Put Your Thinking Cap On: Detecting Cognitive Load using EEG during Learning

Caitlin Mills[1], Igor Fridman[2], Walid Soussou[2], Disha Waghray[3],
Andrew M. Olney[4], & Sidney K. D'Mello[5]

[1] University of British Columbia; [2] Quantum Applied Science and Research (QUASAR) Inc.;
[3] Indiana University; [4] University of Memphis; [5]University of Notre Dame
[1]2136 West Mall
Vancouver, BC, V6T 1Z4, Canada
caitlin.s.mills@psych.ubc.ca I sdmello@nd.edu

ABSTRACT

Current learning technologies have no direct way to assess students' mental effort: are they in deep thought, struggling to overcome an impasse, or are they zoned out? To address this challenge, we propose the use of EEG-based cognitive load detectors during learning. Despite its potential, EEG has not yet been utilized as a way to optimize instructional strategies. We take an initial step towards this goal by assessing how experimentally manipulated (easy and difficult) sections of an intelligent tutoring system (ITS) influenced EEG-based estimates of students' cognitive load. We found a main effect of task difficulty on EEG-based cognitive load estimates, which were also correlated with learning performance. Our results show that EEG can be a viable source of data to model learners' mental states across a 90-minute session.

CCS Concepts

• H.5.m. Information interfaces and presentation (e.g., HCI) → Miscellaneous • K.3.1 computers and Education: Computer Uses in Education

Keywords

EEG; intelligent tutoring systems; engagement; cognitive load

1. INTRODUCTION

What if your learning environment could assess how deeply you were thinking? What if it could anticipate your learning struggles without you having to click on the help button, idle on the screen, or before you answered questions incorrectly. This kind of learning technology would know the perfect time to help, both when you are having trouble or when you are barely paying attention.

Although such an omniscient learning technology is not yet available, students' cognitive and affective states have been reliably inferred using a variety of indirect data sources – eye gaze, mouse movements and keystrokes, click stream data, facial features, and speech and language, to name a few [3, 4, 8, 9, 20, 31, 43]. In fact, a few intelligent tutoring systems (ITSs) have already successfully implemented cognitive and affect detection systems to improve learning [15, 17]. However, access to more direct neurophysiological measures of students' cognitive processes would likely improve current systems. Electroencephalography (EEG) is one such measure that is the present focus.

EEG measures the voltage of coordinated neural firing that passes through the scalp. Different patterns of the neural firing activity can be indicative of distinct cognitive states, such as attentional focus, cognitive load, and engagement [1, 7, 14]. EEG is ostensibly the least invasive and most affordable method of accessing brain activity, yet it is rarely used in education due to several complexities involved. However, we believe that with advances in technology, there is considerable potential for EEG-based measures of students' cognitive states. As an initial step, we focus on modeling cognitive load because of its well-established relationship with both EEG measures and learning outcomes (see [2, 23]).

Cognitive load theory suggests that working memory capacity is limited and cognitive load is essentially a measure of how much "space" in working memory is currently being used [49, 54]. Cognitive load can take two different forms: intrinsic and extraneous (although see [33, 34, 49] for debate about a third type of load, called germane load) . Intrinsic load is imposed by the basic structure of the learning task and is related to the number and interactivity of informational elements in working memory (e.g., three-digit addition imposes more load than single digit addition). In contrast, extraneous load is imposed by the way information is presented and is considered ineffective load in that it does not directly contribute to schema construction [49]. Importantly, the different types of load are additive forces in memory. Critically, if cognitive load exceeds memory resources, learning will be stifled. However, increases in cognitive load are not necessarily negatively related to learning. In fact, imposing cognitive load that contributes to schema formation can positively relate to performance [33].

The overarching goal of our project is to build a system that can automatically measure cognitive load in real-time and optimize its instructional strategies to promote intrinsic cognitive load, while avoiding "cognitive overload." This requires establishing that we can reliably model cognitive load during learning – which is the aim of the present study.

1.1 Related Work

Previous work has explored the utility of EEG in predicting cognitive load in a variety of real-world tasks [1, 2, 7, 14, 24, 38].

LAK '17, March 13-17, 2017, Vancouver, BC, Canada
© 2017 ACM. ISBN 978-1-4503-4870-6/17/03⋯$15.00
DOI: http://dx.doi.org/10.1145/3027385.3027431

Kohlmorgen et al. [36] used EEG signals to optimize automobile drivers' cognitive load in a driving simulation study, finding that lower load was associated with increased responsiveness. Other studies have shown that differences in cognitive load can be detected in memorization and vigilance tasks, as well as while solving arithmetic problems [7, 24, 52]. Cognitive load has also been shown to predict performance on hand written math problems and in memory tasks [13, 14].

