

The cross-subject regression for workload prediction, introduced in section 10.3.2, is reapplied in the current study. Therefore, the pre-processing, as well as the analysis for the newly recorded EEG data were kept similar to the previous study described in 10.2. For analyzing the EEG data, only the 5 sec time frame of the calculation phase was used, to ensure that it did not contain any motor artifacts or perceptual confounds, which could be picked up by a classifier. The power spectrum was calculated for this time window for each trial by using ARMs with a model order of 32, based on Burg's maximum entropy method [139]. As frequency bands were not consistent and varied between subjects, a wide frequency range of $4\,\text{Hz} - 30\,\text{Hz}$ in 1 Hz bins was used [51]. To predict the amount of workload, the calculated power spectra were used as features. To correct for inter-subject variability in the subjects baseline EEG power, the data was *z-score* normalized along the channels. Hence, for each trial, the mean of each frequency bin equals zero. In the EEG-adaptive learning environment these features were used to adapt the difficulty level of the presented learning material online. For the control group, the EEG data was merely recorded to have the same conditions during the learning phase and to be able to analyze the neurological processes, as well as the workload prediction, subsequently offline.

11.3 Cross-subject regression for online workload prediction

Efficient classification and prediction methods are necessary for utilizing EEG-based adaptive learning environments in real-world settings. To be able to completely avoid the training phase of regression models, the cross-subject regression method from section 10.3.2 was applied.
Therefore, normalized EEG data from 10 subjects recorded in the previous mentioned study (see section 10.1) were used for calibrating a linear ridge regression model with a fixed regularization parameter of $\lambda = 0.001$, to estimate the difficulty in real-time on a single-trial basis for new participants. The number of electrodes used for online adaptation was reduced to 16 inner electrodes (FPz, AFz, F3, Fz, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz, POz), to be consistent with the electrode positions used in the previous cross-subject study (see section 10.1). Furthermore, only trials with a Q smaller than 6 were used to train the regression model, since there can be non-linear effects in the EEG data, induced by disengagement due to tasks being too difficult (see section 10.6.2.2). Afterwards, the

previously trained regression model was applied in the EEG-adaptive learning environment, to predict the amount of workload for independent, new subjects in real-time. The result was an EEG-based learning environment, which could adapt the difficulty level of each task to the necessity of each individual subject online. Furthermore, the participants did not need to go through an additional calibration phase for the regression model.

11.4 Adaptation methods for learning environments

In this study two adaptation methods were comparatively applied. The new and innovative EEG-based adaptation of learning material was compared with the state of the art error-based adaptation, to evaluate the efficiency and performance.

11.4.1 EEG-based adaptation used by the experimental group

For the experimental group, the EEG data served as workload indicator. Depending on the estimated amount of workload a participant required for solving each task, the learning material got more difficult or easier. Each session started with an exercise of difficulty level $Q = 2$. A linear ridge regression was used for predicting the amount of workload. If the predicted workload was less than $Q = 0.8$, the presented task difficulty was assumed to be too easy. Thus the following Q-value ($= \textit{target } Q$) was increased by 0.2. Vice versa, the *target* Q of the subsequent task decreased by 0.2 when the predicted workload was greater than $Q = 3.5$. In this case, the presented task difficulty was assumed to be too difficult. If the predicted workload was between $Q = 0.8$ and $Q = 3.5$, the Q-value for the next presented task remained the same and the difficulty level was kept constant.
These thresholds were defined based on the results in section 10.4. Trials with $Q < 1$ were solved correctly in all cases, while none of the subjects were able to solve trials with a $Q \geq 6$. As can be seen in Figure 10.2, 50 % of the trials with a $Q = 3.5$ were successfully solved on average.
To allow a smooth adaptation, the presented tasks were selected from a variety of tasks with a Q-value in the range of $\textit{target } Q \pm 0.5$. For instance, given an arithmetic problem with $Q = 2$ and the predicted amount of workload is 0.7 the adaptation algorithm assumes the presented addition task was too easy. Therefore, the *target* Q for the following assignment will increase by 0.2 and the subsequently presented addition task will have a difficulty level with an information content Q ranging between 1.7 and 2.7. This is calculated as follows:

$$\textit{target } Q = Q + 0.2 = 2 + 0.2 \tag{11.1}$$

$$\text{information content range of } Q = [\textit{target } Q - 0.5,\ \textit{target } Q + 0.5] = [1.7, 2.7] \tag{11.2}$$

11.4.2 Error-based adaptation used by the control group

An error-based adaptive learning environment was developed for the control group. The number of wrong answers served as performance and adaptation measure. This is the current state of the art and permits an assessment of learning efficiency. As for the experimental group (section 11.4.1), each learning session started with an exercise of difficulty level $Q = 2$. When subjects solved five consecutive tasks correctly, the difficulty level increased by factor 1. Vice versa, the difficulty level decreased by factor 1 when participants made three errors in a row. Otherwise, the Q-value did not change and the difficulty level was held constant. The adaptation scheme was kept similar in the control group, as in common tutoring systems. Because of the repetitions till the difficulty level has been changed, the presented Q-values increased or decreased by steps of size 1 and the calculated Q-values were rounded to the next integer. For the control group, EEG signals were additionally recorded to keep the conditions between both studies constant and to enable an offline analysis, but they were not used for online adaptation.

11.5 Neurophysiological features

In previous studies described in section 6.3.1 and 10.5, it could be shown, that task difficulty and thereby the individual's workload capacity is reflected in the power spectrum. To estimate at which electrodes and frequencies the EEG is most affected by task difficulty, the squared correlation coefficient (r^2) between the power at each frequency bin (for each electrode) and the information content Q of the corresponding trials were calculated offline. The analysis of neurophysiological features is stated in the following sections for the experimental group, as well as for the control group. The EEG features described for the experimental group are used for the online adaptation. The neurophysiological features of the control group are merely used for evaluation and comparison reasons, offline.

11.5.1 Analysis of the recorded EEG data from the experimental group

In Figure 11.2 r^2-values for each frequency bin and channel are shown as heatmap for each subject of the experimental group. As postulated in literature, a strong difficulty related effect in the alpha-frequency over the parieto-occipital electrodes (channels 10-16) can be detected in 8 out of 13 subjects. The effect is also prominent in the delta- and theta-frequency band for 9 out of 13 subjects, whereas rather the frontal-central electrodes are affected. Diverse patterns are shown in the lower and higher beta-frequency band, which might be caused due to muscle artifacts. For 8 out of 13 subjects, channel 16 (i.e., electrode POz) has an independent pattern compared to adjacent electrodes. This leads to the suggestion that the electrode was broken and interfering signals were measured during this recording.
To ensure the broken electrode POz does not influence the workload prediction erroneously, the weights of each frequency bin at each electrode are plotted in Figure 11.3. The weights reflect the importance of a feature for the workload prediction. The higher the weight of a feature, the more it has an influence on the online workload prediction.

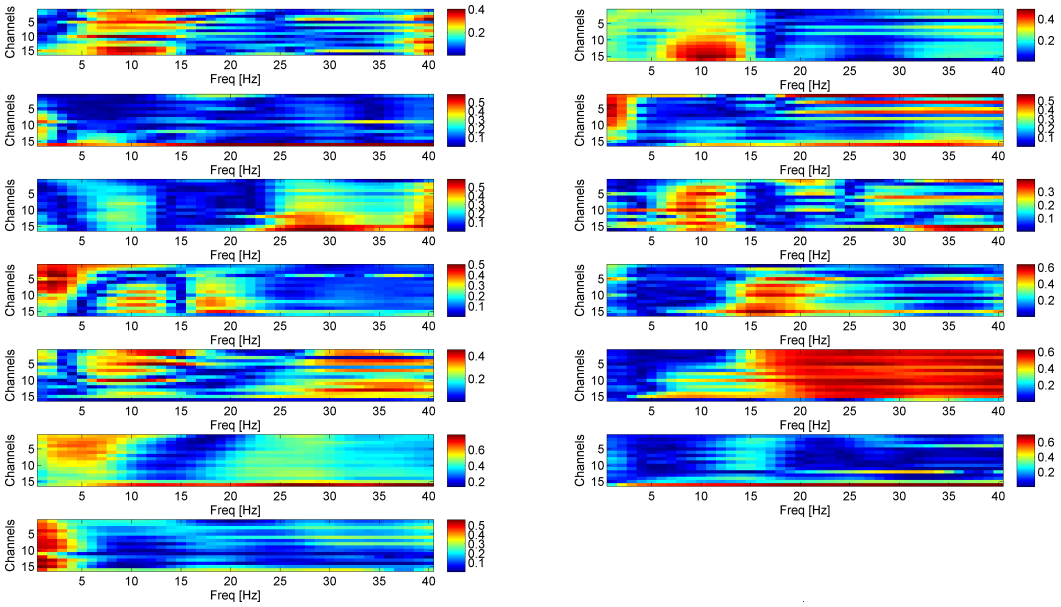


Figure 11.2: Heatmap for each subject of the experimental group, presenting r^2-values between the power at each frequency bin for each electrode and the information content Q. High squared correlations are indicated by red, while low squared correlations are indicated by blue. Participants shown here are enumerated from top left to bottom right line by line.

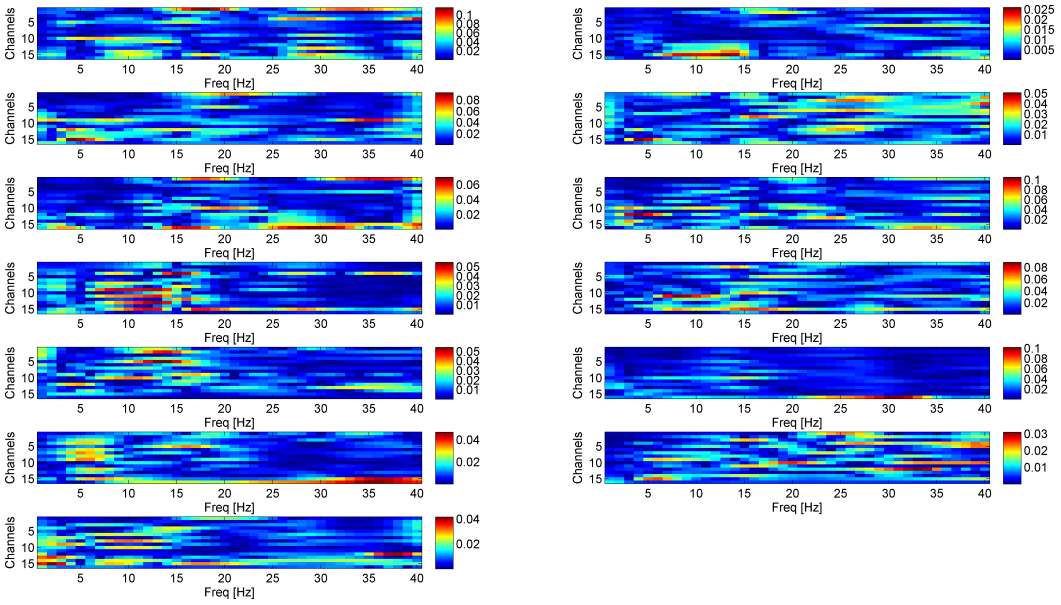


Figure 11.3: Heatmap for each subject of the experimental group, presenting the regression weight of each frequency bin at each electrode. Red indicates features which are strongly weighted and thus important for the regression model, whereas features plotted in blue have lower weights and are thus not important for the online workload prediction. Participants shown here are enumerated from top left to bottom right line by line.

