# Chapter 8
# Brain-Computer Interfaces for Educational Applications

**Martin Spüler, Tanja Krumpe, Carina Walter, Christian Scharinger, Wolfgang Rosenstiel, and Peter Gerjets**

## Introduction

Currently, there is an ongoing debate in education research on how to optimally support learners during learning (Calder 2015; Askew 2015). Obviously, learning outcome is most promising if the training program and learning content is tailored to the learner's specific needs (e.g., Gerjets & Hesse 2004; Richards et al. (2007); for numerical interventions see Dowker 2004; Karagiannakis & Cooreman 2014). To optimally support the learner's efforts, it is important that the learning content is neither too easy nor too difficult. Therefore, it is crucial for successful learning to keep the cognitive workload in the individual optimal range for each learner (Gerjets, Scheiter, & Cierniak 2009; Sweller, Van Merrinboer, & Paas 1998). This can be achieved by adapting the difficulty of the training content to the individual competencies of the learner.

Computer-supported learning (Kirschner & Gerjets 2006) seems specifically suited for implementing adaptivity, because these informational environments can be easily extended with algorithms that change the difficulty of the presented material based on the learner's behavioral response. This allows for an easy personalization of the learning environment to the user's individual needs, which is assumed to be key for efficient learning. So far, adaptive computer-supported learning environments rely on the user's interaction behavior for adaptation, e.g.,

M. Spüler (✉) • T. Krumpe • C. Walter • W. Rosenstiel
Department of Computer Engineering, Eberhard-Karls University Tübingen, Tübingen, Germany
e-mail: spueler@informatik.uni-tuebingen.de

C. Scharinger • P. Gerjets
Knowledge Media Research Center, Tübingen, Germany

J. Buder, F.W. Hesse (eds.), *Informational Environments*,
DOI 10.1007/978-3-319-64274-1_8

the difficulty of the training content is adapted based on the number of correct responses (Corbett 2001; Graesser & McNamara 2010; Käser et al. 2013). These behavioral measures are rather indirect and distal measures that are not very specific with respect to inferring the cognitive processes required for performing the task at hand. For instance, more errors in a row may not only be caused by the difficulty of the task itself but also by task-unspecific processes (e.g., lapses of attention, fatigue, or disengagement).

As already outlined in Chap. 7 of this book (Soltanlou et al. 2017), there are several neurocognitive processes involved in learning, which can be found in recorded neurophysiological data. With the advent of Brain-Computer Interfaces (BCIs), which translate brain activity into control signals (Wolpaw, Birbaumer, McFarland, Pfurtscheller, & Vaughan 2002), there seems to be a new technology that could use these neurocognitive processes in the context of educational learning environments. For measuring brain activity, electroencephalography (EEG) is a widely used method for BCIs. The measured activity is then processed, filtered, and certain features are extracted out of the brain signals. For translation into control signals, different machine learning methods are used. Those methods need a certain amount of training data (e.g., preprocessed EEG data) and class labels (e.g., user thinks left or right). Each data point has to be assigned one of the two class labels. Based on the data and the corresponding labels, the algorithms find patterns in the data that belong to either of the classes. These patterns can be used to create a prediction model which allows to predict to which class a new data point belongs.

While traditional BCIs allow a user to communicate or control a computer by brain activity only (Spüler 2015), they are more recently also used to extract information about the user and infer his or her mental states (e.g., workload, vigilance).

As BCIs can be used to measure cognitive processes, this allows for a more direct and implicit monitoring of the learner's state and should thereby allow for a better adaptation of the training content to improve learning success of the user. That the amount of cognitive workload can be measured by EEG has been shown by multiple studies (Gevins, Smith, McEvoy, & Yu 1997; Murata 2005; Berka et al. 2007; Wang, Hope, Wang, Ji, & Gray 2012). While workload is reflected in the amplitude of event-related potentials in the EEG (Brouwer et al. 2012; Causse, Fabre, Giraudet, Gonzalez, & Peysakhovich 2015), it also affects the oscillatory activity observed in an EEG (Kohlmorgen et al. 2007; Brouwer et al. 2012). Specific to arithmetic tasks, it was shown that the cognitive demand results in a power increase in the theta band and a decrease in the alpha band (Harmony et al. 1999). As there are known effects of workload in the EEG, these can potentially be detected by a BCI and used for adaptation of a learning environment.

In the course of this chapter, we will give an overview of the progress made in the last 7 years, while trying to develop an EEG-based learning environment. Based on the lessons learned from the first studies for workload assessment, we present results that show that EEG can be used to assess workload during arithmetic learning. By developing a BCI that detects workload in real-time, we could show that this approach can also be used for content adaptation in a learning environment. While this system only detected workload as one component, we could also show that different sub-components of workload can be differentiated and thereby be used to gain more detailed information about the learner's mental state. Finally, we discuss these findings, show open problems, and present ideas how this line of research can be extended in the following years.

## The First Steps Are the Hardest: What We Have Learned in the Early Days

When we started the project to develop an EEG-based learning environment, we performed two studies in which we addressed the workload assessment during learning of realistic instructional material and the use of cross-task workload assessment to assess workload in a variety of different tasks. While the results of these studies were mixed, they uncovered several issues in designing an EEG-based learning environment. Here we briefly describe these studies and summarize the lessons we have learned to avoid those issues in the future.

### *Workload Detection During Studying Realistic Material*

In our first study (Walter, Cierniak, Gerjets, Rosenstiel, & Bogdan 2011), we analyzed workload in ten subjects (age 12–14 years) who studied instructional material while EEG was recorded. There were two types of instructional material, which involved processes of learning and comprehension with different levels of workload. The material inducing a high workload consisted of graphical representations and explanations of mathematical angle theorems, while the low workload material consisted of comic strips in which the subjects had to understand the story. While both materials contained a complex graphical display, the comic strips were easy to understand, while the angle theorems were harder and required more effort.

An analysis of the EEG data showed that spectral power in the alpha band (8–13 Hz) over the occipital area showed a strong desynchronisation (decrease of power) for high-workload material. Using a support vector machine (SVM), we tried to classify the data into categories of high and low workload on a single-trial level and obtained an average accuracy of 76%, which shows that EEG can be used to detect workload.

Although the results are promising for developing an EEG-based learning environment, they have to be interpreted with caution as some methodological critique can be applied to that study regarding perceptual-motor confounds. As comic strips and angle theorems differed largely in their appearance, this might lead to differences in perceptual processing beyond imposing different levels of workload. The different perception might also lead to different motor behavior (e.g., different eye movement patterns), which could lead to different motor-related brain activity. Although the results are in line with the literature, some doubts remain if the effects are solely caused by workload modulation.

## *Cross-Task Workload Prediction*

While the previous results have shown that machine learning methods can be used to detect workload from EEG data, another problem becomes apparent when trying to use this approach in a learning environment. For the application of machine learning, training data are needed, which also includes labels that assign each datapoint to one class (such as easy or hard). In the previous study, data from the same task was used. In a learning environment this approach can not be used, due to learning effects of the subject. These learning effects (task is hard at the beginning, easy after learning) change the labels of the data, which makes it difficult to apply machine learning approaches.