Only a handful of studies have examined cognitive load during learning with educational technology [13, 30, 42, 43]. This omission can be attributed to a lack of scalability of EEG, stemming from the cost of EEG technologies as well as the (historically) intrusive nature of the electrodes, which can require dozens of wet electrodes. However, recent technological developments have improved EEG in three key ways: the cost is lower, a smaller number of electrodes can be used, and dry electrodes are just as effective as wet ones [37, 38, 42]. Thus, as the equipment limitations subside, it is important to consider what types of information EEG can provide during learning from technology.

Initial work using EEG during learning with an educational technology has yielded some promising results [16, 29, 30, 42]. For example, EEG waves (alpha, beta, gamma, and theta) were shown to be correlated with motivation during a serious game [16] and with affect during a true-false question task [12]. In another study, machine learning classifiers trained on EEG data were used to predict the correctness of students' answers during reading [29], though these results may be biased since cross validation was not done at the student level, which can lead to overfitting.

Two particularly relevant studies provide evidence that EEG measures of cognitive load are related to the difficulty of the instructional materials [10, 11]. The first study recorded EEG signals while students engaged in multiple tasks, including basic tasks known to induce cognitive load (e.g., digit span, logic tasks) as well as complex problem solving tasks [11]. EEG signals were used to measure cognitive load, using Gaussian Process Regression for each participant. Cognitive load was positively correlated with the difficulty of the task and was negatively related to performance on the basic tasks, but not the complex problem solving task.

A second study collected EEG data while children and adults read difficult and easy passages with an ITS, either silently or out aloud [10]. Easy passages were taken from the K-1 common core database, whereas difficult passages were from the GRE and GED exams. The passages ranged from 62 to 83 words in length. The EEG signal from a single electrode over the frontal lobe was used to train a reader-specific classifier using leave-one-story-out validation and a reader-independent classifier using leave-one-participant-out validation. Separately, the adult and children classifiers were inconsistent in predicting the difficulty of the text. Accuracies for the adult models ranged from 39% to 58% for the reader-specific models and from 49% to 60% for the reader-independent models (chance = 50%). None of the models were above chance for the children (accuracies ranging from 42% to 50%) [10]. When adult and child data were combined, only the reader-independent models reached above-chance levels (accuracy = 56%). Moreover, classifiers were also trained to predict whether comprehension questions would be answered correctly. However, classifiers were only successful for silent reading (not reading aloud) when trained on EEG data collected *while* the question was being answered. Performance was below chance when the classifiers were trained on data while students actually read the texts. Overall, this study provides some evidence that EEG can index difficulty of the material, but the classification results were modest and the dataset was limited to reading short passages, which may not generalize to more interactive learning environments.

Taken together, previous work suggests that EEG might be a valuable tool for assessing students' cognitive processing during learning. Further, EEG may also be a good candidate for real-time optimization and feedback [36, 53]. However, it is unclear if the findings generalize across multiple learning interactions (e.g., reading, listening to a lecture) and in a domain-independent fashion.

1.2 Current Study

Cognitive load theory suggests that learning can be either facilitated or hindered by the amount and type of cognitive load introduced by the instructional design [49]. The current project aims to develop an ITS that can dynamically tailor its instructional strategies to impose an appropriate amount of cognitive load for each participant. We begin by developing an EEG-based detector of cognitive load and testing its sensitivity to an experimental manipulation of instructional difficulty embedded in an ITS called Guru [46]. Difficulty was manipulated in an Instruction phase where material is initially introduced and explained and a Scaffolding phase where students answer questions about target concepts and receive immediate feedback.

We used a recently developed QUASAR [37, 38] headset featuring dry electrodes that fits on the head similar to a hat (see Figure 1). High and low cognitive load states were trained from data collected in a separate training phase using an adaptive algorithm that leverages EEG spectral features using partial least squares regression. The models were then used to predict cognitive load during interactions with an ITS. Whereas some of the tasks were domain-specific (i.e. related to difficulty manipulations in Guru), others were domain-independent (e.g., one vs. three-digit addition for easy vs. difficult conditions, respectively). This was done in order to address the critical research question of how specific the training tasks need to be to accurately predict cognitive load?