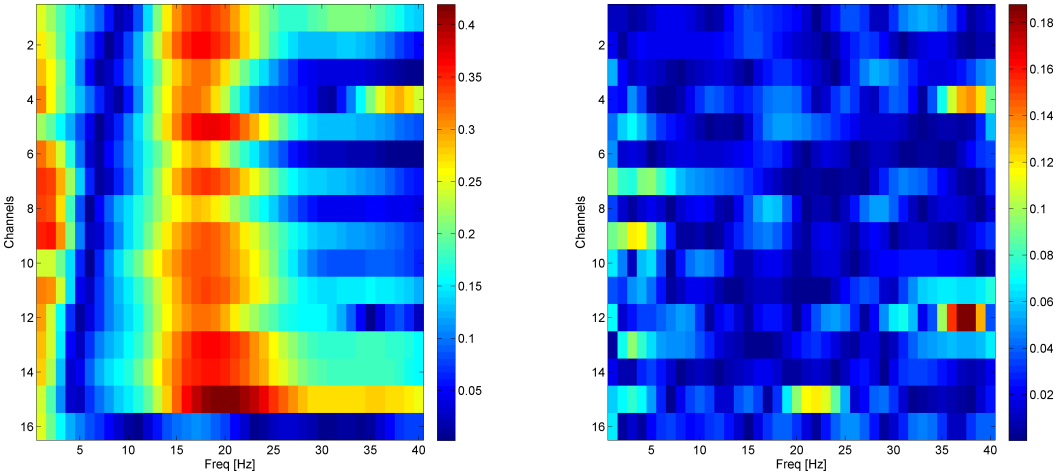


Figure 11.4: Heatmaps averaged over all subjects of the experimental group. **Left:** r^2-values between the power at each frequency bin for each electrode and the information content Q. High squared correlations are indicated by red, while low squared correlations are indicated by blue. **Right:** The regression weight of each frequency bin at each electrode. Red indicates features which are strongly weighted and thus important for the regression model, whereas features plotted in blue have lower weights and are thus not important for the online workload prediction.

For the major part of the subjects, the features in the theta- and alpha-frequency band are strongly weighted, regardless of channel 16. Merely in subject 4, 5, 6, 10 and 11, a strong weighting of features from channel 16 in the lower and upper beta-frequency bands can be detected, which might influence the online workload prediction negatively. Averaged over all subjects (see Figure 11.4) especially features from the delta- and theta-frequency bands across the central, parietal and occipital electrodes are strongly weighted, whereas only a few features from the beta-frequency band are used for the online workload prediction.
In Figure 11.2 the alpha- and theta-frequency bands show an effect related to task difficulty and thus to workload. These strong effects cannot be observed in the heatmap (see Figure 11.4 left) and topography plot (see Figure 11.5), where the average power spec-

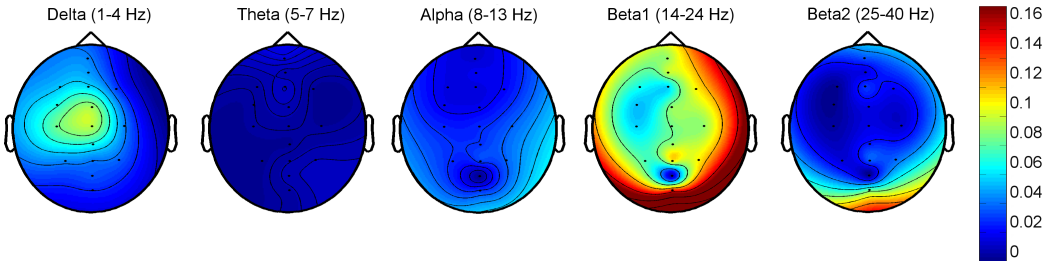


Figure 11.5: Topography plot averaged over the power spectra of all subjects of the experimental group, presenting r^2-values between the power at each frequency bin for each electrode and the information content Q. High squared correlations are indicated by red, while low squared correlations are indicated by blue.

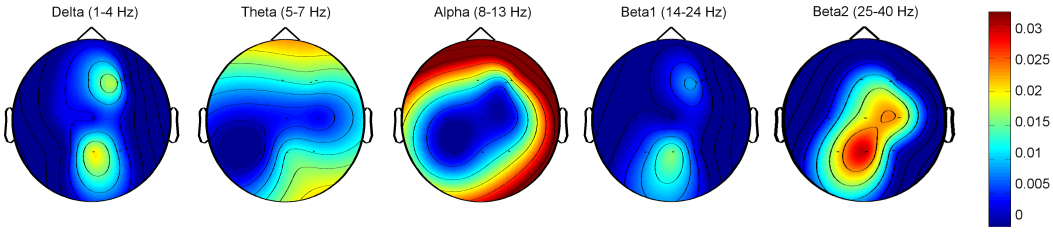


Figure 11.6: Topography plot averaged over the power spectra of all subjects in the control group, presenting r^2-values between the power at each frequency bin for each electrode and the information content Q. High squared correlation is indicated by red, whereas a low squared correlation is presented in blue.

trum over all subjects is shown. In Figure 11.5, a clear difficulty related effect over the central electrodes can be recognized ($r^2 = 0.1$) in the delta-frequency band, while there is no effect detectable in the theta-frequency band. A small effect ($r^2 = 0.04$) can be measured over the parieto-occitiptal electrodes in the alpha-frequency band. As in the heatmap (Figure 11.2), strong diverse patterns are shown in the lower and less prominent in the upper beta-frequency band ($r^2 = 0.16$). This might be caused due to muscle artifacts. The increased r^2-value in the theta- and alpha-frequencies, while averaging the power spectra over all subjects, can be explained by varying frequency band boundaries between subjects. The patterns in the delta-frequency band seem to be consistent over the cross-subject studies (see chapter 10) and therefore a robust objective measurement to predict the amount of workload during addition tasks.

11.5.2 Analysis of the recorded EEG data from the control group

The cognitive load theory points out, that in a learning phase, levels of workload should be held in an optimal range for successful learning. Therefore, the recorded EEG data of the control group was analyzed offline, to be able to estimate the amount of workload for each trial and subject subsequently.

The calculated r^2-values averaged over the power spectrum of all subjects of the control group are shown as topography plots in Figure 11.6. A strong difficulty related effect over the parieto-occipital and frontal-central electrodes can be recognized in the delta ($r^2 = 0.02$), as well as in the lower and upper beta-frequency band ($r^2 = 0.03$). The effects measured in the theta- and alpha-frequency band might be caused due to eye-movements and muscle artifacts.

11.6 Behavioral results

The following sections report on the behavioral results of the experimental and control group, while using the two diverse learning environments.

11.6.1 Performance results of the experimental group

On average, 45.51 % of all 180 assignments were solved correctly by the experimental group, see Table 11.1. The best subject answered 92.22 % of all 180 tasks correctly, whereas the least performing subject merely solved 10 % of the assignments successfully. Averaged over all subjects, a maximum Q-value of 5.85 was reached by adapting the learning material based on EEG data. Each subject achieved at least the difficulty level of $Q = 3.2$.
In Figure 11.7, the presented difficulty levels are shown with the number of correctly solved trials for each subject, including a smoothing over six trials for better illustration. An increase of task difficulty can be recognized for all subjects from the starting point till the end of the learning phase. The number of correctly solved trials was constantly high for subjects 2, 3, 9, 12 and 13. For subjects 1, 5, 6, 7 and 10 an alternating increase and decrease of correct answers was detectable, which might be caused by adapting the learning material too late. Further, the learning processes in subjects could have led to these findings, when the presented level of difficulty was not changing for a longer time. Subjects 4, 8 and 11 showed just a small number of correctly solved tasks after the first trials. For two of them (subject 4 and 11) the difficulty level linearly increased to the highest Q-value ($Q = 6.6$) without decreasing till the end of the learning phase. A wrong workload prediction based on the broken electrode POz could be the reason for this effect (see section 11.7.1.2).

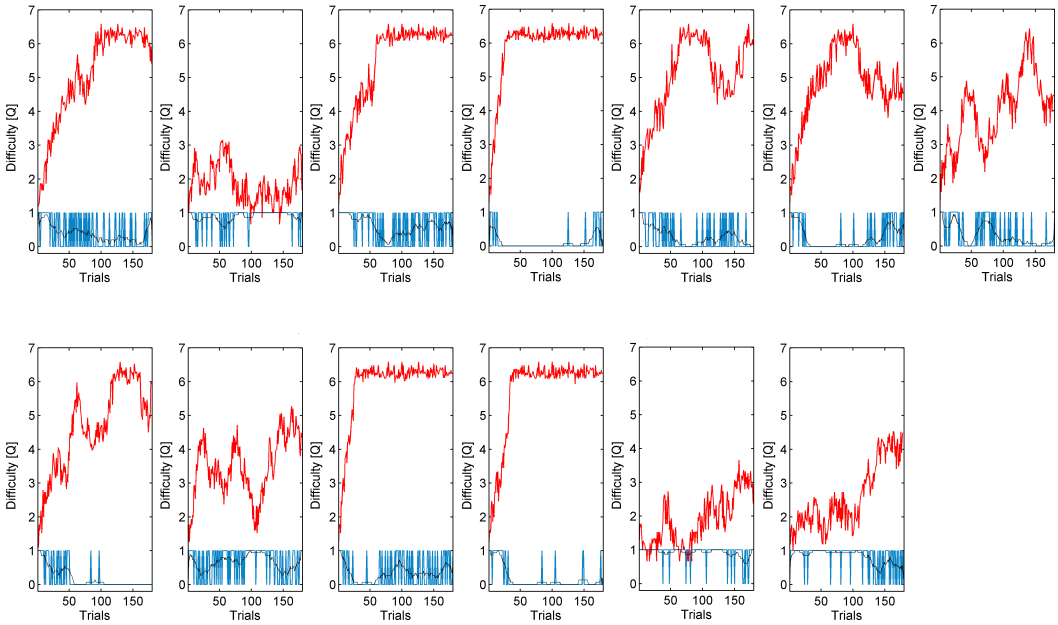


Figure 11.7: The actual difficulty level of each trial (red line), the correctness of each result for each trial (blue line), whereas 1 is correct and 0 is wrong, as well as the smoothed curve of correctness (black line) for each subject of the experimental group. Participants shown here are enumerated from top left to bottom right line by line.

Table 11.1: Relative amount of correctly solved trials in % and the maximum Q-value each subject of the experimental group reached, while performing the learning session. The maximal reachable Q-value is 6.6.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	mean
Correct	38.33	88.89	60.56	10.00	28.89	25.56	31.11	17.78	66.11	33.33	16.67	92.22	82.22	45.51
max. Q	6.6	3.2	6.6	6.6	6.6	6.6	6.6	6.6	5.2	6.6	6.6	3.6	4.6	5.85

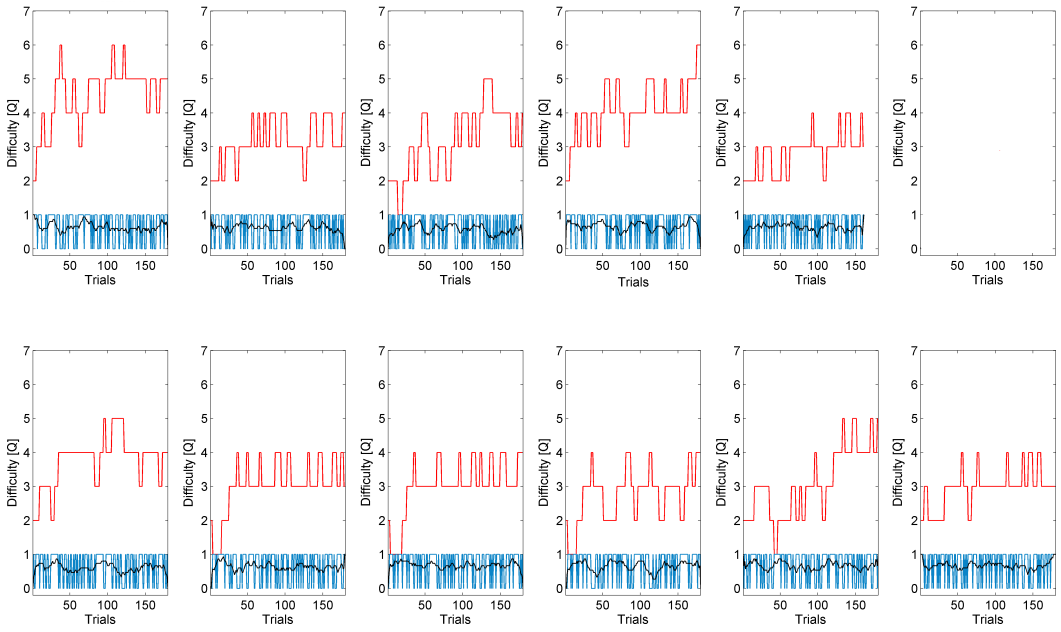


Figure 11.8: The actual difficulty level of each trial (red line), the correctness of each result for each trial (blue line), whereas 1 is correct and 0 is wrong, as well as the smoothed curve of correctness (black line) for each subject of the control group. Participants shown here are enumerated from top left to bottom right line by line. Subject 5 quit after 160 trials. Because of technical problems, no Q-values and given answers were saved for subject 6.