As an alternative approach to calibrate an EEG-based learning environment, we wanted to use basic psychological tasks (n-back, go-nogo, reading span) that induce workload and use data from these tasks to train a classifier, which can afterwards be used in a realistic learning scenario. Therefore, we recorded EEG data from 21 subjects who performed the three basic psychological tasks for workload induction and two realistic tasks. The two realistic tasks were comprised of algebra and arithmetic tasks with different levels of difficulty. The tasks were designed to avoid perceptual motor confounds. More details can be found in (Walter, Schmidt, Rosenstiel, Gerjets, & Bogdan 2013).

When trying to classify the data, we found that a within-task classification (by cross-validation) works very well for the three basic tasks with an average accuracy >95%, while accuracies of 73% were reached for the arithmetic task and 67% for the algebra task. For the cross-task classification, in which the classifier was trained on data from one task, and tested on another task, we reached accuracies around 50% (chance level) regardless of the combination of the tasks. These results show that the EEG data contains workload-related activity that can be detected by machine learning methods within a task, but not across different tasks. To identify reasons why the cross-task classification failed, we also investigated the labels. We used the objective difficulty, which ranges from 1 (e.g., 0-back) to 3 (e.g., 2-back). This approach assumes that a 2-back task is as difficult as an arithmetic task of level 3. As this is a bold assumption that may not necessarily be true, we also investigated the subjective labels, as each subject rated the subjective difficulty of each trial. Unfortunately, the subjective ratings could not be used for all subjects, as some subjects had a strong bias in their response. However, for the remaining subjects, the cross-task classification clearly improved. When using two basic tasks for training, and testing on either the algebra or the arithmetic task, classification worked significantly above chance level for 16 out of 27 tests with accuracies $\geq$70%.

## *Pitfalls in Designing EEG-Based Learning Environments*

As we have discussed in a previous paper (Gerjets, Walter, Rosenstiel, Bogdan, & Zander 2014), the results from the first two studies led to some insight, which can be summarized as four lessons that we have learned.

The first lesson was to avoid perceptual-motor confounds. If there are possible perceptual-motor confounds it is difficult to tell if classification occurs solely on the basis of workload-related changes in the brain activity or if the classifier is unintentionally trained to detect perceptual- or motor-related brain activity. By using simple material, which rules out any perceptual-motor confounds, one can ensure that classification occurs solely on workload-related brain responses. However, this raises another question, how well a classifier trained on workload data obtained from a simple task is able to predict workload in a realistic and complex scenario, where perceptual-motor confounds can not be eliminated.

The second lesson is about the task order in the context of learning. Randomizing the task order with regard to task difficulty is usually a good choice to avoid confounds. However, when it comes to learning, this is hard to implement due to learning effects. A task with medium difficulty might induce high workload in the beginning. With time, the subject learns how to perform the task and a medium difficulty will induce low workload after the subject has mastered the task. This shows not only that the labeling of the data during learning is problematic, but also that data from the same task and subject should not be used for prediction in a learning environment.

Lesson three is about using different tasks. As a cross-task classification avoids some problems (lesson 2), it has some different issues. Although perceptual-motor confounds should be avoided, there is always task-related brain activity present, which might be different when performing different tasks. Thereby, for a cross-task classification to work, it is advisable to use many different tasks. Thereby the classifier can be trained on the overall workload pattern, without overfitting to task-specific activity patterns.

Lesson four is also related to the labeling. As we have seen that a cross-task classification using multiple tasks for training is advisable, the questions of labeling arise. As the difficulty of a task is hard to compare between two tasks, and might also be perceived differently by different subjects, an objective label might not be the best choice. It would be better to use subjective labeling to assess the perceived difficulty of the task. While this might lead to a better quality of the data, it is more difficult to work with, as some subjects tend to give highly biased ratings (leading to imbalanced data) and additional time is needed to acquire the subject rating.

## EEG-Based Workload Assessment During Arithmetic Learning

As a next step towards the development of an EEG-based learning environment, we performed a study in which the subjects performed arithmetic addition tasks. The EEG data collected during that experiment was then used to train a classifier to predict the difficulty of the task, which we used as measure for the expected mental workload.

### *Study Design*

Ten students (six female; mean age: $24.9 \pm 5.3$ years, 17–32) participated voluntarily in this study and received monetary compensation for participation. All participants reported normal or corrected-to-normal vision and no mathematical problems. Participants were chosen randomly. Nevertheless, all were university students (with different fields of study) and can thus be considered as having a high educational background.

The tasks consisted of simple arithmetic exercises in the form $x + y =?$, where the addends ranged from single-digit to four-digit numbers. Difficulty of a task was assessed with the $Q$-value (Thomas 1963), which ranged from 04 (e.g. 1+1) to 7.2 (addition of two four-digit numbers). The main parameters for problem difficulty in addition are the size of the summands involved (i.e., problem size Stanescu-Cosson et al. 2000) and whether or not a carry operation is needed (Kong et al. 2005). As $Q$-value takes into account both problem size and the need for a carry operation it is a more comprehensive measure of task difficulty as compared to using only one parameter such as problem size. Moreover, the $Q$-value also considers additional aspects, which should affect task difficulty (e.g., the reduced difficulty of specific problems such as $1000 + 1000$). Therefore, the $Q$-value is an adequate measure of task difficulty. In Spüler et al. (2016) more detail on how we calculated the $Q$-value is given.

Each participant had to solve 240 addition problems while EEG was measured. Due to possible learning effects, the tasks were presented with increasing difficulty. The time course of the experiment is illustrated in Fig. 8.1. The calculation task was shown for 5 s, in which the participants could solve the addition problem. After the 5 s, presentation of the task disappeared and the subjects had up to 3.5 s to enter their answer. The activation interval during the calculation phase was later used for analysis of the EEG signal. By separating the response from the calculation phase, we tried to avoid motor confounds in the EEG data during the calculation phase.

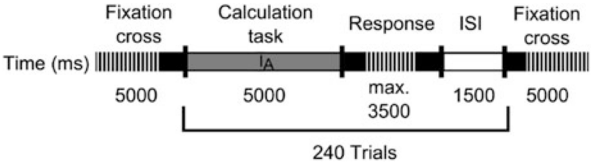

**Fig. 8.1** Schematic illustration of the time course of the experiment. The *grey area* indicates the considered activation interval ($I_A$), whereas the *white area* represents response and inter-stimulus interval (ISI). At the beginning and the end of the experiment, a fixation cross was displayed

## EEG Recording

A set of 28 active electrodes (actiCap, BrainProducts GmbH) attached to the scalp placed according to the extended International Electrode 10–20 Placement System (FPz, AFz, 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 used to record an electrooculogram (EOG); two of them were placed horizontally at the outer canthus of either the left or the right eye, respectively, to measure horizontal eye movements, and one was placed in the middle of the forehead between the eyes to measure vertical eye movements. Ground and reference electrodes were placed on the left and right mastoids. EOG- and EEG-signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp, g.tec) and sampled at a rate of 512 Hz. EEG data was high-pass filtered at 0.5 Hz and low-pass filtered at 60 Hz during the recording. Furthermore a notch-filter was applied at 50 Hz to filter out power line noise.

## Analysis of EEG Data Regarding Workload-Related Effects

To remove the influence of eye movements we applied an EOG-based regression method (Schlögl et al. 2007). For analyzing the data we estimated the power spectrum during the calculation phase of each trial. We used Burg's maximum entropy method (Cover & Thomas 2006) with a model order of 32 to estimate the power spectrum from 1 to 40 Hz in 1 Hz bins.