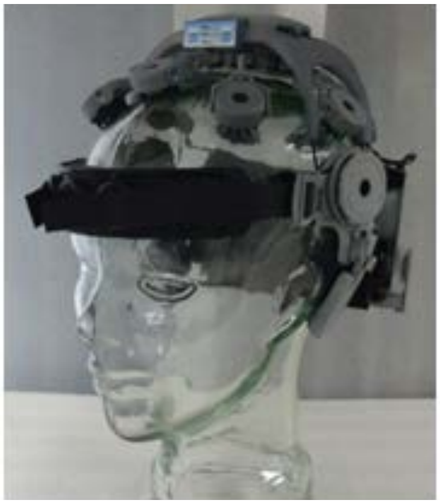

Figure 1. Example of QUASAR Headset

We also address the issue of EEG viability over the course of a learning session. This is an important question when students need to sustain attention for extended periods of time. Although it is desirable to collect data to infer cognitive states, it is equally important that students feel comfortable and unconstrained by the sensors. Thus, one of the challenges of collecting EEG data is minimizing interference with the learning session, while maximizing data validity. We evaluate data viability by quantifying how much valid data can be collected over the course of two 20-minute back to back tutoring session after setup and training tasks are completed.

2. Implementation

2.1 Guru

Guru is a dialogue-based ITS in which an animated tutor agent engages the student in a collaborative conversation that references a multimedia workspace. Guru is distinct from most dialogue-based ITSs, such as AutoTutor [27, 56] or Why-Atlas [55] because it is modeled on 50-hours of observations of *expert* human tutors that reveal markedly different pedagogical strategies from novice tutors [18]. The computational models of expert tutoring embedded in Guru are multi-scale, ranging from tutorial modes (e.g., scaffolding), to collaborative patterns of dialogue moves (e.g., information-elicitation), to individual moves (e.g., direct instruction) [46]. An in-school evaluation study has shown that Guru is as effective as human tutors in improving learning outcomes, and is more effective than classroom instruction alone [47].

Guru covers 120 biology topics aligned with the Tennessee Biology I Curriculum Standards in 15 to 40-minute tutoring sessions. The topics are organized around core *concepts* (e.g. "proteins help cells regulate functions") that Guru attempts to get students to articulate over the course of the session.

Guru's interface (see Figure 2) consists of a multimedia panel, a 3D animated agent, and a response box. The agent speaks, gestures, and points using motion capture driven animation. Throughout the dialogue, the tutor gestures and points to areas on the multimedia panel that are most relevant to the current discussion and are slowly revealed as the dialogue advances. Student typed input is analyzed using natural language processing techniques to maintain a student model that is used to tailor instruction to individual students.

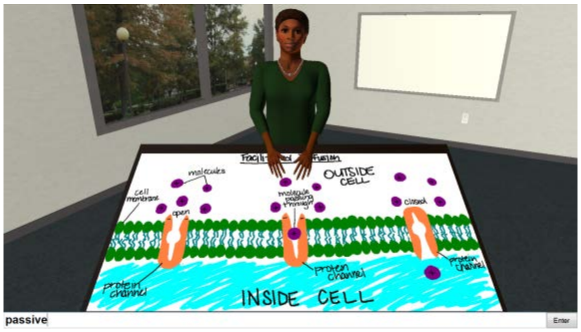

Figure 2. Screenshot of typical Guru content.

A typical Guru session is typically ordered in phases: Preview, Common Ground Building Instruction (CGB Instruction), Summary, Concept Maps I, Scaffolding I, Concept Maps II, Scaffolding II, and Cloze Task. We focused on the two dialog-oriented learning phases (CGB Instruction and Scaffolding) in the current study. **CGB Instruction** (sometimes called collaborative lecture) is where basic information and terminology is covered. This step is essential because biology involves considerable specialized terminology that needs to be discussed before more collaborative knowledge building activities can proceed. **Scaffolding** in the typical version uses a Prompt → Feedback → Verification Question → Feedback → Elaboration cycle to cover target concepts. In the adapted version of Scaffolding, the tutor would ask a question, process the student's answer, and provide feedback/explanation.

We used two topics in the study that were selected in consultation with the high school where data was collected. The topics were selected to minimize overlap with previous topics covered in classroom lecture (based on the syllabus) in order to minimize the effects of prior knowledge. The topics were: *Maintaining Temperature* (Topic A) and *Trophic Levels* (Topic B). The Maintaining Temperature topic focused on how humans and other animals regulate their internal body temperature in order to stay alive during both hot and cold environments. The Trophic Levels topic focused on how energy is transferred across multiple levels of the food chain.