11.6.2 Performance results of the control group

The control group answered 64 % of all 180 assignments correctly on average. Since the error-rate was used for adapting the difficulty level of the presented learning material, the number of correctly solved trials was similar across subjects. On average, a maximum Q-value of 4.64 was reached (see Table 11.2). The best subjects reached a maximum Q-value of 6, whereas each participant achieved at least the difficulty level of $Q = 4$. While the difficulty level of the presented tasks increased compared to the starting point, the smoothed correctness curve was constantly high for each subject (Figure 11.8). It can therefore be assumed, that subjects participated conscientiously and tried to solve the tasks correctly. Because of technical problems, no Q-values and given answers were saved for subject 6. Subject 5 quit the experiment after 160 trials.

Table 11.2: Relative amount of correctly solved trials in % and the maximum Q-value each subject of the control group reached, while solving the learning session. The maximal reachable Q-value is 6. Because of technical problems no Q-values and given answers were saved for subject 6.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	mean
Correct	64.44	62.22	57.22	66.67	63.35	–	61.11	63.33	67.78	65.00	66.11	66.67	64.00
max. Q	6	4	5	6	4	–	5	4	4	4	5	4	4.64

11.7 Workload prediction results

The cognitive load theory recommends to keep the amount of workload in an optimal range while presenting and studying new learning material (in this study $0.8 < Q < 3.5$), to enable a successful learning support. Therefore, a workload prediction for each subject was calculated, using the calibrated linear ridge regression model as stated in section 11.3. Frequencies ranging form $4\,\text{Hz} - 30\,\text{Hz}$ were used as features. In the following section, the workload prediction results are stated for the experimental group, as well as for the control group. For the experimental group, the predicted workload results were calculated in real-time and used for the online adaptation of the presented learning material. For the control group, the predicted workload results were calculated offline for evaluation and comparison reasons, but not for adapting the presented learning material.

11.7.1 Workload prediction results of the experimental group

In Figure 11.9, the actual task difficulty and the predicted performance during the online experiment is shown for each subject of the experimental group. Furthermore, the workload prediction based on 16 channels is compared to the workload prediction based on 15 channels, to analyze how strongly the broken electrode POz influenced the prediction results in the online study. Therefore, an additional offline analysis was conducted, where the electrode POz was excluded in order to determine its impact on the workload prediction and thus the online adaptation of the learning material.

11.7.1.1 Online workload prediction with 16 channels

During the learning session, the amount of workload was held in the predefined optimal state over time for all 13 subjects. Thus, the predicted workload for each task ranged mainly from $0.8 < Q < 3.5$. But taking the observations of the neurophysiological features and the performance results into account (see section 11.5 and 11.6), especially for subject 4 and 11, it is debatable whether this workload prediction is valid, or due to technical artifacts based on the broken electrode POz. Therefore, an additional offline analysis was performed, where the electrode POz was excluded in order to determine its impact on the workload prediction and thus on the online adaptation of the learning material.

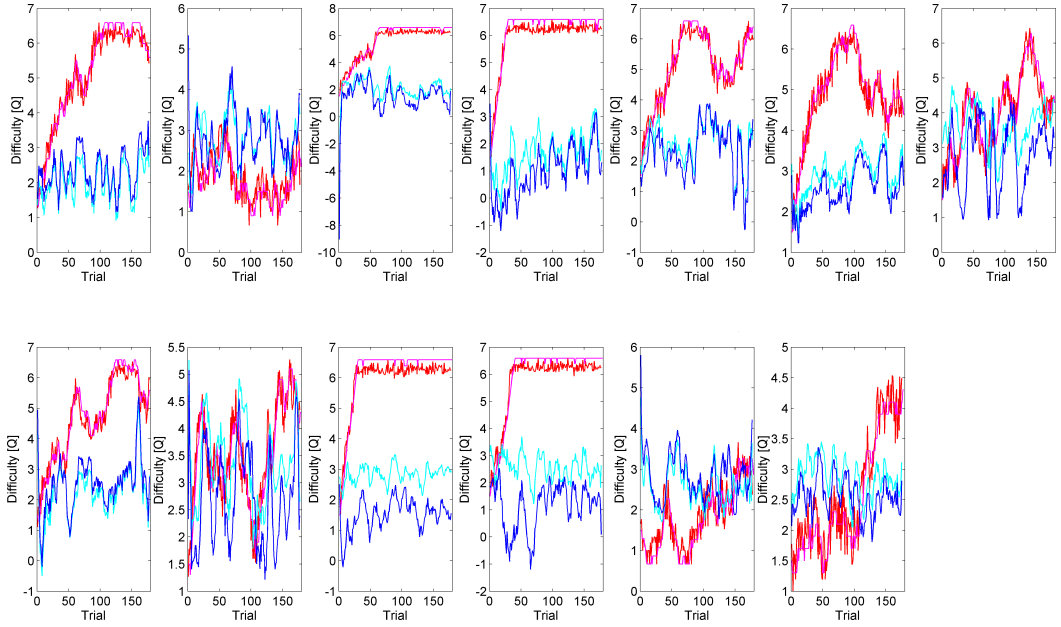


Figure 11.9: Workload prediction performance during the online EEG-adaptive study with 16 channels (blue line) and during an offline simulation with 15 channels, excluding electrode POz (light blue line). For better visualization, the predicted workload curves were smoothed. Furthermore, the presented task difficulty (red line) and the corresponding *target Q* (pink line) are shown. Participants are enumerated from top left to bottom right line by line.

11.7.1.2 Offline workload prediction with 15 channels

For the offline workload prediction, the regression model was the same as for the online experiment, but with 15 channels (FPz, AFz, F3, Fz, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz), omitting channel POz.

As the global deviation (*GD*) is a good metric for capturing prediction bias errors, it was calculated to analyze the difference of workload prediction based on 15 and 16 channels (see Table 11.3). The smaller the *GD*-value, the smaller the predicted bias error. For subjects 1, 3, 4, 5, 6, 8, 10 and 11 a $GD > 3.5$ is measurable. As subjects 4, 5, 6, 10 and 11 exhibit high regression weights for channel 16 (see Figure 11.3), it can be assumed, that the learning material adaptation during the online learning session was negatively influenced by the broken electrode POz. Particularly for subject 10 and 11 a big difference in workload prediction can be recognized when using 15 versus 16 channels (see Figure 11.9). The *GD* reached 9.52 for subject 10 and 10.1 for subject 11. For both subjects, the workload prediction results were shifted to a higher *Q*-value in the offline simulation, excluding electrode POz. This might be the reason for low performance results of both subjects (see Table 11.1) based on underrated workload leading to erroneous adaptation of the learning material.

Table 11.3: Performance results of the cross-subject workload prediction utilizing the *RMSE* and the *GD* between the predicted Q using 16 channels and excluding the electrode POz for each trial and subject of the experimental group.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
RMSE	3.37	1.25	3.73	4.38	2.84	2.13	1.31	2.73	1.03	3.23	3.47	1.19	1.08
GD	9.39	0.964	11.6	18.2	6.13	3.67	0.153	5.85	0.0214	9.52	10.1	0.503	0.0592

For subjects 4, 5 and 6 the prediction bias were highest in the first trials (see Figure 11.9). As for subject 10 and 11, excluding electrode POz led to a higher workload prediction. Again, the actual workload at the beginning of the online learning phase seems to be underrated, which might be the reason for the linearly increasing difficulty. Negative Q-values for workload prediction can be noticed in Figure 11.9 for subjects 3, 4, 5, 8, 10 and 11 during the online learning phase. For subjects 3, 4, 5 and 8, negative Q-values occurred using 15, as well as 16 electrodes for workload prediction. Possibly these values indicate subjects being disengaged due to the task being too difficult or too easy.

11.7.2 Offline workload prediction results of the control group

For the control group, the offline workload prediction was calculated based on the same regression model as for the experimental group in the online setup (see section 11.3). Merely nine electrodes (FC3, FCz, FC4, C3, Cz, C4, P3, Pz, P4) were used as features.
In Figure 11.10, the distribution of the actual task difficulty and the offline predicted workload for cross-subject regression is shown. For subject 6 no Q-values were saved, due to technical problems. For 8 out of 12 subjects (subjects 1, 2, 3, 5, 8, 10, 11 and 12) the amount of workload was held in a constant range over time. For subjects 1, 3, 8, 10 and 11, the predicted workload range was $0.8 < Q < 3.5$. These participants were kept in the predefined optimal workload state during the whole learning phase. The subjects 2, 5 and 12 needed a constant amount of workload, whereas the workload range for the remaining subjects 4, 7 and 9 was highly distributed and ranged in $0 \leq Q < 5.5$. For subjects 7 and 9 a decrease of predicted workload was recognized over time. This observation led to the assumption, that both subjects were overwhelmed ($Q > 3.5$) at the beginning of the learning phase and after more than 100 trials, they reached a good workload state for learning. For subject 4 an increase of the predicted workload correlated with increasing task difficulty. This subject seemed to exceed the predefined perfect workload range for learning after solving more than 100 trials.

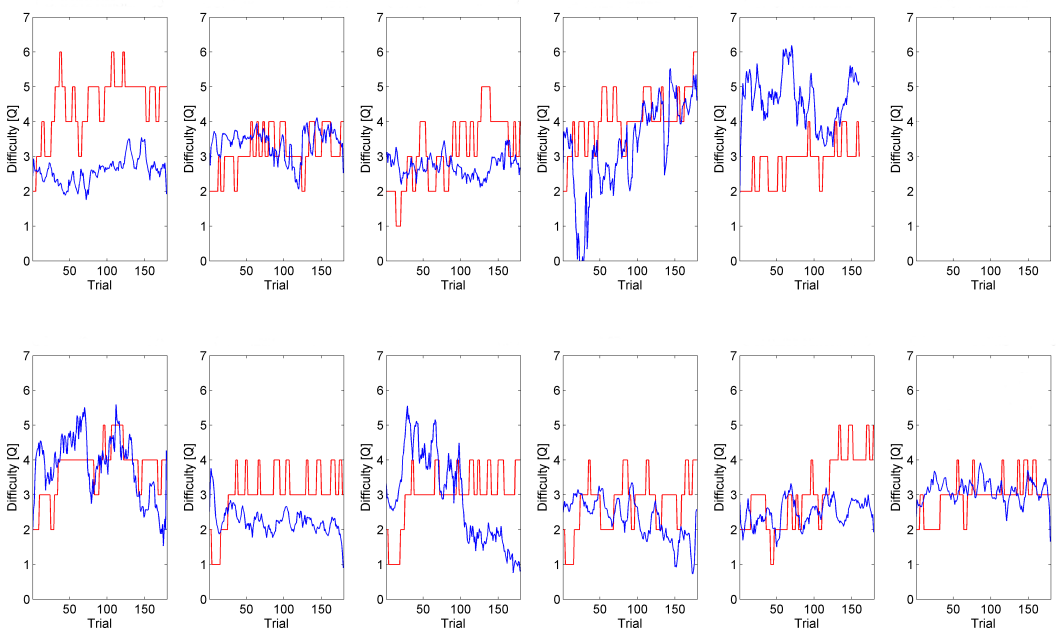


Figure 11.10: Difficulty prediction using a linear regression on features from 9 EEG channels (blue line) and actual difficulty level (red line) of the current math problem for all subjects of the control group. Participants shown here are enumerated from top left to bottom right line by line. Subject 5 quit after 160 trials. For subject 6 no Q-values were saved, due to technical problems.

11.8 Learning effect using adaptive learning environments

In the previous study (see chapter 10), common performance metrics as the *CC* and *RMSE* between the presented and predicted Q-values were used for analyzing the workload prediction. These performance measurements are unsuitable for the EEG-based online adaptive learning environment, since in an adaptive setting the amount of workload remains the same, while the task difficulty increases. Thus, these measures are not meaningful. Therefore, the learning effect after completing the learning phase serves as performance measure for the adaptation and is used as an indicator of how successful each subject was supported during learning. Hence, each individual subject of the experimental group, as well as of the control group had to perform a pre-test before the learning phase started. This was used to assess the prior knowledge of each user. After the learning session, each participant had to solve a post-test, which was necessary for measuring the learning effect.

11.8.1 Learning effect using an EEG-based adaptive learning system

Table 11.4 reports the learning effects of each individual subject of the experimental group. Furthermore, average values calculated for all subjects, as well as for all subjects besides subjects 4, 5, 6, 10 and 11 (= select) are shown. As for these five subjects the adaptation was partially based on the broken electrode POz (see section 11.7.1.2) and therefore confounded with noise, no reasonable EEG-based adaptation can be guaranteed.