As an indication of workload-related effects in the power spectrum we calculated $R^2$-values (squared Pearson's correlation coefficient) between the $Q$-value and the power at each electrode and frequency bin. In Fig. 8.2, the results for different frequency bands are shown. While workload-related changes can be observed in the delta and lower beta frequency band, changes are strongest in the theta and alpha band. Location of activation changes slightly between theta and alpha, with theta having a more centro-parietal distribution and alpha being stronger over the parieto-occipital region.

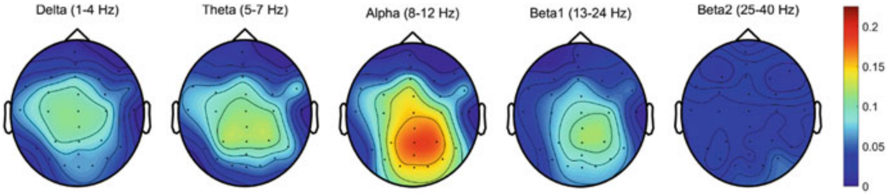

**Fig. 8.2** Topographic display of $R^2$ values averaged over all participants, showing the influence of the $Q$-value for each electrode in different frequency bands. $R^2$ color coded with increasing *red reflecting* higher $R^2$ values

## *Prediction of Task Difficulty*

To predict the task difficulty based on the EEG data, we processed the data similarly to that for the analysis in the previous section. In an earlier attempt to predict workload using this data (Walter, Wolter, Rosenstiel, Bogdan, & Spüler 2014), which is not presented in this section, we did not use an EOG-based artifact correction and found that the prediction performance is largely influenced by eye-movement artifacts. Thus, here we used an EOG-based regression method (Schlögl et al. 2007) to eliminate eye-movement artifacts from the data. We again used Burg's maximum entropy method (Cover & Thomas 2006) with a model order of 32 to estimate the power spectrum from 1 to 40 Hz in 1 Hz bins.

To correct for inter-participant variability in the participant's baseline EEG power, the first 30 trials (easy trials, with $Q < 1$) were used for normalization. The mean and 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 (either for training or for testing the model). Based on the normalized data, we trained a linear ridge regression model (regularization parameter $\lambda = 10^3$ determined by cross-validation on the training data). To reduce the number of features we used only every second frequency bin (1 Hz, 3 Hz, 5 Hz,...) to train the model. When not specified otherwise, we used the 17 inner electrodes (F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, POz, O1, Oz, O2) resulting in a total of 340 features used for prediction. However, other electrode subsets were also tested.

The prediction was tested in two different approaches. For the first approach, the within-subject prediction, data from the same subject were used for training and testing the prediction model. To ensure that data do not overlap, a 10-fold cross-validation was performed. The within-subject prediction was done only to get an estimate of the data quality and how well task difficulty can be predicted in an optimal scenario. The second approach, the cross-subject prediction, used data from nine subjects for training the model and was tested on the one remaining subject. The procedure was repeated so that data from each subject was used for testing once. With regard to a future application of the method in an online learning environment, the within-subject approach is less interesting, because it would mean that the learning environment needs to be calibrated to the user, which means that the user has to use the system for a while to collect a certain amount of data needed for training. The cross-subject prediction, on the other hand, allows the use of data from other subjects for training a prediction model. Using this approach in a learning environment would enable the subject to start using the system right away without the need for the collection of training data.

For evaluation of the prediction model, we used different metrics: correlation coefficient (CC) and root-mean-squared-error (RMSE). As both metrics assess different kinds of prediction error, it is important to look at both metrics (Spüler, Sarasola-Sanz, Birbaumer, Rosenstiel, & Ramos-Murguialday 2015).

As a regression predicts the $Q$-value, we assume that a precise prediction of task difficulty will not be necessary in an adaptive learning environment. It is more important to identify when the task gets too easy or too difficult to keep the learner in his optimal range of cognitive workload, which is why we also used the output of the regression model for a classification into three difficulty levels. For $Q < 2$ the trial was considered easy, for $Q > 4$ it was considered too difficult, while it had a medium difficulty level when $Q$ was between 2 and 4. These ranges were estimated based on the relationship between $Q$ and the behavioral task performance. To estimate the validity of our classification procedure, we applied this thresholding to the actual $Q$ and the predicted $Q$ and evaluated the classification accuracy (CA), which is given as the percentage of trials correctly classified.

## *Performance of Workload Prediction*

The cross-subject prediction performance averaged over all subjects reached an average CC of 0.74, RMSE of 1.726, and a CA of 49.8% (chance level would be 33%). Regarding the classification accuracy, trials of normal difficulty ($2 < Q < 4$), were correctly classified with an accuracy of 74.2%, while it was difficult to differentiate easy 4.8% accuracy) and difficult (30.4%) trials. A closer look at the prediction revealed that trials with $Q > 6$ were mostly classified as easy, indicating that easy trials and trials with $Q > 6$ share similar patterns in EEG. To improve performance we used only trials with $Q < 6$ for training and testing the model, which resulted in a significantly better CC of 0.82 ($p < 0.01$, t-test), RMSE of 1.343 ($p < 0.001$, t-test), and an improved classification accuracy of 55.8% ($p < 0.05$, t-test) (Table 8.1).

As it would be beneficial for an EEG-based learning environment to use as few EEG electrodes as possible, we also investigated how performance changes if the number of electrodes is reduced. The results with different electrodes are summarized in Table 8.1. When using only one electrode (besides the ground and reference electrode), POz was the position that delivered best results. While using POz performs significantly worse than using 17 electrodes in terms of CA ($p < 0.05$, t-test), there was no significant difference in CC and RMSE ($p > 0.05$). There

**Table 8.1** Prediction performance averaged over all participants for different electrode subsets using only trials with $Q < 6$

| Number | Description | Positions | CC | RMSE | CA |
|---|---|---|---|---|---|
| 28 | All | See section "EEG Recording" | 0.807 | 1.292 | 55.6% |
| 17 | Inner | See section "Prediction of Task Difficulty" | 0.820 | 1.343 | 55.8% |
| 7 | Central | Fz,FCz,Cz,CPz,Pz,POz,Oz | 0.839 | 1.352 | 54.7% |
| 9 | Frontocentral | F3,Fz,F4,FC3,FCz,FC4,C3,Cz,C4 | 0.725 | 1.465 | 45.5% |
| 7 | Parietooccipital | P3,Pz,P4,POz,O3,Oz,O4 | 0.831 | 1.383 | 54.1% |
| 1 | Best electrode | POz | 0.788 | 1.479 | 50.5% |

In each line one subset is described by the number of electrodes, a short description referring to the approximate location of the electrodes as well the position of all electrodes in the subset according to the 10–20 system. As performance metric we showed the correlation coefficient (CC), the root mean squared error (RMSE), and the average classification accuracy (CA)

was also no significant difference in prediction performance between using the 17 inner channels, all 28 channels, or a subset of frontocentral, central, or parietocentral electrodes ($p > 0.05$).

The within-subject prediction achieved an average CC of 0.90 and an RMSE of 0.95, which is significantly better for CC (<0.01) and RMSE ($p < 0.001$), but, as discussed earlier, the reduced calibration and training effort in a cross-subject prediction seems to outweigh the performance increase of the within-subject prediction.