2.2 Difficulty Manipulation

We manipulated difficulty by creating two versions the CGB instruction and Scaffolding sections of Guru. Easy and difficult versions of CGB Instruction were created by manipulating the complexity of the tutor's spoken content based on dimensions that are known to play an important role in text complexity [25, 28]: narrativity, syntactic ease, and referential cohesion. Easy versions consisted of shorter, simpler sentences with higher frequency words (e.g., replacing the low-frequency word "modicum" with a higher-frequency word like "small"). Difficult versions had more complex, longer sentences with lower frequency words. As an example, consider the difficult and easy versions of the following sentence: (difficult) "Once the brain detects increased heat, it instigates the pumping of blood nearer to the skin." vs. (easy) "The brain will realize that it is too hot. Then it will begin to pump blood up close to the skin." Importantly, the content and number of words was kept consistent across the easy and difficult versions (p = .86 for a paired t-test comparing length).

We assessed the linguistic features of the easy and difficult versions of the CGB Instruction content using Coh-Metrix, a computational text analyses tool [26]. We focused on the three linguistic features (narrativity, referential cohesion, and syntactic simplicity) that have been linked to text difficulty [25] (see Table 1). We conducted paired samples t-tests to compare the easy and difficult versions of the CGB Instruction text. There were four easy-difficult pairs, one for each of the two sections (first or second) within each topic (2 x 2 = 4). The easy and difficult versions were significantly different (*ps* < .05), with effect sizes ranging from 1.88 to 6.03 sigma.

In addition, Flesch Kincaid Grade Level (FKGL), a widely used measure of text difficulty [35], of the easy version was over 3 grade levels lower than FKGL for the difficult version (4.9 and 8.5, respectively); the difference was also statistically significantly different ($p < .05$).

Table 1. Descriptive statistics for difficulty manipulation metrics

	Easy	**Difficult**
Metric	*M (SD)*	*M (SD)*
Narrativity	58.6 (8.15)	45.5 (5.64)
Syntactic Ease	96.6 (2.63)	84.4 (6.93)
Referential Cohesion	55.2 (7.78)	30.8 (3.68)
FKGL	4.90 (.680)	8.50 (.500)

Note. FKGL = Flesch Kincaid Grade Level.

For Scaffolding, we manipulated difficulty based on evidence that recall is more effortful than recognition [40, 44]. Thus, we created two versions of every question. The prompt question was phrased as a true-false question for the easy version (e.g., "True or false: The second trophic level is composed of primary consumers."), whereas the difficult version were open-ended questions that required students to recall the answer in the absence of answer choices (e.g., "What term is used to describe the organism s place on the food chain?") Students' answers to the difficult questions were scored by comparing their typed answers to a list of possible predefined correct answers (e.g., correct answer: "trophic level[s]").

2.3 EEG Headset

EEG was collected using a prototype QUASAR 24-channel EEG headset (see Figure 3). The headset uses ultra-high impedance dry-electrode technology, which has been demonstrated to record high-quality EEG without the need for skin preparation of any kind [37, 37, 39]. It was designed to be a light, low cost unit, specifically developed for ecological data collection. The system is self-contained, including data acquisition, data storage, cable and wireless data output, and batteries. The headset can be put on the head similar to wearing a baseball cap, while reliably positioning the sensors.

Previous work suggests that this system produces reliable data. First, signal quality recorded from the dry sensors has been shown to have a 90% correlation to data obtained from clinical wet electrodes attached by a technician [38]. Another study compared wet versus dry electrodes by correlating the output from the two types across three different conditions (eyes closed/open, n-back task, and an artifact condition where participants were instructed to move their jaw, head, eyes, and shoulders one at a time). They reported an average correlation of .85 for the frontal site, and .39 for the parietal site [21]. The authors were encouraged by the high correlations at the frontal sites, and suggested that low parietal correlations may have been due to the close proximity of the reference location and the parietal electrode site (causing errors in noise subtraction). Finally, more promising evidence comes from a recent clinical evaluation of a similar 20-sensor EEG system, which determined that data quality from a dry electrode system was suitable for diagnosing status epilepticus and seizure activity within 190 seconds of donning it [51]. The present study is the first test of the dry EEG sensors during learning in a classroom setting.