Table 11.4: Number of correctly solved trials in the pre-and post-test, as well as the learned factor for each subject of the experimental group and overall subjects. Since for subject 4, 5, 6, 10 and 11 a broken electrode might have influenced the adaptation, a mean value based on the remaining 8 subjects was additionally calculated (= select).

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	mean	
														all	select
Pre-Test	3	0	6	0	2	1	0	1	0	5	1	1	0	1.54	1.37
Post-Test	7	0	9	8	2	3	3	2	8	7	7	2	8	5.08	4.88
Learned	4	0	3	8	0	2	3	1	8	2	6	1	8	3.54	3.51

Table 11.5: Number of correctly solved trials in the pre- and post-test, as well as the learned factor for each subject of the control group. Because of technical problems, no Q-values and given answers were saved for subject 6.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	mean
Pre-Test	4	3	1	0	2	–	1	2	1	1	2	0	1.55
Post-Test	8	2	4	7	5	–	8	3	4	4	6	4	5
Learned	4	-1	3	7	3	–	7	1	3	3	4	4	3.45

Furthermore, no individual support for these subjects was possible and thus the learning success could be negatively influenced. That is why only the mean results of the selected 8 subjects (i.e., subjects 1, 2, 3, 7, 8, 9, 12 and 13) will be stated in the following. On average 1.37 of 11 pre-test assignments were solved from the experimental group successfully. The best subject solved 54.54 % of the pre-test tasks correctly, whereas the least performing subjects gave no correct answers. After using the EEG-adaptive learning environment and thus learning how to calculate using base 8, a learning effect can be recognized for almost every subject, except for subject 2 and 5. On average, 4.88 assignments from 11 post-test tasks were solved correctly after completing the learning phase (see Figure 11.11 and Table 11.4), 3.51 more tasks compared to the pre-test. Although subject 4 and 11 did not receive optimal learning support, a high learning effect can be measured after completing the learning phase. Subject 4 solved 8 tasks more after the learning phase, compared to the pre-test and subject 11 answered 7 assignments in the post-test correctly, 6 tasks more than in the pre-test. It can therefore be assumed that these two subjects have learned how to calculate using base 8 regardless of the presented degree of task difficulty. On average a significant learning success can be verified between the pre- and post-test ($p = 0.008$, two-sided Wilcoxon test). Hence, the EEG-based learning environment seems to successfully support nearly all participants in their learning process of how to calculate using base 8.

11.8.2 Learning effect using an error-based adaptive learning system

The control group solved 1.55 tasks from the 11 pre-test assignments on average correctly, very similar to the experimental group. Thus implies equal prior knowledge, which makes both groups comparable. The best subject performed 36.36 % of the pre-test tasks accurately, whereas the worst subjects gave no correct answers (see Table 11.5). For almost every subject, a learning effect of how to calculate using base 8 is noticeable after the error-

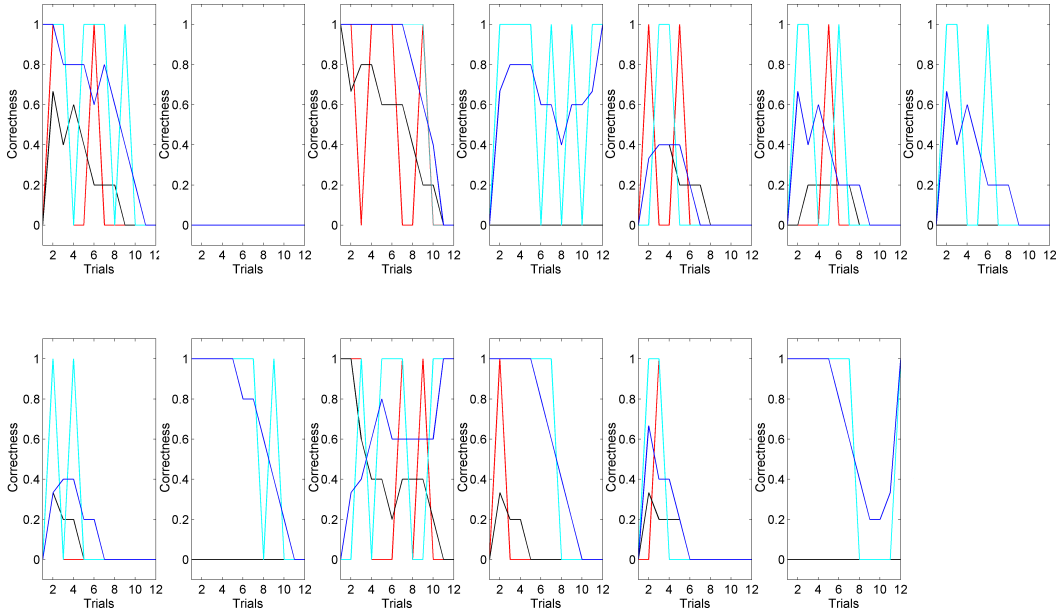


Figure 11.11: The amount of correctly solved tasks in the pre-test (red line) compared to the number of trials solved correctly in the post-test (turquoise line) from the experimental group, whereas 1 means correct and 0 means wrong. The results are plotted as smoothed lines to see a clearer trend, for the pre-test (black line) as well as for the post-test (blue line). Participants shown here are enumerated from top left to bottom right line by line.

adaptive learning session, except for subject 2. On average, 3.45 more tasks were solved correctly in the post-test, compared to the pre-test (see Table 11.5 and Figure 11.12). As for the experimental group a significant learning success can be verified between the pre- and post-test ($p = 0.0016$, two-sided Wilcoxon test). Thus, the error-based learning environment seems to support nearly all participants in their learning process of how to calculate using base 8.

In the following section the two adaptive learning environments will be compared. Although both adaptation schemes yield similar results, the benefits and innovative applications of using EEG data for adapting a learning environment, will be pointed out in chapter 12.

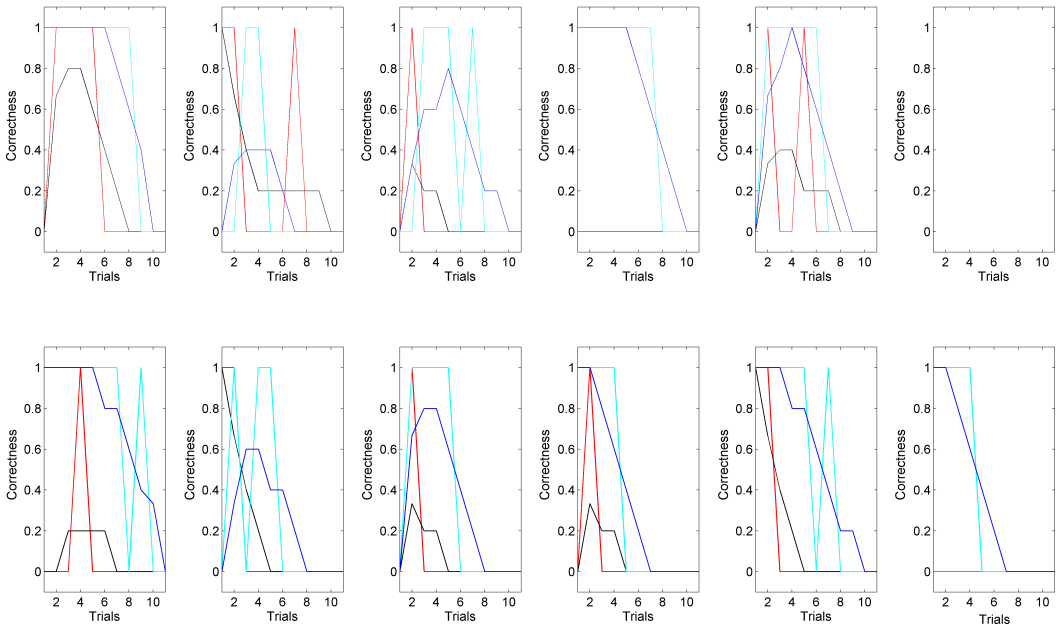


Figure 11.12: The amount of correctly solved tasks in the pre-test (red line) compared to the number of trials solved correctly in the post-test (turquoise line) of the control group. Whereas 1 means correct and 0 means wrong. The results are plotted as smoothed lines to see a clearer trend, for the pre-test (black line) as well as for the post-test (blue line). Participants shown here are enumerated from top left to bottom right line by line. For subject 6, no answers were saved, due to technical problems.

11.9 Discussion

In the present chapter it is shown that EEG data can be utilized to adapt a learning environment in real-time, by using a cross-subject regression method without individual calibration phase. Furthermore, the learning effect of the EEG-based adaptive learning environment was compared with performance results after using an error-adaptive learning environment, which is state of the art. The comparison revealed, that both learning environments induce learning effects on how to calculate using base 8. In the following, the benefits and drawbacks, as well as further ideas for adaptive learning environments will be discussed.

11.9.1 EEG-based vs. error-based adaptive learning environment

To evaluate the performance of the innovative EEG-based learning environment, the learning factor of each subject is compared with the learning effect for participants using the error-based adaptive learning environment. Thus, the difference of correct answers in the pre- and post-test after solving the learning phase with the EEG-based or error-based adaptive learning environment are studied.

First, it is necessary to prove that subjects across both groups have similar prior knowledge about calculating using base 8. As there are no significant differences ($p > 0.05$, two-sided Wilcoxon test) between the prior knowledge of both groups, the groups can be assumed as homogenous and the learning factors between the learning environments are comparable. Second, the learning factors are evaluated. No significant differences ($p > 0.05$, two-sided Wilcoxon test) can be documented for both groups. Therefore, it is uncertain if the EEG- or error-based adaptive learning environments support subjects better.
Comparing the mean of successfully solved tasks for the pre- and post-test data for each group leads to the assumption, that the learning effect in the EEG-based learning environment ($\varnothing = 3.5$) is slightly higher than for the error-based learning environment ($\varnothing = 3.45$), although it is not significant. This might be an indicator that the EEG-based adaptation is preferable to the error-based adaptive system.
To intensify this effect, a larger subject size for both groups is necessary. Optimally, the group of participants should be aged-matched and should have related workload capacity ranges, because these properties might be influencing factors for the performance of the learning environment.

11.9.2 Influencing factors for the performance of a learning environment

Diverse individual workload capacities and different prior knowledge can be influencing factors for the performance of a learning environment. Furthermore, technical problems, as well as unknown features can lead to wrong adaptation of the learning material. Therefore, these factors will be discussed in the following sections in more detail.

11.9.2.1 Adaptation errors due to individual workload capacities

An inappropriate adaptation of the learning material can be caused by diverse workload capacities of each individual learner. In the experimental group, subjects 1, 5, 6, 7 and 10 might have smaller or larger workload capacities as the remaining subjects, so that the workload range of $0.8 < Q < 3.5$ is not suitable for supporting them in their learning process successfully. Thus, the optimal difficulty level of the presented learning material has to be lower as they have no workload capacity left, or higher as they have more workload capacity left. To overcome this problem, a workload capacity test can be accomplished for each subject at the beginning of each session, to adapt the Q-values defining the optimal workload range individually. Hence, the varying workload capacities might be taken into account for the learning material adaptation and the learning environment can work more precisely for each individual subject.

11.9.2.2 Effect of broken electrode on online workload prediction

As stated in section 11.7.1.2, the damaged electrode POz led to wrong workload prediction in the EEG-based adaptive learning environment and thus to erroneous adaptation of the learning material. Predefined, a subject with a predicted workload of $Q > 3.5$ was assumed to be overwhelmed by the current task. Therefore, the difficulty level of the subsequent addition task decreased. For participants 10 and 11 it is recognizable, when using

16 electrodes for workload prediction as in the online setting, the predicted workload was constantly $Q < 2$. But excluding electrode POz led to workload predictions with $Q > 3$. This causes the assumption, that the difficulty level of the presented learning material at the beginning of each learning phase was optimal for these two learners. But the regression model based on 16 channels predicted an insufficient workload state and increased the level of difficulty due to the damaged electrode.