## A Passive BCI for Online Adaptation of a Learning Environment

As we have shown in the previous section, that it is possible to predict workload during solving mathematical exercises by using a cross-subject regression model, we wanted to apply this model in an online study. This study should serve as a proof-of-concept that a digital learning environment can adapt its content in real-time based on the EEG of the user. To evaluate the effect of the learning environment, we assessed the learning success of the subjects and compared it to a control group.

### *Study Design and Participants*

The participants in this study were divided into two groups: an experimental group using the EEG-based learning environment and a control group using an error-adaptive learning environment, which can be considered as state-of-the-art. In both groups, the subjects learned arithmetic addition in the octal number system (e.g., $5 + 3 = 10$), which was a completely new task to all subjects.

To evaluate the learning success, each subject did a pre-test and a post-test, before and after using the learning environment for approximately 45 min (180 exercises). The tests consisted of 11 exercises, with varying difficulty. Although difficulty of the exercises were the same for the pre- and post-tests, the exercises themself were different.

The participants of both groups were university students of various disciplines, reported to have normal or corrected-to-normal vision, and participated voluntarily in the EEG experiment. Thirteen subjects (7 male and 6 female; mean age: $28.1 \pm 4.3$ years, 21–35) participated in the experimental group using the EEG-based learning environment. The control group consisted of 11 subjects (7 male and 4 female; mean age: $23.4 \pm 1.4$ years, 22–27), using an error-adaptive learning environment.

## *Cross-Subject Regression for Online Workload Prediction*

Based on the EEG data obtained from the previous study (see section "EEG-Based Workload Assessment During Arithmetic Learning"), we created a prediction model that was used to predict the cognitive workload online and further was able to adjust the learning environment accordingly. In terms of usability, we wanted the EEG-based learning environment to be useable out-of-the-box without the need for a subject-specific calibration phase, which is why we used the cross-subject regression method.

Therefore, EEG data from the previous study were used for training a linear ridge regression model with a regularization parameter of $\lambda = 10^3$. 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). Furthermore, only trials with a $Q$-value smaller than 6 were used to train the regression model, since trials with a higher $Q$-value showed similar EEG patterns as very easy trials, which is most likely due to a disengagement of the subjects. While the trained regression model allows workload prediction on a single-trial level, the accuracy of the prediction was improved by averaging the prediction output using a moving average window of the last six trials. The moving average leads to a more robust prediction, but also makes the system react slower to sudden changes in workload, which is feasible since it is not recommended to adapt the difficulty of an online learning environment too rapidly.

The prediction model thus trained was then applied in the EEG-based learning environment, to predict the amount of cognitive workload for novice subjects in real-time.

### *Real-Time Adaptation of the Learning Environment*

For the experimental group, the EEG data served as a workload indicator. Therefore, we used the output of the previously trained regression model to predict the current workload state of each learner and differentiated three difficulty levels. 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 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 relationship between error rate and $Q$-value (Spüler et al. 2016). Trials with $Q < 1$ were solved correctly in all cases, while none of the subjects were able to solve trials with a $Q > 6$. 50% of the trials with a $Q = 3.5$ were successfully solved on average.

For the control group, an error-adaptive learning environment was used. The number of wrong answers served as performance and adaptation measure. When subjects solved five consecutive tasks correctly, the difficulty level increased by 1. On the other hand, the difficulty level decreased by a factor of 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 was changed, the presented $Q$-values increased or decreased by steps of size 1 and the calculated $Q$-values were rounded to the next integer. The learning session for the experimental group as well as for the control group started with an exercise of difficulty level $Q = 2$.

To compare the learning success of the two groups, the learning effect after completing the learning phase serves as a performance measure and is used as an indicator of how successfully each subject was supported during learning. Hence, each subject 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, and the difference in score between the two tests served as an indicator of the learning effect.

### *Task Performance Results*

The behavioral results, how well the subjects performed the octal arithmetic task, are shown in Table 8.2 for the experimental group and in Table 8.3 for the control group. For the experimental group, 45.5% of the 180 exercises were solved correctly on average. Averaged over all subjects, a maximum $Q$-value of 5.85 was reached by using the EEG-based learning environment. Each subject achieved at least the difficulty level of $Q = 3.2$ (see Table 8.2).

**Table 8.2** Task performance for the experimental group

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Correct | 38.3 | 88.9 | 60.6 | 10.0 | 28.9 | 25.6 | 31.1 | 17.8 | 66.1 | 33.3 | 16.7 | 92.2 | 82.2 | 45.5 |
| 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 |

The relative amount of correctly solved trials in % and the maximum $Q$-value for each subject using the EEG-based learning environment is shown

**Table 8.3** Task performance for the control group

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Correct | 64.4 | 62.2 | 57.2 | 66.7 | 63.3 | 61.1 | 63.3 | 67.8 | 65.0 | 66.1 | 66.7 | 64.0 |
| max. $Q$ | 6 | 4 | 5 | 6 | 4 | 5 | 4 | 4 | 4 | 5 | 4 | 4.64 |

The relative amount of correctly solved trials in % and the maximum $Q$-value for each subject using the error-adaptive learning environment is shown

**Table 8.4** 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

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pre-Test | 3 | 0 | 6 | 0 | 2 | 1 | 0 | 1 | 0 | 5 | 1 | 1 | 0 | 1.54 |
| Post-Test | 7 | 0 | 9 | 8 | 2 | 3 | 3 | 2 | 8 | 7 | 7 | 2 | 8 | 5.08 |
| Learned | 4 | 0 | 3 | 8 | 0 | 2 | 3 | 1 | 8 | 2 | 6 | 1 | 8 | 3.54 |

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 8.3). The best subjects reached a maximum $Q$-value of 6, whereas each participant achieved at least the difficulty level of $Q = 4$.

## *Learning Effect*

To evaluate if the EEG-based learning environment works and how it compares to an error-adaptive learning environment, we analyzed the learning effect of each subject by pre- and post-tests. Furthermore, the learning effects between the experimental and control group were compared.

Table 8.4 reports the learning effect of each individual subject of the experimental group, as well as the overall values. After using the EEG-based learning environment and thus learning how to calculate in an octal number system, a learning effect can be recognized for almost every subject, except for subject 2 and 5. On average, 5.08 assignments from 11 post-test tasks were solved correctly after completing the learning phase (see Table 8.4), 3.54 more assignments compared to the pre-test. On average a significant learning effect can be verified between the pre- and post-test ($p = 0.0026$, two-sided Wilcoxon test).

**Table 8.5** Number of correctly solved trials in the pre- and post-tests, as well as the learned factor for each subject of the control group and overall subjects

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | 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 |

The control group solved 1.55 tasks from the 11 pre-test assignments on average correctly. As the difference to the experimental group is not significant ($p > 0.05$, two-sided Wilcoxon test), equal prior knowledge can be implied for both groups. The best subject of the control group performed 36.36% of the pre-test tasks accurately, whereas the worst subjects gave no correct answers (see Table 8.5). For almost every subject, a learning effect for calculating in the octal number system is noticeable after the error-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 8.5), which also shows a significant learning effect between the pre- and post-tests ($p = 0.0016$, two-sided Wilcoxon test).

Although the learning effect for the group using the EEG-based learning environment is higher than for the control group, this difference is not significant ($p > 0.05$, two-sided Wilcoxon test). These results show that an EEG-based adaptation of a digital learning environment is possible and that results are similar to error-adaptive learning environments, which can be considered state-of-the-art.