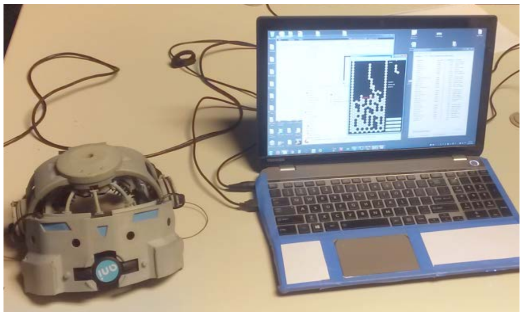

Figure 3. Image of EEG and computer setup (Tetris task shown)

2.4 Training Tasks

EEG recordings of the following tasks were then used to train the cognitive workload model.

Forward digit span (FDS-*x*): A sequence of digits is flashed one-by-one on the screen (for one second). The participant is then asked to recall the sequence in the order presented and enter it using the keyboard. They are shown a short (2 sec) correctness feedback message followed by the next sequence of digits. This task taxes short-term working memory. Difficulty (*x*) varies from 4 digits (easy) to 9 digits (difficult). Typical adults are 100% accurate on FDS-4, 75% on FDS-7, and 25% on FDS-9.[1]

Column addition (CA-*x*): The participant is shown three numbers with *x* digits, and asked to add them using the column addition method. They are given 15 seconds to answer each problem. Participants enter the answer using the keyboard from left to right, upon which they are given correctness feedback (2-sec display message). There is a delay of 10 secs for CA-1 and 1 sec for CA-3 before the next problem appears. This task tests numerical manipulation and short-term working memory. Typical adults complete a CA-1 problem within 3 seconds and a CA-3 problem within 15 sec.

***n*-Back:** Participants are shown a continuing sequence English letters at 2.5 sec intervals with a .3 sec blank screen between each letter. They are asked to press a key on the keyboard when the current letter matches one that appeared *n* steps back. This task is commonly used to assess working memory. Typical adults are approximately 100% accurate on 1-back and 75% accurate on 3-back.

Tetris-*x*: Participants play a Tetris-style game. Difficulty (*x*) is varied by changing the level from 1 (easy) to 9 (difficult), which increases the speed of the pieces traversing the board by a factor of around 5. To ensure that participants' prior familiarity with Tetris does not confound the results, we used a game with non-traditional pieces made of hexagonal instead of square blocks. Participants typically were able to play comfortably at level 1, but lost the game within roughly 45 secs on level 9, after which the game automatically restarted.

Guru Intro: A short (~100 sec) session where the participant is first introduced to Guru's interface and the animated tutor. This task involves no significant mental activity, but was used to control for novelty effects.

Guru Train: Whereas the other training tasks, focus on easy and difficult versions of traditional cognitive load inducing tasks (e.g. involving digits and speed), the final training task was task-dependent. The task presented students with two short (60 sec) segments on the (unrelated) topic of Exponential Growth. The segments were similar to CGB Instruction and each was followed by an on-screen quiz. Participants first completed the easy version (Guru Train-E) followed immediately by the difficult version (Guru Train-D). The idea behind this training task was that the ITS itself might impose cognitive load demands that are not inherent in the more traditional tasks. Thus, mimicking easy and difficult versions of Guru might yield especially informative training data. Data Collection.

Eyes Train: Closed (EC): The participant holds eyes closed.

Eyes Train: Open (EO): The participant fixes gaze on a static target cross on the screen.

2.5 Participants

Twelve students were recruited from a high school in the U.S. under a protocol approved by Aspire IRB. Students, 5 males and 7

[1] FDS-7 was substituted with FDS-9 for two participants as the former was too easy for them.

females, were enrolled in 9th grade biology class in Fall 2014. Participation was voluntary and students received no classroom credit for their participation. The students were not informed of the study protocol and difficulty manipulations prior to the study.

To enroll in the study, parents and students attended an informational session and signed informed consent / minor assent forms prior to the study. Communication with students was coordinated with a local teacher, who facilitated student enrolment and data collection.

2.6 Design

The difficulty manipulation was implemented using a within-subjects design, where the easy and difficult versions of CGB Instruction and Scaffolded Dialogue were interleaved at the midpoint of each activity corresponding to each topic. For example, the first half of the CGB Instruction was presented using the easy version of the content followed by the difficult version in the second half for the first topic and the presentation order (i.e., difficult-easy) was reversed for the second topic. Similarly, the first half of the scaffolding prompts was true-false questions (easy) for the first topics and the second half open-ended (difficult). We used an interleaving rather than a blocking strategy (i.e., each for topic 1 and difficult for topic 2; and difficulty for topic 1 and easy for topic 2) to mitigate topic and fatigue effects. Order of difficulty (difficult-easy vs. easy-difficult) was also counterbalanced across topics, such that each participant received both orderings of difficulty. Topic order was counterbalanced across participants.