11.9.2.3 Adaptation errors due to unknown EEG features

A negative Q-value prediction was noticeable during the online learning phase (see section 11.7.1.2). Predicting a negative Q-value leads to the assumption, the calibrated regression model is not aware of the features in the specific trial. As mentioned in the previous chapter 10.7.1, this effect in the EEG data is likely induced by disengagement due to a task being too difficult or too easy. Disengagement leads to an U-shaped EEG pattern, meaning neural signatures of disengagement are similar to EEG patterns for low cognitive workload [146, 154]. Thus, to adapt the learning material correctly, the trial with a negative Q-value prediction should not be interpreted as too easy, but as too difficult. For subjects 10 and 11, the negative predicted Q-values are based on channel 16 and disappeared when excluding electrode POz. As the negative Q-values for these subjects are obviously induced based on the broken electrode and not based on brain signals, they have no further meaning.

11.9.3 Drawbacks of missing calibration phase per subject

As described in chapter 10 the possibility to calibrate a generalized regression model independently from each subject is an essential benefit for realizing an adaptive learning environment. Still, some drawbacks of missing calibration phases occur.

11.9.3.1 EOG regression on individual basis

In section 10.7 it was demonstrated that eye-movements can be a confounding factor in EEG-based adaptation of numerical learning material. Nevertheless, EOG correction was not used in this study. Technically, EOG regression is practicable across subjects, but it is unknown how this will affect the EEG signal quality in an online cross-subject study. Applying the EOG regression for each subject would result in recording training data per subject for the EOG artifact correction. The data size would not be as large as for calibrating the workload prediction regression, but the benefit in completely avoiding the calibration phase would not exist anymore.
The topography plots (see Figure 11.5), over all subjects clearly showed a difficulty related effect over the central electrodes, in the delta-frequency band, while a small effect can be measured over the parieto-occitiptal electrodes in the alpha-frequency band. As these patterns are similar to the patterns in section 10.6.2.2 after EOG correction, it can be assumed, that the neurophysiological features which were used for the adaptation in EEG-based learning environment, are partially based on workload related brain signals in the delta- and alpha-frequency bands.

11.9.3.2 Subject specific normalization of the EEG data

The normalization step from section 10.6.2.2 is not practicable in an online adaptive learning environment without calibration phase. In the previous study, the first 30 trials were used for a baseline correction. Since in the current study no calibration session is conducted, no such trials exist. To improve the performance of the learning environment, as well as to avoid artifact contaminated EEG data, a transfer learning approach could be applied. Thus, each subject had to fulfill 30 trials that do not induce learning effects. These data could then be used for EOG regression as well as for baseline correction.

11.9.3.3 Individualized workload range

Each subject has an individual cognitive workload capacity range. In this study, a fixed predefined range of $0.8 < Q < 3.5$ was assumed to be optimal for all subjects, based on the results in section 10.4. But it is recognizable, that the performance varies over subjects. Some subjects can solve 50 % on average when assignments with $Q = 3$ are presented, whereas others can still solve tasks with $Q = 4$ with 50 % correctness. By accomplishing a span task at the beginning of each learning phase, the individual's workload capacity of each learner could be measured. Based on the results, the Q-thresholds could be defined for each subject individually. Furthermore, the prior knowledge of each user should be taken into account. The starting difficulty can be adapted individually, based on the prior knowledge.

11.9.4 Comparison with the state of the art

In previous studies, either EEG data recorded in a real learning setting was used for offline analysis [116, 117] or EEG data was utilized for adapting material in well controlled working-memory tasks online [50]. These studies have nothing to do with real-world learning settings. There are a few studies, adapting video games online based on diverse workload levels or affective states, identified in EEG data [122, 123]. Furthermore, isolated studies are dealing with real time adaptive tutoring systems reacting on affective states, detected in EEG data [124, 125, 126] to support students in there learning process. Though, up to my knowledge, there are no studies dealing with online workload detection based on EEG data and complex learning material adaptation.

11.10 Conclusion

The results have shown, that EEG-based adaptive learning environments do not necessarily lead to better learning effects than the common state of the art (error-adaptive learning environments), but for the error-based adaptive learning environment the required amount of workload is not always in the predefined optimal range for each subject. The EEG-based adaptive learning environment is a promising method to measure workload states in a non-obtrusive and objective way. Despite using the novelty of the online cross-subject regression, as well as the continuous workload measurement, this is the first study where

EEG data is used in a real learning context to adapt the difficulty of the presented material online. Furthermore, the workload level is constantly kept in a predefined range. The EEG-based learning environment supported the learners successfully in calculating using base 8, as the learning effect was significant ($p = 0.008$, two-sided Wilcoxon test). Although the results are promising, further modification could result in an even more precise and successful adaptation of the learning environment. Defining individual Q-value thresholds, depending on individual workload capacities as well as interpreting negative Q-value predictions as disengagement, could increase the performance of the EEG-based adaptive learning environment. Furthermore, a repetition of smaller learning phases over an extended period of time might be recommendable, to induce larger learning effects. Thus, using EEG-based learning environments as in this study, it is indeed possible to predict the user's workload in real-time by using an non-obstructive, objective measurement and adapt the presented complex learning material accordingly, to support students successfully in their learning process.
To conclude, the error-based learning environment supports the learner in calculating using base 8 well, but the required amount of workload is not always in the predefined optimal range for each subject. Furthermore, most subjects were not able to reach higher difficulty levels as $Q > 4$. Thus, a repetition of smaller learning phases over an extended period of time is recommendable, to induce larger learning effects. This led to the assumption, that the error-adaptive method is promising, but can be improved by additional workload indicators.

12 Summary and conclusion

The aim of this thesis is reached. An EEG-based adaptive learning environment was developed, which can predict the actual workload of a learner, adapt the presented learning material based on the cognitive workload state in real-time and support each individual user in his/her learning process successfully.
Before discussing the challenges involved in developing an EEG-based adaptive learning environment, the previous results will be summarized, to give an overview regarding the developed and evaluated methods for this thesis.
Conducting the first study (see chapter 5), where subjects had to learn angle theorems and reading comic-strips, demonstrated, it is possible to differentiate diverse cognitive states in the EEG signals. Furthermore, a major challenge for the development of a real-world learning environment became clear. Presenting complex and too diverse material causes artifacts and induces confounds in the EEG data. Therefore, the design and topic of the instruction material in the learning environment has been carefully deliberated.
By utilizing a variety of pre-processing and feature extraction methods, as connectivity or ICA, the most suitable features for an online workload detection were specified in the second study, reported in chapter 6.
The following studies were essential to enable the development of an efficient learning environment based on EEG data. To reduce the calibration time of a classifier, as well as to generate a generalized classifier based on workload specific features in the EEG data, the cross-task classification method was developed (chapter 7). Subsequently, the influence of subjective cognitive workload labeling (chapter 8) and the task order effect during cross-task classification (chapter 9) were analyzed. As this classification method caused low workload detection accuracies around chance level and to completely avoid the calibration session of a classifier, a second classification method called cross-subject regression was utilized (see chapter 10). By using a leave-one-subject-out validation, a precise workload prediction across subjects was achieved offline. Since the results from the cross-subject study reported in chapter 10 were promising, the cross-subject regression was applied in an online study (see chapter 11). Thereby the performance of the innovative EEG-based adaptive learning environment was compared with an error-based adaptive learning environment used by a control group. The usage of cross-subject regression in an EEG-based adaptive learning environment was successful. The difficulty of the presented learning material was adapted in real-time, dependent on the predicted workload of each subject. Furthermore, the workload range of each user was held in an optimal span. Compared to an error-based adaptive learning environment, which is state of the art, the learning effect induced by the EEG-based learning environment is better, but not significant.

12.1 Suitable learning material for an adaptive learning environment

Developing learning material which is suitable for an EEG-based adaptive learning environment causes some challenges which have to be considered. First, complex learning material can induce more artifacts in EEG data than easy learning material. Furthermore, tasks which are perceptually not identical can lead to differences in semantic processing or eye-movements. Workload states induced by controlled experimental working-memory tasks can differ from workload states induced by realistic learning tasks. These instructional as well as perceptual confounds can be picked up by a classifier, possibly causing a classification to not being based on diverse workload states, but on artifacts. Therefore, assignments with an increasing level of difficulty should not differ in their task design and complexity, but only in the amount of workload needed for solving a task successfully. Second, it should be possible to evaluate the presented learning material depending on the degree of difficulty, utilizing objective criteria. The categorization of learning material into diverse levels of difficulty is necessary to enable a supervised classification training. Depending on the objective associated difficulty level, each trial gets labeled for classifier training. Therefore, the criteria should be an objective and stable measurement across a variety of subjects. Third, learning material should always be presented in an increasing level of difficulty, due to the level of expertise of each learner. Thus, confounding task difficulty and task order is unavoidable for learning material. Providing an opportunity to normalize and baseline correct the EEG data for further analysis is helpful, counteracting negative fixed order effects.

12.2 Generalizable classification methods

Developing an adaptive learning environment based on EEG data for online workload prediction across subjects, is the goal of this thesis.

Using within-task classification methods for real-world learning environments are not feasible. The collection of training data for the classifier in combination with a learning environment remains a big challenge. The recording is time consuming for the participants, as a high number of training trials is needed. This can cause frustration or loss of motivation, which are crucial factors for successfully performing in a learning phase. Furthermore, the complex learning material cannot be used as training data, as subjects will develop solving strategies in the calibration phase before solving the actual learning phase. Due to the increasing knowledge during the classifier calibration phase, the amount of required workload will decrease over time, for the same degree of difficulty. Thus, EEG patterns used for classifier training are not reproducible for classifier testing. Finally, generalizable classifiers are desirable for adaptive learning environments. The calibrated classifier should be able to detect different workload states task independently. Two modified classification methods like cross-task classification and cross-subject regression can counteract these challenges.

Cross-task classification reduces the calibration time of a classifier and enables the development of a generalized classifier. By using well defined working-memory tasks for classifier training, a precise variation of workload states is possible. But applying the cross-task classification on real-learning material leads to classification accuracies around chance level. Thus, the transfer of the trained classification models to independent test data from the application phase does not work successfully so far. Furthermore, a calibration phase is still necessary to enable the classifier training. To conclude, the intention of training a classifier based on a mixture of specific workload components, to apply the calibrated model in a complex learning context, is sophisticated. Because of training and test material being too diverse, this generalization is not practicable so far. For a first improvement, the transfer step from simple working-memory tasks to complex learning tasks should not be too difficult. Thus, optimal learning material has to be developed, varying merely the amount of workload during the testing phase which is also manipulated during the simple training tasks.
The second method, cross-subject regression, completely avoids the individual classifier training, as models are trained on datasets from independent subjects. By using a leave-one-subject-out validation, a precise workload prediction across subjects was reached. Compared to the cross-task classification, high prediction accuracies can be achieved. Furthermore, using cross-subject regression in an EEG-based adaptive learning environment worked as successful as the current standard. Enabling the accurate workload prediction in real-time across subjects is an important step to reach the aim of this thesis.

12.3 Non-linear effects in EEG data

Analyzing the EEG data in the conducted studies causes an interesting finding when solving difficult tasks. An U-shaped EEG pattern can be recognized in the power spectra. Since this pattern is detectable in three studies (see chapter 6, chapter 7 and chapter 10), independent of the presented learning material or the methods of pre-pocessing or analyzing the EEG data, it leads to the assumption that the characteristic can just be attributed to task difficulty of the presented learning material. Increasing task difficulty leads to a desynchronization in the alpha-frequency band and a synchronization in the theta-frequency band. These characteristics can be precisely classified across subjects. However, a sharp increase of the difficulty level might cause disengagement detectable by a reversed EEG pattern, alpha-synchronization and theta-desynchonization. If learning tasks are too difficult and thus require an amount of cognitive workload higher than their optimal workload range, results in disengagement. The neural signatures of disengagement are similar to EEG patterns for low cognitive workload. These non-linear effects in the EEG data cause imprecision in workload prediction, as it is impossible to differentiate between the cognitive workload state being too high or too low for learners.
Therefore, taking additional parameters into account when adapting the learning material, seems to be advisable. Besides workload, also vigilance and engagement are important factors for solving a task correctly and learn efficiently. Decreasing vigilance is often specified as the decline in attention-requiring performance over an extended period of time. Further-

more, vigilance increases steeper in the context of difficult, compared to easy tasks. Tiwari et al. [164] describe the interrelationship of vigilance and workload. An increasing workload is accompanied by a vigilance decrement. Engagement and workload increased as a function of task difficulty during learning and memory tasks [165]. The results from previous studies [164, 165, 166] showed, the detection of various vigilance and engagement states is possible. Therefore, measuring EEG data and extracting components specific for workload, vigilance and motivation is recommendable. Using workload as basic component to keep the subjects in their optimal workload range is unavoidable. Combining this EEG patterns with additional information about vigilance and engagement might enable the classifier to differentiate between disengagement and low workload. Thus, the postulated U-shaped EEG pattern induced by increasing workload can be decoded using additional information from the EEG signal. The ability to continuously and unobtrusively measure vigilance, engagement and cognitive workload in a learning environment can be useful in identifying the optimal adaptation scheme for the presented learning material, which can support learners successfully.