## Towards More Differentiated Workload Estimation

We have shown in the previous section that it is possible to use a passive BCI approach to adapt content of a digital learning environment in real-time, based on the user's EEG recordings. While the EEG-based learning environment worked, it showed results that were as good as an error-adaptive learning environment. This raises the question of how we can improve the EEG-based learning environments and what can we do with an EEG-based learning environment that we can not do with an error-adaptive learning environment. In the proposed EEG-adaptive learning environment we adapted according to the user's workload, which was initially measured by the $Q$-value. As the $Q$-value was shown to highly correlate with the error-rate (Spüler et al. 2016), it is not surprising that the results are similar to the ones obtained with an error-adaptive learning environment.

As we used a very simply concept of workload in the previous work, a more detailed definition of workload could help to improve the EEG-based learning environment. Miyake and colleagues (Miyake et al. 2000) defined three components of workload: Updating, Inhibition, and Shifting. Updating is described as a process of (un-)loading, keeping, and manipulating information in memory during a certain period of time (Ecker, Lewandowsky, Oberauer, & Chee 2010), whereas Shifting can be described as the adjustment of a current mind set to new circumstances or new rules of a task (Monsell 2003). Inhibition comprises blending out (irrelevant) information or withholding actions while being involved in a cognitive task (Diamond 2013). Miyake and colleagues describe those three core functions as sharing many properties but also having distinct diversities that characterize the individual functions. If the diversities between the functions lead to different patterns of EEG activity, those patterns could be detected and used to gain more detailed information about the user. Instead of only assessing the mental workload, it was also possible to assess how demanding a task is in terms of updating, inhibition, and shifting, and the content could be adapted more precisely to match the individual needs of the user.

With the aim of developing a prediction model that is able to differentiate between workload components, we performed a study to find evidence of two of the components (Updating and Inhibition) in EEG data and see how well machine learning methods can be used to differentiate those components based on the EEG data.

### *Task Design: n-Back Meets Flanker*

The “n-back meets flanker” task (Scharinger, Soutschek, Schubert, & Gerjets 2015b) combines an n-back task with simultaneous presentation of flanker stimuli. The task was used to find and investigate interactions between the two executive functions Updating and Inhibition in terms of behavioral data as well as on a neurophysiological level. The n-back task is known to induce Updating demands in different intensities depending on the value of $n$ (Miyake et al. 2000; Jonides et al. 1997). In this experiment $n$ was altered between 0, 1, and 2 in a block design. In each trial, at each stimulus presentation, the subject had to react as accurately and quickly as possible to the following question: “Is the currently presented letter the same as the one I saw $n$ before?” A yes or no decision was required, which was recorded via a button press. In addition to the given answer, the reaction time was recorded. The stimuli consisted of a set of four letters, S, H, C, and F. The different values of $n$ indicated how many letters need to be constantly updated in memory throughout the specified block. For $n = 0$ the subjects had to remember one predefined stimulus letter, which remained the same throughout the block, which is why no updating of memory content was necessary.

The flanker task is known to induce Inhibition demands (Eriksen 1995; Sanders & Lamers 2002). A flanking stimulus can either be congruent or incongruent to the accompanied n-back task stimulus. An incongruent flanker usually introduces conflicts in the decision making, reactions eventually need to be inhibited even though the reflex urges the subject to react otherwise. The flanker stimuli in this experiment were presented simultaneously with the n-back stimulus and they were either congruent (identical) or incongruent (different) with the presented n-back stimulus. The central letter used for the n-back task was flanked by three letters on the right and three on the left (all identical). The subjects were instructed to not react to the flanker stimuli at all. Each n-back condition was presented twice, in which 154 trials where presented per block. Half of the stimuli of each block were targets, half were non-targets. The flanker items were chosen to be incongruent in one-third of all trials, congruent in the other two-thirds of all trials. By design an incongruent—incongruent sequence of stimuli was prevented, to avoid undesired effects. In addition to that in each block, a sequence of ten targets and non-targets was randomly chosen in which the central n-back stimulus was removed and only the flanker letters were presented. This was done to recreate awareness of the flanker items.

## *Data Recording and Preprocessing*

Twenty-two subjects (10 male and 12 female; mean age: 28.5 ± 4.3 years, 21–35) participated in the study. All participants had normal or corrected-to-normal vision, none reported neurological disorders and all were right-handed (with one exception). Twenty-two electrodes (Acticap brain products) were used for the recording at a sampling rate of 500 Hz and placed according to the 10/20 system (Jasper 1958) with the reference at right mastoid and the ground electrode at AFz.

The data was bandpass filtered between 0.4 and 40 Hz and re-referenced to the common average. To remove artifacts, a threshold of 100 $\mu$V was chosen and all trials exceeding this level were discarded, whereas trials including EOG artifacts were corrected using independent component analysis (ICA), in which artefactual independent components were removed by visual inspection. After these steps, trials were sorted and more trials were discarded. The first four trials of each block, for example, as they were always congruent non-targets, were removed as well as all congruent trials following incongruent trials because the so-called Gratton effect (the effect caused by the incongruence is bigger if the incongruent trial follows a congruent trial compared to an incongruent trial) should not influence the analysis. Furthermore, all trials in which the subject did not respond correctly to the n-back task were removed from the dataset. On average this resulted in 215 trials per subject that could be used for the analysis. More details can be found in the publication (Scharinger et al. 2015b).

## *Neurophysiological Analysis*

To investigate if Updating and Inhibition show different patterns in the EEG activity, the data were divided into trials containing Baseline (BL), Updating 1 (Up1), Updating 2 (Up2), and Inhibition (Inh). The first 0-1,000 ms after stimulus onset were used as epoch of interest. Updating demands are supposed to be induced for

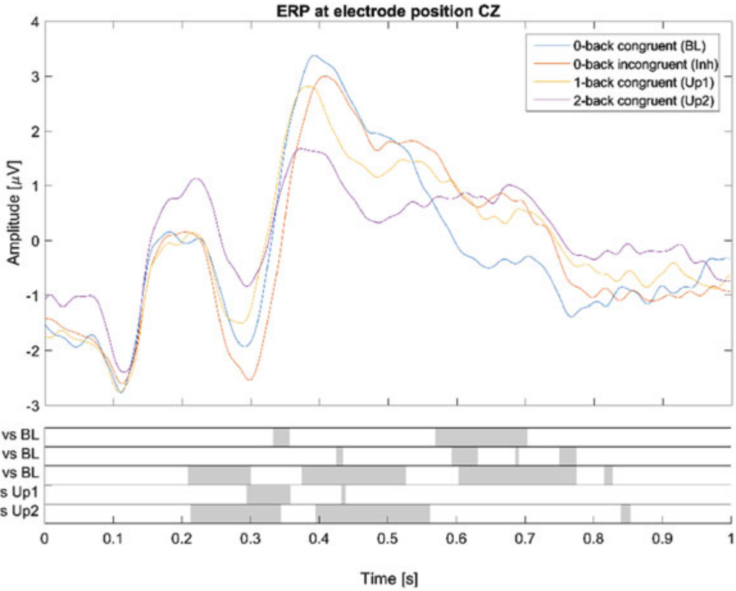

**Fig. 8.3** ERPs from 0-1,000 ms from stimulus onset for the four conditions: BL, Inh, Up1, and Up2 at position Cz. The values represent the grand average over all subjects and trials. The *grey bars* (*bottom*) state that the amplitude differs significantly ($p < 0.05$, corrected for multiple comparisons) for the respective comparison

1-back and 2-back, while Inhibition demands are expected for incongruent flanker. Due to this, the data were divided into four conditions:

- Baseline (BL): 0-back with congruent flanker
- Inhibition (Inh): 0-back with incongruent flanker
- Updating1 (Up1): 1-back with congruent flanker
- Updating2 (Up2): 2-back with congruent flanker

To find evidence of diversity between Inhibition and Updating in the EEG, we first looked at the event-related potentials (ERPs) in the data. As the amplitude of the P300 at electrode positions Fz, Cz, Pz is supposed to be correlated with workload, we investigated the ERPs at those positions. While nearly no effects were found at electrode Fz, we found significant differences in amplitude between all conditions in the time range of 200–750 ms for electrodes Cz and Pz. The results for electrode Cz are shown in Fig. 8.3.

As this shows that Inhibition and Updating show different patterns in the ERPs, we further analyzed the data with regard to the power spectrum. Therefore, we used Burg's maximum entropy method (Burg 1972) with a model order of 32 to estimate the power spectrum between 4 and 13 Hz (theta and alpha band). Significant ($p < 0.05$) workload-related changes could be observed in the power spectrum in both frequency bands at Fz and Cz, while Pz showed only a significant effect in the alpha

**Table 8.6** Classification accuracies averaged over all subjects using either evoked response potentials (ERPs) for classification or the power spectrum

| Features | Inh vs BL | Up vs BL | Up2 vs BL | Up vs Inh | Up2 vs Inh |
|---|---|---|---|---|---|
| RP | 61.12%*** | 61.20%*** | 73.38%*** | 68.23%*** | 74.73%*** |
| Spectrum | 52.48%* | 52.68%* | 63.17%*** | 55.46%*** | 64.13%*** |

Significance regarding the difference to chance level accuracy is indicated as follows: *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$

band, which therefore was much stronger than at the other electrodes. However, the only significant effects ($p < 0.05$) between Updating (up1) and Inhibition were found in the theta band at electrode Cz.

## *Classification*

As there is evidence that Inhibition and Updating show different patterns of EEG activity, we tried to use machine learning methods to classify the data in order to detect if a task induced Updating or Inhibition demands. For classification we used the ERP data, as well as the power spectral data. Data from 17 inner channels were used for classification by a support vector machine (SVM) with a linear kernel and a hyperparameter of $C = 1$. For the classification of the ERPs, we further used a CCA-based spatial filtering method (Spüler, Walter, Rosenstiel, & Bogdan 2014) to improve classification accuracy.

Classification results are shown in Table 8.6. While all results were significantly above chance level accuracy, classification using ERPs worked better than using the power spectrum. While Inhibition and Updating could individually be classified against baseline data, it was also possible to discriminate between Inhibition and Updating demands. While this is a clue that there might be differences between Inhibition and Updating, the results could also be explained by the different conditions inducing different amounts of workload and the classifier detecting that different amounts of workload. To rule out that hypothesis we performed a classification where the classifier was trained on one condition (e.g., BL vs. Inh) and tested on a different condition (e.g., Up1). If the patterns between the conditions were similar and only varied in strength due to different amounts of workload, we would expect that the conditions would still be classified correctly as a high-workload condition. However, most of the results were not significantly above chance ($p > 0.05$), showing that classification did not work. Only for one test, in which the classifier was trained on *Up2 vs. BL*, we obtained a classification accuracy significantly different from chance level ($p < 0.05$). As this classifier classified Inhibition trials mostly as BL, this indicates that Inhibition has a pattern more similar to Baseline than to the Up2 condition.

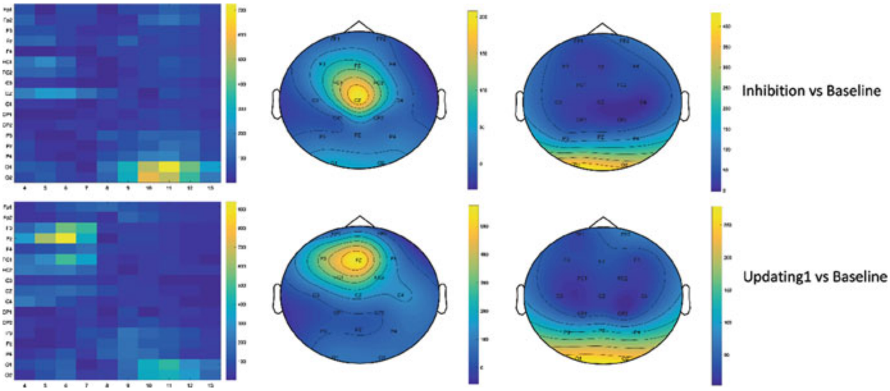

**Fig. 8.4** Activation patterns obtained by SVM classification for Inhibition vs. Baseline (*top*) and Updating1 vs. Baseline (*bottom*). The activation patterns for all electrodes and frequencies are shown(*left*), a topographic display of the activation pattern for the theta-band (*middle*) and for the alpha-band (*right*)

In addition to the previous neurophysiological analysis regarding the changes in the power spectrum, we looked at the classifier weights in order to visualize the differences in the power spectra between Inhibition and Updating. As weights of a linear classifier should not be interpreted directly, we used the feature covariance matrix to calculate neurophysiologically interpretable activation patterns (Haufe et al. 2014) from the classifier weights. The activation patterns for Inh vs. BL and Up1 vs. BL are shown in Fig. 8.4. It can be seen that occipital alpha activity is a strong indicator of changes in workload for both components. In the theta-band, the pattern differed between Inhibition and Updating. While the theta activity in the central region changed with workload for the Inhibition component, the Updating component showed workload-related changes in the frontal area. Not only does the strength of the effect vary between the two components, also the location does; this is a further argument that the two components of the Miyake-model are also different on a neurophysiological level. With regard to an EEG-based learning environment, this opens up new possibilities as detecting the individual Inhibition and Updating demands of a task could lead to a more specific adaptation of the content than only detecting the overall workload.

## Discussion and Outlook

In this chapter, we have outlined the progress of our project and our efforts to develop an EEG-based learning environment. While the lessons learned from our first studies were already outlined in the section "Pitfalls in Designing EEG-Based Learning Environments", the main message of these studies was that cross-task classification is not feasible. Although the cross-task classification could be improved by using subjective labels, a subjective labeling introduced other problems (difficulty to collect labels and subjective bias), which is why the cross-task classification was not investigated any further.

As an alternative, we employed a cross-subject classification, in which a prediction model is trained on a number of subjects, and used to predict workload in a different subject. Thereby, it is possible to use the workload prediction on a subject without the need for a time-consuming subject-specific calibration phase. As this method performed well when tested offline, it was tested online to adapt the content of a digital learning environment in real-time based on the EEG signals. This study was a successful proof-of-concept that an EEG-based adaptation of a digital learning environment is possible. However, we could not show any benefit of the system as it performed similarly to an error-adaptive learning environment. As this work also revealed new difficulties that arise when trying to evaluate a digital learning environment in an online study, these two issues need to be discussed in more detail.