12.4 Possibilities to improve workload prediction

The workload prediction across subjects works precise and is a promising step for the development of an efficient classification method for an EEG-based adaptive learning environment. Although, it is desirable to avoid the individual training phase for each classifier, it offers disadvantages for the following reasons.

EOG artifact correction

Eye-movements can be a confounding factor in EEG-based adaptation of numerical learning material (see section 10.7). Until the time of writing, the effect of using online EOG regression across subjects is unknown. To explore how precise an EOG regression across subjects works and how large the benefit is, additional studies have to be conducted.

Subject specific normalization of the EEG data

Normalization of the EEG data is recommended, because of confounds in the EEG data induced by the fixed order of learning material, or inter-subject variances in the EEG power baseline. An individual normalization step as in section 10.6.2.2 is not practicable in an online adaptive learning environment without an individual calibration phase. Avoiding the classifier training completely will not allow for a subject specific normalization. But performing a short calibration phase, where individual EEG power baseline is recorded, might improve the performance of the workload prediction. Thus, a transfer learning approach could be applied, to train a classifier or regression model based on a database of a group of subjects and recalibrate it for each individual subject independently.

Individualized workload range

Each subject has an individual cognitive workload capacity range. Learners can broadly be categorized in three classes: highly intelligent, normal and learning-disabled. In the conducted studies, a fixed predefined range was assumed to be optimal for all subjects, but the performance varies across subjects. Highly intelligent subjects can solve more complex

tasks with the same amount of workload as learning-disabled persons need for solving easy tasks. Therefore, the determination of a personal workload range might improve the adaptation of learning material. By accomplishing a span task or an intelligence test at the beginning of each learning phase, the workload capacity of each learner can be measured. Thus, the Q-thresholds can be defined for each subject individually.
Furthermore, having a great number of participants, three regression models could be created using the cross-subject regression method. The model calibration would be based on EEG data from highly intelligent subjects, or normally intelligent participants, or persons with learning disabilities. Defining the workload capacity range for each subject at the beginning of the learning phase, the optimal cross-subject trained regression model can then be chosen to support the learner in his/her optimal learning speed.

12.5 Benefits of an online workload indicator

Applying online adaptive learning environments as in chapter 11 support learners in their learning process successfully. EEG is, as of today, the most promising physiological signal which can accurately reflect cognitive states, e.g., workload, in a non-obstructive and objective way. Identifying and quantifying workload on a second-by-second time-frame can be used for more than adapting learning material, as reported in this thesis. By using information about additional cognitive states as vigilance or engagement an optimal support for learners can be ensured. Whereas varieties in workload are attended by error making, vigilance and engagement cannot be explained by the error rate. The interplay of these three components can be used to individualize a learning phase for each user. Measuring the cognitive states continuously can lead to optimal break time recommendations for each individual participant. Furthermore, specific cognitive state based advices for task solving can be inserted, to draw the attention of the learner and support her/him optimally.

12.6 Challenges of evaluating adaptive learning environments

In non-adaptive learning environments or offline studies, common performance metrics as the *CC* and the *RMSE* between the presented and predicted workload levels (i.e., Q-values) are used for analyzing the prediction performance. In the online adaptive learning environment these performance measurements are unsuitable. During a non-adaptive setting, an increase of task difficulty leads to an increase of workload. Thus the *CC* and *RMSE* are reliable measurements for evaluating the performance. But in an adaptive setting, where the aim is to increase the task difficulty while the amount of workload remains the same, these measures are not meaningful. Therefore, the learning effect is a preferable indicator of how successful each subject is supported during learning. The learning effect can be measured by measuring the prior knowledge with a pre-test before solving the learning phase. Furthermore, the knowledge has to be determined with a post-test after completing the learning phase. Calculating the difference of both learning states (before and after solving the learning phase) of each subject leads to a degree measure, the learning effect

can be detected with. Although it is a good indicator to identify the learning success of each subject, it is entangled with the amount of prior knowledge each subject has.

12.7 Comparison with the state of the art

Developing generalized classification methods to adapt learning environments in real-time, based on EEG data, is at the cutting edge of research. By the time of writing, only a few studies are dealing with cross-task classification. They either reached classification results just around chance level [110], used similar tasks for training and testing the classifier, as spatial n-back and verbal n-back [111], or had just a small number of participants. Compared to the state of the art, cross-task classification led to good classification results in this thesis, while using a combination of diverse working-memory tasks for classifier training. Applying subjective cognitive workload ratings for class labeling can further improve the cross-task classification results. Modifying the class labeling process for SVMs is a new and innovative idea which has potential to improve the cross-task classification results. As far as I know, the reported study in this thesis is the first one, dealing with it for workload classification during learning. Although the high cross-task classification accuracies of this thesis can be induced by task order effects, the results are comparable with the classification accuracies form Gevins et al. [111] and Baldwin and colleagues [110], since they do not even control for task order effects.
Utilizing cross-subject regression, the second method for developing generalized classifiers in this thesis, caused precise workload prediction in real-time, across subjects. The prediction results are at least 6 % better than the results from previous studies [111, 112, 113]. While the study reported in this thesis only used workload specific EEG features, the other studies [111, 112, 113] needed a higher number and more widespread EEG features to reach good classification results. Furthermore, none of these studies used realistic learning tasks. By my knowledge, the cross-subject regression stated in this thesis, is the most successful method for predicting levels of workload online, while solving a complex learning task.
Using EEG data to adapt the presented material in an adaptive learning environment, instead of error counting (i.e., state of the art [1, 2]), led to better learning effects, but not significantly. Nevertheless, despite the precise and continuous workload prediction, EEG data can further be used to detect additional parameters as vigilance and engagement in real-time. Combining the EEG patterns of workload with additional information about vigilance and engagement might improve the classifier in differentiating even more mental states. Moreover, using additional EEG characteristics might enhance the adaptation process of an EEG-adaptive learning environment. Parameters as vigilance and engagement cannot be measured by an error-based learning environment. Therefore, EEG signals are a recommendable and promising measurement for developing good adaptive learning environments, to support learners perfectly.

12.8 Outlook

Some open questions still remain and should be explored in further studies. The findings of this thesis suggest, that a transfer learning approach might improve the performance of the learning environment and avoid artifact contaminated EEG data. Thus, each subject has to fulfill 30 trials which do not induce learning effects. Math tasks based on study 4 reported in chapter 10 can be used. The maximal Q-value can be determined while calculating using the base 10, but no learning effects of how to calculate using the base 8 are induced. Furthermore, these data can then be used for EOG regression, as well as for individual baseline correction. By using an additional pre-test directly before completing the learning phase, the prior knowledge of each subject can be discovered and thus the starting Q-value individually be adapted.
Furthermore, combining the EEG-based adaptive learning environment with the state of the art method seems to be promising. Finding the individual workload range of each user by analyzing the EEG data, predicting the actual workload states and validating the prediction based on error making can lead to a successful adaptation method. The addition of further information about cognitive states as vigilance or engagement can improve the individual support enormously.
By using additional parameters for workload detection, as eye-tracking (e.g., analyzing pupillometry) or NIRS, additional input channels might be used for adapting the presented learning material more precisely.
For the instructional design, the comparison of adaptive and adaptable learning environments is a highly interesting research question. In this context, adaptive means a learning environment accommodates directly to the cognitive workload state of the user by changing the presented material. Adaptable does not change the difficulty level automatically but gives feedback to the user about the measured workload state. This method gives subjects the opportunity to change their vigilance, engagement and workload state actively themselves.
The most promising learning environment should subsequently be applied to students with learning disabilities or attention deficit hyperactivity disorders, so they can be supported in their learning process successfully.

Bibliography

[1] Alber Corbett. Cognitive computer tutors: Solving the two-sigma problem. *Proceedings of the 8th International Conference on User Modeling*, pages 137–147, 2001.

[2] Arthur Graesser and Danielle McNamara. Self-regulated Learning in Learning Environments with Pedagogical Agents that Interact in Natural Language. *Educ. Psychol.*, 2010.

[3] Tian Lan, Deniz Erdogmus, Andre Adami, Santosh Mathan, and Misha Pavel. Channel selection and feature projection for cognitive load estimation using ambulatory EEG. *Computatinal Intelligence and Neuroscience*, page 74895, 2007.

[4] G. Pavel, M. Wang and K. Li. Augmented cognition: allocation of attention. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*, Bis Island, Hawaii, USA, 2003.

[5] Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. Classifying single trial EEG: Towards brain computer interfacing. *Advances in neural information processing systems*, 1:157–164, 2002.

[6] J. Sweller, J. J. G. Van Merriënboer, and F. Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10:251–296, 1998.

[7] L.S. Vygotsky and M. Cole. *MIND IN SOCIETY*. Harvard University Press, 1978.

[8] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O Zander. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*, 8, 2014.

[9] Nelson Cowan. Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, 26(2):197–223, June 2014.

[10] A. Baddeley. *Working Memory*. Clarendon Press., 1986.

[11] Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423, November 2000.

[12] A. Baddeley. Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63:1–29, January 2012.

[13] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller. The expertise reversal effect. *Educational Psychologist*, 38:23–31, 2003.

[14] Fred Paas, J.E. Tuovinen, Huib Tabbers, and Pascal W.M. Van Greven. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, 38(1):63–71, 2003.

[15] T.D. Wagner and E.E. Smith. Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective and Behavioral Neuroscience*, 3(4):255–274, 2003.

[16] Alan Baddeley. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839, 2003.

[17] W. Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2-3):169–195, Apr 1999.

[18] E. Klein, H.C. Nuerk, G. Wood, and K. Knops, A. Willmes. The exact vs. approximate distinction in numerical cognition may not be exact, but only approximate: How different processes work together in multi-digit addition. *Brain and Cognition*, 69:369–381, 2009.

[19] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[20] Roland Brunken, Jan L Plass, and Detlev Leutner. Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1):53–61, 2003.

[21] Gabriele Cierniak, Katharina Scheiter, and Peter Gerjets. Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, 25(2):315 – 324, 2009. Including the Special Issue: State of the Art Research into Cognitive Load Theory.

[22] Krista E. DeLeeuw and Richard E. Mayer. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100(1):223–234, Feb 2008.

[23] L. Mihalca, R.J.C.M. Salden, G. Corbalan, F. Paas, and M. Miclea. Effectiveness of cognitive-load based adaptive instruction in genetics education. *Journal in Computers in Human Behavior*, 27(1):82–88, January 2011.

[24] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1477–1480. ACM, 2004.

[25] Arthur F Kramer. Physiological metrics of mental workload: A review of recent progress. Technical report, DTIC Document, 1990.

[26] Willem B Verwey and Hans A Veltman. Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of experimental psychology: Applied*, 2(3):270, 1996.

[27] A. Gevins and M. Smith. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, 4(1):113–131, 2003.

[28] P. Antonenko and D.S. Niederhauser. The influence of leads on cognitive load and learning in a hypertext environment. *Computers in Human Behavior*, 26:140–150, 2010.

[29] P. Antonenko, F. Paas, R. Grabner, and T. van Gog. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438, 2010.

[30] I. Gerlic and N. Jausovec. Multimedia: Differences in cognitive processes observed with EEG. *Journal of Technology Research and Development*, 47(3):5–14, 1999.

[31] Hans Berger. Über das Elektroenzephalogramm des Menschen. *Archiv für Psychatrie und Nervenkrankheiten*, 87:527–570, 1929.

[32] D. Heger, F. Putze, and T. Schultz. Online Workload Recognition from EEG Data during Cognitive Tests and Human-Machine Interaction. In ; DilBeyerer, J.; Beyerer, U. D.; Hanebeck, and T. Schultz, editors, *KI2010*, volume 6359, pages 410–417. Lecture Notes in Computer Science, Springer, Heidelberg, 2010.

[33] S. Sanei and J.A. Chambers. *EEG Signal Processing*. Wiley, 2008.

[34] Ernst Niedermeyer and FH Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[35] G. Pfurtscheller and F.H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110:1842–1857, 1999.

[36] H. Jasper. The ten twenty electrode system of the international federation. *Electroencephalographic and Clinical Neurophysiology*, 10:371–375, 1958.

[37] Jaakko Malmivuo and Robert Plonsey. *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, 1995.

[38] H. Aurlien, I.O. Gjerde, J.H. Aarseth, G. Eldoen, B. Karlsen, H. Skeidsvoll, and N.E. Gilhus. EEG background activity described by a large computerized database. *Clinical Neurophysiology*, 115(3):665 – 673, 2004.

[39] BrainProducts. Brain products press release, December 2011.

[40] Thorsten Zander. *Utilizing Brain-Computer Interfaces for Human-Machine Systems*. PhD thesis, Technischen Universität Berlin, 2012.

[41] Florin Popescu, Siamac Fazli, Yakob Badower, Benjamin Blankertz, and Klaus-R. Müller. Single Trial Classification of Motor Imagination Using 6 Dry EEG Electrodes. *PLoS ONE*, 2(7):e637, 07 2007.

[42] Pavel Bobrov, Alexander Frolov, Charles Cantor, Irina Fedulova, Mikhail Bakhnyan, and Alexander Zhavoronkov. Brain-computer interface based on generation of visual images. *PLoS ONE*, 6(6):e20674, 06 2011.

[43] Dimitii Gribkov and Valentina Gribkova. Learning dynamics from nonstationary time series: Analysis of electroencephalograms. *Physical Review*, 61(6):6538–6545, 2000.

[44] J. Kohlmorgen, G. Dornhege, M. Braun, B. Blankertz, K.R. Müller, G. Curio, K. Hagemann, A. Bruns, M. Schrauf, and W.E. Kinces. Improving human performance in a real operating environment through real-time mental workload detection. In *Toward Brain-Computer Interfacing*, pages 409–422. MIT Press, Cambridge, MA, 2007.

[45] Michael Teplan. Fundametals of EEG measurement. *Measurement Science Review*, 2(2):1–11, 2002.

[46] A. Gevins and M.E. Smith. Electroencephalography (EEG) in neuroergonomics. In R. Parasuraman and M. Rizzo, editors, *Neuroergonomics: The Brain at Work*, pages 15–31. Oxford University Press, Oxford, 2007.

[47] Wolfgang Klimesch, Roman Freunberger, Paul Sauseng, and Walter Gruber. A short review of slow phase synchronization and memory: evidence for control processes in different memory systems? *Brain Res*, 1235:31–44, Oct 2008.

[48] Paul Sauseng, Birgit Griesmayr, Roman Freunberger, and Wolfgang Klimesch. Control mechanisms in working memory: a possible function of EEG theta oscillations. *Neuroscience and Biobehavioral Reviews*, 34(7):1015–1022, Jun 2010.

[49] Roman Freunberger, Markus Werkle-Bergner, Birgit Griesmayr, Ulman Lindenberger, and Wolfgang Klimesch. Brain oscillatory correlates of working memory constraints. *Brain Research*, 1375:93–102, Feb 2011.

[50] A. Gevins, M.E. Smith, L. McEvoy, and D. Yu. High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cereb Cortex*, 7(4):374–385, Jun 1997.

[51] M. Pesonen, H. Hämäläinen, and C. M. Krause. Brain Oscillatory 4-30 Hz responses during a visual n-back memory task with varying memory load. *Brain Research*, 1138:171–177, 2007.

[52] A. Stipacek, H. Grabner, R., C. Neuper, A. Fink, and C. Neubauer, A. Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load. *Neurosci Letter*, 353(3):193–196, December 2003.

[53] C. M. Krause, M. Pesonen, and H. Hämäläinen. Brain oscillatory 4-30 Hz electroencephalogram responses in adolescents during a visual memory task. *Neuroreport*, 21:767–771, 2010.

[54] W. Klimesch, H. Schimke, and G. Pfurtscheller. Alpha frequency, cognitive load and memory performance. *Brain Topography*, 5(3):241–251, 1993.

[55] A. Mecklinger, A.F. Kramer, and D.L. Strayer. Event related potentials and EEG components in a semantic memory search task. *Psychophysiology*, 29(1):104–119, Jan 1992.

[56] R.W. Backs, A.M. Ryan, and G.F. Wilson. Psychophysiological measures of workload during continuous manual performance. *Hum Factors*, 36(3):514–531, September 1994.

[57] F. Boiten, J. Sergeant, and R. Geuze. Event-related desynchronization: the effects of energetic and computational demands. *Electroencephalography and Clinical Neurophysiology*, 82(4):302–309, Apr 1992.

[58] K. Dujardin, P. Derambure, L. Defebvre, L. Bourriez, J., M. Jacquesson, J., and J. D. Guieu. Evaluation of event-related desynchronization (erd) during a recognition task: effect of attention. *Electroencephalography and Clinical Neurophysiology*, 86(5):353–356, May 1993.

[59] G. Pfurtscheller and F. Lopes Da Silva. EEG Event-Related Desynchronization (ERD) and Event-Related Synchronization (ERS). In E. Niedermeyer and L. Da Silva, F., editors, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, chapter 51, pages 1003–1016. Lippincott Williams & Wilkins, 2004.

[60] G. Pfurtscheller. Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest. *Electroencephalography and Clinical Neurophysiology*, 83(1):62–69, Jul 1992.

[61] M.P. Norton and D.G. Karczub. *Fundamentals of Noise and Vibrtion Analysis for Engineers*. Cambridge University Press, 2003.

[62] Nai-Jen Huan and Ramaswamy Palaniappan. Neural network classification of autoregressive features from electroencephalogram signals for brain-computer interface design. *Journal of Neural Engineering*, 1(3):142–150, Sep 2004.

[63] N.-J. Huan and R. Palaniappan. Classification of Mental Tasks Using Fixed and Adaptive Autoregressive Models of EEG Signals. In *Proceedings of the 2nd International IEEE EMBS.*, 2005.

[64] Dean J Krusienski, Dennis J McFarland, and Jonathan R Wolpaw. An evaluation of autoregressive spectral estimation model order for brain-computer interface applications. In *Proceedings of the 28th IEEE EMBS Annual International Conference*, pages 1323–1326. IEEE, 2006.

[65] K.J. Blinowska, L.T. Czerwosz, W. Drabik, P.J. Franaszczuk, and H. Ekiert. EEG data reduction by means of autoregressive representation and discriminant analysis procedures. *Electroencephalography and clinical Neurophysiology*, 51(6):650–658, 1981.

[66] L. Marple. Resolution of conventional fourier, autoregressive, and special arma methods of spectrum analysis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, volume 2, pages 74–77, May 1977.

[67] Martin Spüler. *Assessing the Benefit of Adaptive Brain-Computer Interfacing*. PhD thesis, Eberhard Karls Universität Tübingen, 2012.

[68] M.B. Priestley. *Spectral analysis and time series.* Academic Press, London, 1981.

[69] G. Pfurtscheller and F.H. Lopes da Silva. Event-related desynchronization (ERD) and event-related synchronization (ERS). In E. Niedermeyer and F. H. Lopes da Silva, editors, *Electroencephalography: Basic Principles, Clinical Applications and Related Field*, pages 1003–1016. Lippincott, Williams and Wilkins, Philadelphia, PA, fifth edition, 2005.

[70] Francisco Varela, Jean-Philippe Lachaux, Eugenio Rodriguez, and Jacques Martinerie. The brainweb: phase synchronization and large-scale integration. *Nature reviews neuroscience*, 2(4):229–239, 2001.

[71] Giulio Tononi and Gerald M Edelman. Consciousness and complexity. *science*, 282(5395):1846–1851, 1998.

[72] Karl J. Friston. Beyond phrenology: what can neuroimaging tell us about distributed circuitry? *Annual review of neuroscience*, 25(1):221–250, 2002.

[73] F. Kalaska, John and J. Crammond, Donald. Cerebral cortical mechanisms of reaching movements. *Science*, 255(5051):1517–1523, 1992.

[74] V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine*, 2011. doi:10.1016/j.compbiomed.2011.06.020.

[75] Karl J. Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78, 1994.

[76] Paul L. Nunez and Ramesh Srinivasan. *Electric Fields of the Brain*. Oxford University Press, 2006.

[77] Katarzyna J Blinowska, Rafał Kuś, and Maciej Kamiński. Granger causality and information flow in multivariate processes. *Physical Review E*, 70(5):050902, 2004.

[78] MJ Kaminski and KJ Blinowska. A new method of the description of the information flow in the brain structures. *Biological cybernetics*, 65(3):203–210, 1991.

[79] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, Terrence J Sejnowski, et al. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, pages 145–151, 1996.

[80] Terrence D. Lagerlund and Gregory A. Worrell. *Electroencephalographic: Basic PPrincipal, Clinical Applicaltions, and Related Fields*, chapter EEG Source Localization (Model-Dependent and Model Independent Methods), pages 829–844. Lipponcott Williams & Wilkins, 5 edition, 2005.

[81] Roberto D Pascual-Marqui, Christoph M Michel, and Dietrich Lehmann. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, 18(1):49–65, 1994.

[82] Thorsten O Zander, Christian Kothe, Sabine Jatzev, and Matti Gaertner. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. In *Brain-computer interfaces*, pages 181–199. Springer, 2010.

[83] Thorsten O. Zander and Christian Kothe. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *J Neural Eng*, 8(2):025005, Apr 2011.

[84] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.

[85] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[86] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[87] Fabien Lotte, Marco Congedo, Anatole Lécuyer, and Fabrice Lamarche. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, 4, 2007.

[88] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Report Microsoft Research-TR-98-14*, 1998.

[89] Chang Chih-Chung and Lin Chih-Jen. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[90] Pai-Hsuen Chen, Rong-En Fan, and Chih-Jen Lin. A study on smo-type decomposition methods for support vector machines. *Neural Networks, IEEE Transactions on*, 17(4):893–908, 2006.

[91] A.W. Kuhn and H.W. Tucker. Aw (1951) nonlinear programming. In *2nd Berkeley Symposium. Berkeley, University of California Press*, 1951.

[92] George A.F. Seber and Alan J. Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

[93] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[94] Keith E Muller and Paul W Stewart. *Linear model theory: univariate, multivariate, and mixed models*. John Wiley & Sons, 2006.

[95] Hesham Sheikh, Dennis J. McFarland, William A. Sarnacki, and Jonathan R. Wolpaw. Electroencephalographic(EEG)-based communication: EEG control versus system performance in humans. *Neuroscience Letters*, 345(2):89–92, Jul 2003.

[96] M. Spüler, W. Rosenstiel, and M. Bogdan. A fast feature selection method for high-dimensional meg bci data. In *In Proceedings of the 5th International Brain-Computer Interface Conference.*, pages 24–27, Graz, Austria,, 2012.

[97] Martin Spüler, Andrea Sarasola, Niels Birbaumer, Wolfgang Rosenstiel, and Ander Ramos-Murguialday. Comparing metrics to evaluate performance of regression methods for decoding of neural signals, 2015.

[98] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[99] Carina Walter, Stephanie Schmidt, Wolfgang Rosenstiel, Peter Gerjets, and Martin Bogdan. Using cross-task classification for classifying workload levels in complex learning tasks. In *Affective Computing and Intelligent Interaction (ACII), 2013*, pages 876–881. IEEE, 2013.

[100] C. Walter, P. Wolter, W. Rosenstiel, M. Bogdan, and M. Spüler. Towards Cross-Subject Workload Prediction. In *Proceedings of the 6th International Brain-Computer Interface Conference*, Graz, Austria, 09 2014.

[101] Michel Besserve, Karim Jerbi, Francois Laurent, Sylvain Baillet, Jacques Martinerie, and Line Garnero. Classification methods for ongoing EEG and MEG signals. *Biol Res*, 40(4):415–437, 2007.

[102] Nick F. Ramsey, Martijn P. van de Heuvel, Kuan H. Kho, and Frans S. Leijten. Towards human BCI applications based on cognitive brain systems: an investigation of neural signals recorded from the dorsolateral prefrontal cortex. *IEEE Trans Neural Syst Rehabil Eng*, 14(2):214–217, Jun 2006.

[103] M. Dyson, F. Sepulveda, and J. Q. Gan. Localisation of cognitive tasks used in EEG-based BCIs. *Clin Neurophysiol*, 121(9):1481–1493, Sep 2010.