## *Evaluating an EEG-Based Learning Environment*

Commonly, if neural signals (like EEG) are used to estimate a user's mental state or a user's intention, the performance of the prediction method is assessed in terms of accuracy or other metrics that try to quantify how well the prediction model is working. While we have also evaluated the prediction performance of our model in the offline study (see section "EEG-Based Workload Assessment During Arithmetic Learning"), this kind of assessment is no more feasible with an online learning environment.

The reason for this is the lack of an objective measure for the user's workload. For the creation of the prediction model we used EEG data from a task that all subjects were carrying out fluently (addition in a decimal system). As no learning effects are expected in this case, the difficulty of the task (measured by $Q$) was used as a subjective measure of expected workload. But the sole purpose of an online learning environment is to induce learning effects, and we have shown that our system does so. Thereby, the relationship between task difficulty ($Q$) and workload is invalidated. At the beginning, when the task is unknown to the user, even easy exercises will induce a high workload. After using the learning environment, the user may have mastered arithmetic tasks in the octal number system, and even exercises with moderate difficulty will result in a low workload. This example shows that the relationship between task difficulty and workload is invalidated if learning effects are present.

As the task difficulty measured by the $Q$-value is the only objective measure we have, and the relationship to workload is invalid in the online scenario, there are no means to objectively assess the prediction performance of our model. Although we cannot assess the performance of the prediction model in the online scenario, we can evaluate the EEG-based learning environment with regard to its effect on learning success. Learning success was defined as the difference in score between the pre- and post-tests, which were performed by the subjects before and after using the learning environment. As it is not important for the learner how accurate the workload-prediction works, but it is important how much the learning success can be improved, learning success is also the most user-centered metric.

Due to these facts, that other usually used metrics are not applicable for the scenario of an online learning environment and that learning success is the most user-centered metric, we think that learning success should be used as prime outcome measure for this kind of study.

### *Benefits of an EEG-Based Learning Environment*

As subjects using the EEG-based learning environment showed a significant learning effect, we could show that the EEG-based learning environment works. When comparing the experimental group using the EEG-based learning environment with the control group using the error-based learning environment, the learning success was higher for the experimental group, but the difference was not significant. This basically indicates that an EEG-based learning environment is an alternative to the state-of-the-art approach but offers no benefits in its current form. However, the presented system is meant as a proof-of-concept, and the parameters could be tweaked to optimize workload prediction, as well as the adaptation of the learning material.

Although we aimed at a high usability for the presented system by using a cross-subject prediction model to omit a subject-specific calibration phase, the usability of the current system is still rather low. The use of gel-based EEG needs some time and effort to prepare, thereby making it impractical to use an EEG-based learning environment on a wide basis. As we have shown in the section "EEG-Based Workload Assessment During Arithmetic Learning", the number of electrodes could be drastically reduced to reduce preparation time. Another approach would be the use of dry EEG electrodes, which would reduce the preparation time even more, as they don't require any gel. However, dry electrodes have a lower signal-to-noise ratio and would thereby lead to a lower prediction accuracy. Regardless of the preparation time, the costs of an EEG setup are high, which raises the question: what are the benefits of an EEG-based adaptive learning environment that would justify the additional cost and time investment compared to a normal digital learning environment?

It is likely that the answer is that there are no benefits when using an EEG-based learning environment in a broad population (e.g., in a class room). However, this technique could prove helpful for special cases in which the user suffers from learning disability or other problems. In the presented study, one subject of the experimental group suffered from test anxiety and reported feeling very comfortable using the EEG-based learning environment, as the system turned down the difficulty every time the subject was close to feeling overwhelmed, thereby providing a good learning experience. Depending on the individual case, the good learning experience can outweigh the disadvantages of the setup time and costs and thus render the usage feasible.

## *Improving an EEG-Based Learning Environment*

As the EEG-based learning environment merely served as a proof-of-concept to show that the idea is feasible, it also inspires one to think about possible improvements that could be made to enhance the prediction performance. On one side, the system could be improved by using better classification and signal processing methods from the area of BCI research. While we used a simple normalization method of the EEG signal, which helped to improve results, results could be further improved by using more advanced methods to normalize EEG data like PCA-based covariate shift adaptation (Spüler, Rosenstiel, & Bogdan 2012b) or a classifier adaptation (Spüler, Rosenstiel, & Bogdan 2012a), which improves the classifier while using the learning environment and adapts it to changes in the EEG data. Regarding the cross-subject classification approach, there are also approaches from the field of transfer learning (Tu and Sun 2012; Samek, Meinecke, & Müller 2013) that could be used to improve the cross-subject (and possibly also a cross-task) classification, which should be evaluated in future studies.

Despite improving only the prediction performance, one could also try to increase the information about the user, by not only measuring workload, but also how demanding a task is regarding different subcomponents of workload. As each learner is different, one might have more problems with tasks that have a high Updating demand, while others might have more problems with high Inhibition demands. By being able to discriminate between those two, the learning content could be adapted more specifically to the individual needs of the user. Measuring the Inhibition demand could potentially also be used as an indicator of distractions outside of the learning environment and would thereby be an interesting measure for the application when applied in a classroom setting.

Measuring workload or differentiating workload into subcomponents are not the only things that could be done with an EEG-based learning environment. It was also shown that mental states like vigilance, alertness, or attention can be detected in EEG (Oken, Salinsky, & Elsas 2006; Berka et al. 2007). If those can be detected by an EEG-based learning environment this could lead to a more comprehensive monitoring of the user and not only allow for better adaptation of the learning content, but also introduce breaks or place motivating rewards based on the mental state of the user to improve learning effort and learning success of the user.

Lastly, such a BCI-based learning environment could also be improved by not only using EEG to measure neurocognitive processes, but also use other physiological signals, such as Eye-Tracking. As was shown by Scharinger, Kammerer, & Gerjets (2015a), cognitive workload influences pupil dilation, which can be measured in real-time and could be used as additional information to improve an adaptive learning environment. In the following chapter of this book (Scheiter et al.

2017), a learning environment is presented that measures eye-movements of the student and uses this additional information for adaptation. As the combination of eye-tracking information with neurophysiological signals (like EEG) would give a more comprehensive view of the current state of the learner, combining EEG with other physiological signals like eye tracking could allow for an even better adaptation of the system.

**Acknowledgements** This research was financed by the Leibniz ScienceCampus Tübingen "Informational Environments". It was further supported by the *Deutsche Forschungsgemeinschaft* (DFG; grant SP-1533\2-1) and the LEAD Graduate school at the Eberhard-Karls University Tübingen, which is funded by the Excellence Initiative of the German federal government.

## References

Askew, M. (2015). Numeracy for the 21st century: A commentary. *ZDM: The International Journal on Mathematics Education, 47*(4), 707–712.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T.,. . ., Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space and Environmental Medicine, 78*(Supplement 1), B231–B244.

Brouwer, A.-M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., & Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering, 9*(4), 045008.

Burg, J. P. (1972). The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics, 37*(2), 375–376.

Calder, N. (2015). Student wonderings: Scaffolding student understanding within student-centred inquiry learning. *ZDM: The International Journal on Mathematics Education, 47*(7), 1121–1131.

Causse, M., Fabre, E., Giraudet, L., Gonzalez, M., & Peysakhovich, V. (2015). EEG/ERP as a measure of mental workload in a simple piloting task. *Procedia Manufacturing, 3*, 5230–5236.