[104] Anne-Marie Brouwer, Maarten A Hogervorst, Jan BF Van Erp, Tobias Heffelaar, Patrick H Zimmerman, and Robert Oostenveld. Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[105] C. Berka, D.J. Levendowski, and M.M. Cvetinovic. Real-Time Analysis of EEG Indexes of Alertness, Cognition and Memory Acquired With a Wireless EEG Headset. *International Journal of Human-Computer Interaction*, 17:151–170, 2004.

[106] I. Chaouachi, M. Jraidi and C. Frasson. Modeling Mental Workload Using EEG Features for Intelligent Systems. In J.A. Konstan, editor, *UMAP2011*, volume 6787, pages 50–61. LnCS, Springer, Heidelberg, 2011.

[107] F. Putze, J. Jarvis, and T. Schultz. Multimodal Recognition of Cognitive Workload for Multitasking in the Car. In *Proceeding of the International Conference on Pattern Recognition 2010*, 2010.

[108] Dongrui Wu, Brent J Lance, and Thomas D Parsons. Collaborative filtering for brain-computer interaction using transfer learning and active class selection. *PloS one*, 8(2):e56624, 2013.

[109] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[110] C.J Baldwin and B.N. Penaranda. Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage*, 59:48–56, 2012.

[111] Alan Gevins, Michael E Smith, Harrison Leong, Linda McEvoy, Susan Whitfield, Robert Du, and Georgia Rush. Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(1):79–91, 1998.

[112] Z. Wang, R. M. Hope, Z. Wang, Q. Ji, and W. D. Gray. Cross-subject workload classification with a hierarchical bayes model. *NeuroImage*, 59(1):64–69, 2012.

[113] Jing Jin, Eric W Sellers, Yu Zhang, Ian Daly, Xingyu Wang, and Andrzej Cichocki. Whether generic model works for rapid erp-based bci calibration. *Journal of neuroscience methods*, 212(1):94–99, 2013.

[114] Matthias Krauledat, Michael Tangermann, Benjamin Blankertz, and Klaus-Robert Müller. Towards zero training for brain-computer interfacing. *PloS one*, 3(8):e2967, 2008.

[115] J Mostow. Collaborative research on learning technologies: An automated reading assistant that listens. In *Proceedings of the national science foundation interactive systems grantees workshop*, pages 412–440, 1997.

[116] Federico C. Galán and Carole R. Beal. EEG estimates of engagement and cognitive workload predict math problem solving outcomes. In *User Modeling, Adaptation, and Personalization*, pages 51–62. Springer, 2012.

[117] Jack Mostow, Kai-Min Chang, and Jessica Nelson. Toward exploiting EEG input in a reading tutor. In *Artificial Intelligence in Education*, pages 230–237. Springer, 2011.

[118] Kai-min Chang, Jessica Nelson, Udip Pant, and Jack Mostow. Toward Exploiting EEG Input in a Reading Tutor. *International Journal of Artificial Intelligence in Education*, 22(1):19–38, 2013.

[119] Jack Mostow and Joseph Beck. When the rubber meets the road: Lessons from the in-school adventures of an automated Reading Tutor that listens. *Scale-Up in Education*, 2:183–200, 2007.

[120] Chris Berka, Daniel J Levendowski, Caitlin K Ramsey, Gene Davis, Michelle N Lumicao, Kay Stanney, Leah Reeves, Susan H Regli, Patrice D Tremoulet, and Kathleen Stibler. Evaluation of an EEG workload model in an Aegis simulation environment. In *Defense and security*, pages 90–99. International Society for Optics and Photonics, 2005.

[121] Ronald H Stevens, Trysha Galloway, and Chris Berka. EEG-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills. In *User Modeling 2007*, pages 187–196. Springer, 2007.

[122] Michael C Dorneich, Santosh Mathan, J Creaser, Stephen D Whitlow, and Patricia May Ververs. Enabling improved performance though a closed-loop adaptive system driven by real-time assessment of cognitive state. In *Proceedings of the 11th International Conference on Human-Computer Interaction (Augmented Cognition International)*, pages 22–27, 2005.

[123] Lawrence J. Prinzel, Frederick G. Freeman, Mark W. Scerbo, Peter J. Mikulka, and Alan T. Pope. A closed-loop system for examining psychophysiological measures for adaptive task allocation. *The International journal of aviation psychology*, 10(4):393–410, 2000.

[124] Emmanuel G Blanchard, Boris Volfson, Yuan-Jin Hong, and Susanne P Lajoie. Affective artificial intelligence in education: From detection to adaptation. In *AIED*, volume 2009, pages 81–88, 2009.

[125] Daniel Szafir and Bilge Mutlu. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20. ACM, 2012.

[126] Daniel Szafir and Bilge Mutlu. Artful: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1001–1010. ACM, 2013.

[127] C. Walter, G. Cierniak, P. Gerjets, W. Rosenstiel, and M. Bogdan. Classifying mental states with machine learning algorithms using alpha activity decline. In *Proceed-*

ings of the 19th Europeen Symposium of Neuronal Networks (ESANN 2011), pages 405–410, 4 2011.

[128] R. Schwonke, J. Wittwer, V. Aleven, R. Salden, C. Krieg, and A. Renkl. Can tutored problem solving benefit from faded worked-out examples? In S. Vosniadou, D. Kayser, and A. Protopapas, editors, *Proceedings of the 2nd European Cognitive Science Conference*, pages 59–64. Erlbaum, New York, 2007.

[129] Robert F Schmidt, Florian Lang, and Manfred Heckmann. *Physiologie des Menschen: mit Pathophysiologie*, volume 31. Springer Heidelberg Berlin, 2010.

[130] Per B Sederberg, Michael J Kahana, Marc W Howard, Elizabeth J Donner, and Joseph R Madsen. Theta and gamma oscillations during encoding predict subsequent recall. *The Journal of Neuroscience*, 23(34):10809–10814, 2003.

[131] G Zeller and D Bente. Veränderungen der hirnelektrischen Organisation bei visuellen Such-und Diskriminationsprozessen unterschiedlichen Schwierigkeitsgrades. *EEG-EMG*, 14(4):177–185, 1983.

[132] C. Walter, G. Cierniak, W. Rosenstiel, M. Bogdan, T.O. Zander, and P. Gerjets. Using Passive Brain-Computer Interfaces for cognitive workload assessment. Poster presentation at the 48. Kongress der Deutschen Gesellschaft für Psychologie, Bielefeld, 2012.

[133] C. Walter, A.-A. Pape, W. Rosenstiel, P.; Gerjets, and M. Bogdan. Detecting Working Memory Load from EEG-Data during learning and solving complex tasks. In *BBCI Workshop 2012 on Advances in Neurotechnology*, Berlin, Germany., 09 2012.

[134] J. Schuh. Computerbasierte Vermittlung transferierbaren Fertigkeitswissens zur Lösung mathematischer Textaufgaben. Wissensprozesse und digitale Medien. Logos-Verl., 2006. Wissensprozesse und digitale Medien. Logos-Verl.

[135] D.J. McFarland, L.M. McCane, S.V. David, and J.R. Wolpaw. Spatial filter selection for EEG-based communication. *Electroencephalography and Clinical Neurophysiology*, 103:386–394, 1997.

[136] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, Dec 2000.

[137] Scott Makeig, Tzyy-Ping Jung, Anthony J Bell, Dara Ghahremani, and Terrence J Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, 94(20):10979–10984, 1997.

[138] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computatinal Intelligence and Neuroscience*, 2011:156869, 2011.

[139] T.M. Cover and J.A. Thomas. *Elements of information theory*. Hoboken,NJ: Wiley-Interscience, 2006.

[140] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[141] T.F. Oostendorp and A. van Oosterom. Source parameter estimation in inhomogeneous volume conductors of arbitrary shape. *IEEE Trans Biomed Eng*, 36:382–391, 1989.

[142] R. Kavanagh, T.M. Darcey, D. Lehmann, and D.H. Fender. Evaluation of methods for three-dimensional localization of electric sources in the human brain. *IEEE Trans Biomed Eng*, 25:421–429, 1978.

[143] C.-T. Lin, R.-C. Wu, S.-F. Liang, W.-H. Chao, Chen Y.-J., and T.-P. Jung. EEG-based drowsiness estimation for safety driving using independent component analysis. *IEEE Trans. Circuits Syst. I, Reg. Papers*, page 2726–2738, 2005.

[144] W. Klimesch, M. Doppelmayr, and S. Hanslmayr. Upper alpha erd and absolute power: their meaning for memory performance. *Progress in Brain Research*, 6:151–165, 2006.

[145] Roland H. Grabner and Bert De Smedt. Neurophysiological evidence for the validity of verbal strategy reports in mental arithmetic. *Biol Psychol*, 87(1):128–136, Apr 2011.

[146] C. Walter, S. Schmidt, W. Rosenstiel, M. Bogdan, and P. Gerjets. Alpha- and theta frequencies as indicators for optimal cognitive load during learning. In *6th International Cognitive Load Theory Conference.*, Toulouse, France., 06 2013.

[147] T. S. Redick, A. Calov, C. E. Gay, and R. W. Engle. Working memory capacity and go/no-go task performance: selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2):308–324, 2011.

[148] W. Schnotz, H. Horz, N. McElvany, S. Schroeder, M. Ullrich, and J. Baumer. Das BITEProjekt: Integrative Verarbeitung von Bildern und Texten in der Sekundarstufe, 2010.

[149] Katharina Scheiter, Peter Gerjets, and Julia Schuh. The acquisition of problem-solving skills in mathematics: How animations can aid understanding of structural problem features and solution procedures. *Instructional Science*, 38(5):487–502, 2010.

[150] Alois Schlögl, Claudia Keinrath, Doris Zimmermann, Reinhold Scherer, Robert Leeb, and Gert Pfurtscheller. A fully automated correction method of EOG artifacts in EEG recordings. *Clinical neurophysiology*, 118(1):98–104, 2007.

[151] Bert De Smedt, Roland H Grabner, and Bettina Studer. Oscillatory eeg correlates of arithmetic strategy use in addition and subtraction. *Experimental brain research*, 195(4):635–642, 2009.

[152] Piet MT Broersen. Automatic spectral analysis with time series models. *Instrumentation and Measurement, IEEE Transactions on*, 51(2):211–216, 2002.

[153] N. Penaranda and C.L. Baldwin. More can be less: Selecting EEG features for an artificial neural network-based workload classification. In *Applied Human Factors and Ergonomics Conference, Miami, FL*, 2010.

[154] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, pages 13–17. ACM, 2008.

[155] C. Walter, Scharinger C., W. Rosenstiel, P. Gerjets, and Bogdan M. Adaptive Learning Environments based on passive BCI Methodology. In *Passive BCI Community Meeting*, Delmenhorst, Germany. (Project Presentation), 2014.

[156] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning (synthesis lectures on artificial intelligence and machine learning). *Morgan and Claypool Publishers*, 2009.

[157] S. Lemm, B. Blankertz, T. Dickhaus, and K. Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56:387–399, 2011.

[158] E. Thomas, M. Dyson, and M. Clerc. An analysis of performance evaluation for motor-imagery based bci. *Journal of neural engineering*, 10(3), 2013.

[159] M. Spüler, W. Rosenstiel, and M. Bogdan. Principal component based covariate shift adaption to reduce non-stationarity in a MEG-based Brain-Computer Interface. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–7, 2012.

[160] Abdul Satti, Cuntai Guan, Damien Coyle, and Girijesh Prasad. A covariate shift minimisation method to alleviate non-stationarity effects for an adaptive brain-computer interface. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 105–108. IEEE, 2010.

[161] B. H. Kantowitz. *Human Factors Psychology*, chapter Mental Workload, pages 81–121. North-Holland Publishing Co., Amsterdam, The Netherlands., 1987.

[162] H. B. G. Thomas. Communication theory and the constellation hypothesis of calculation. *Quarterly Journal of Experimental Psychology*, 15(3):173–191, 1963.

[163] Wojciech Samek, Motoaki Kawanabe, and Carmen Vidaurre. *Group-wise stationary subspace analysis-a novel method for studying non-stationarities*. na, 2011.

[164] Trayambak Tiwari, Anju L Singh, and Indramani L Singh. Task demand and workload: Effects on vigilance performance and stress. *Journal of the Indian Academy of Applied Psychology*, 35(2):265–275, 2009.

[165] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(Supplement 1):B231–B244, 2007.

[166] Barry S Oken, Martin C Salinsky, and SM Elsas. Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical Neurophysiology*, 117(9):1885–1901, 2006.