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In *Proceedings of the 8th International Conference on User Modeling* (pp. 137–147).

Cover, T., & Thomas, J. (2006). *Elements of information theory*. Hoboken, NJ: Wiley-Interscience.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135.

Dowker, A. (2004). *What works for children with mathematical difficulties?* (Vol. 554). Nottingham: DfES Publications.

Ecker, U. K., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning Memory and Cognition, 36*(1), 170.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition, 2*(2–3), 101–118.

Gerjets, P. H., & Hesse, F. W. (2004). When are powerful learning environments effective? the role of learner activities and of students conceptions of educational technology. *International Journal of Educational Research, 41*(6), 445–465.

Gerjets, P., Scheiter, K., & Cierniak, G. (2009). The scientific value of cognitive load theory: A research agenda based on the structuralist view of theories. *Educational Psychology Review, 21*(1), 43–54.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). 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*, 385.

Gevins, A., Smith, M., McEvoy, L., & Yu, D. (1997, Jun). 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.

Graesser, A., & McNamara, D. (2010). Self-regulated Learning in Learning Environments with Pedagogical Agents that Interact in Natural Language. *Educational Psychologist, 45*, 234–244.

Harmony, T., Ferna'ndez, T., Silva, J., Bosch, J., Valde's, P., Ferna'ndez-Bouzas, A.,. . ., Rodríguez, D. (1999). Do specific eeg frequencies indicate different processes during mental calculation? *Neuroscience Letters, 266*(1), 25–28.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage, 87*, 96–110.

Jasper, H. (1958). The 10/20 international electrode system. *EEG and Clinical Neurophysiology, 10*, 371–375.

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber E. J., Awh, E., Minoshima, S., Koeppe, R.A. (1997). Verbal working memory load affects regional brain activation as measured by pet. *Journal of Cognitive Neuroscience, 9*(4), 462–475.

Karagiannakis, G. N., & Cooreman, A. (2014). Focused MLD intervention based on the classification of MLD subtypes. In *The Routledge International Handbook of Dyscalculia and Mathematical Learning Difficulties* (p. 265).

Käser, T., Baschera, G.-M., Busetto, A. G., Klingler, S., Solenthaler, B., Buhmann, J. M., Gross, M. (2013). Towards a framework for modelling engagement dynamics in multiple learning domains. *International Journal of Artificial Intelligence in Education, 22*(1–2), 59–83.

Kirschner, P., & Gerjets, P. (2006). Instructional design for effective and enjoyable computer-supported learning. *Computers in Human Behavior, 22*(1), 1–8.

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller K.-R., Curio, G.,. . ., Kincses, W. E. (2007). Improving human performance in a real operating environment through real-time mental workload detection. In *Toward Brain-Computer Interfacing* (pp. 409–422). Cambridge, MA: MIT Press.

Kong, J., Wang, C., Kwong, K., Vangel, M., Chua, E., & Gollub, R. (2005). The neural substrate of arithmetic operations and procedure complexity. *Cognitive Brain Research, 22*(3), 397–405.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134–140.

Murata, A. (2005). An attempt to evaluate mental workload using wavelet transform of EEG. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 47*(3), 498–508.

Oken, B., Salinsky M., & Elsas, S. (2006). Vigilance, alertness, or sustained attention: Physiological basis and measurement. *Clinical Neurophysiology, 117*(9), 1885–1901.

Richards, K. C., Enderlin, C. A., Beck, C., McSweeney J. C., Jones, T. C., & Rober son, P. K. (2007). Tailored biobehavioral interventions: A literature review and synthesis. *Research and Theory for Nursing Practice, 21*(4), 271–285.

Samek, W., Meinecke, F. C., & Müller, K.-R. (2013). Transferring subspaces between subjects in brain-computer interfacing. *IEEE Transactions on Biomedical Engineering, 60*(8), 2289–2298.

Sanders, A., & Lamers, J. (2002). The Eriksen flanker effect revisited. *Acta Psychologica, 109*(1), 41–56.

Scharinger, C., Kammerer, Y., & Gerjets, P. (2015a). Pupil dilation and eeg alpha frequency band power reveal load on executive functions for link-selection processes during text reading. *PloS One, 10*(6), e0130608.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015b). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology, 52*(10), 1293–1304.

Scheiter, K., Fillisch, B., Krebs, M.-C., Leber, J., Ploetzner, R., Renkl, A., et al. (2017). How to design adaptive multimedia environments to support self-regulated learning. In *Informational Environments: Effects of Use Effective Designs* (Chap. 9).

Schlögl, A., Keinrath, C., Zimmermann, D., Scherer R., Leeb, R., & Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clinical Neurophysiology, 118*(1), 98–104.

Soltanlou, M., Jung, S., Roesch, S., Ninaus, M., Brandelik, K., Heller, J., et al. (2017). Behavioral and neurocognitive evaluation of a web-based learning platform for orthography and arithmetic. In *Informational Environments: Effects of Use Effective Designs* (Chap. 7).

Spüler, M. (2015). A Brain-Computer Interface (BCI) system to use arbitrary Windows applications by directly controlling mouse and keyboard. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4087–1090).

Spüler, M., Rosenstiel, W., & Bogdan, M. (2012a). Adaptive SVM-based classification increases performance of a MEG-based Brain-Computer Interface (BCI). In *International Conference on Artificial Neural Networks* (pp. 669–676).

Spüler, M., Rosenstiel, W., & Bogdan, M. (2012b). 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.

Spüler, M., Walter, A., Rosenstiel, W., & Bogdan, M. (2014). Spatial filtering based on canonical correlation analysis for classification of evoked or event- related potentials in EEG data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 22*(6), 1097–1103.

Spüler, M., Sarasola-Sanz, A., Birbaumer, N., Rosenstiel, W., & Ramos-Murguialday, A. (2015). Comparing metrics to evaluate performance of regression methods for decoding of neural signals. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1083–1086).

Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller K., & Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: A step towards online adaptation in numerical learning. *ZDM: The International Journal on Mathematics Education ZDM, 48*(3), 267–278.

Stanescu-Cosson, R., Pinel, P., van de Moortele, P.-F., Le Bihan, D., Cohen, L., & Dehaene, S. (2000). Understanding dissociations in dyscalculia. *Brain, 123*(11), 2240–2255.

Sweller, J., Van Merrinboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

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

Tu, W., & Sun, S. (2012). A subject transfer framework for eeg classification. *Neurocomputing, 82*, 109–116.

Walter, C., Cierniak, G., Gerjets, P., Rosenstiel, W., & Bogdan, M. (2011). Classifying mental states with machine learning algorithms using alpha activity decline. In *European Symposium on Artificial Neural Networks*

Walter, C., Schmidt, S., Rosenstiel, W., Gerjets, P., & Bogdan, M. (2013). Using cross-task classification for classifying workload levels in complex learning tasks. In *Affective Computing and Intelligent Interaction (ACII), 2013* (pp. 876–881).

Walter, C., Wolter, P., Rosenstiel, W., Bogdan, M., & Spüler, M. (2014, 09). Towards cross-subject workload prediction. In *Proceedings of the 6th International Brain-Computer Interface Conference*, Graz, Austria.

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

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical Neurophysiology, 113*(6), 767–791.