2.7 Procedure

The study lasted approximately 1.5 hours per student. Four students were sequentially tested per day. Data was collected in a quiet secluded classroom where there was very little interruption or distraction. The researchers and students were the only people present in the room during session.

An overview of the procedure is presented in Figure 4. The study began with basic setup. This included students donning the headset and making sure the students were comfortable while wearing it. The sensors were also checked to ensure they were working properly. Next, students completed all of the traditional training tasks followed by the Guru Intro and the Guru Train tasks. Students then completed the two Guru topics with a 10-minute break in between each topic. They were not required to spend the entire 20 minutes completing each topic, and were not cut off at any point before completing the topic. After the second topic, participants completed the EC and EO training tasks. These tasks were administered at the end because they were unlikely to suffer from fatigue effects compared to training tasks that required more attention and action. Students were then debriefed.

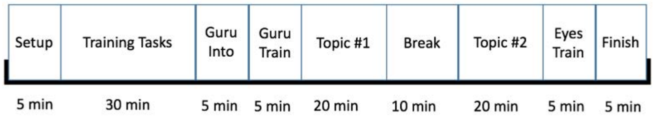

Figure 4. Overview of study procedure

Each Guru session began with a pre-test (~10 multiple choice questions about the topic). Then, students completed an easy and difficult block of CGB instruction and Scaffolding (see Design). At the end of each instruction and scaffolding block, students were prompted to provide subjective assessments of difficulty of the preceding material on a 6-point Likert scale. For example, during the lecture, the tutor would prompt the student to respond by asking, “How difficult are you finding this lecture so far on a scale of one to six? Six being very difficult.” A cloze task (e.g., fill in the blanks) followed the Scaffolding phase, but it did not include any difficulty manipulation and is not analyzed here. Finally, students completed a post-test on the topic they just completed.

3. MODEL BUILDING

EEG patterns are highly variable between individuals [5] and across days [42], so the models are not expected to generalize to new students. Instead, personalized models were constructed for each student using their respective training data. We tested four models that varied based on the training tasks used (see Table 2). Two models used Eyes Open (EO1 and EO2) in order to mimic instances when there would essentially be minimal cognitive load incurred. The corresponding high cognitive load models were either the combination of all other training tasks (EO1) or only the Guru Train tasks (EO2). The only differences between the two of the models is that Task2 includes the Guru Training tasks, while Task1 is complete domain-independent.

These models use an adaptive algorithm, called *Qstates*, for cognitive state classification (see [41] for a detailed overview of Qstates). The models use EEG spectral features to train models on the low and high cognitive load training tasks via partial least squares regression [41]. Model output was determined using a stratified k-fold (k = 6) cross-validation technique. Data epochs were randomized, with 60 secs used for model training and 30 secs for classification. This process was repeated six times until all the data was classified once. Using multivariate normal probability density functions (MVNPDF), the models produce a real-time measure of cognitive load by estimating the likelihood that a given 2 second epoch reflects low or high load. The output is normalized to have values that range from 0 (low) to 1 (high).

Table 2. Description of training tasks used in four models

Model	Low Cognitive Load	High Cognitive Load
EO1	Eyes Open	FDS-4, CA-1, 1-back, Tetris-1, Guru Train-E, FDS-7, CA-3, 3-back, Tetris-8, Guru Train-D, Guru Intro
EO2	Eyes Open	Guru Train-E and Guru Train-D
Task1	FDS-4, CA-1, 1-back, Tetris-1	FDS-7, CA-3, 3-back, Tetris-8
Task2	FDS-4, CA-1, 1-back, Tetris-1, Guru Train-E	FDS-7, CA-3, 3-back, Tetris-8, Guru Train-D

4. RESULTS

The cognitive load models made a prediction once every two seconds. Predicted levels of cognitive load are expected to change throughout the course of a Guru session rather than being exclusively in a low or high state. For example, load might be at high levels for 10 seconds when the tutor introduces a novel concept, but then return to a baseline level afterwards, before going back up again when a vaguely familiar concept is introduced. Thus, we analyze the data by aggregating the models’ predictions across periods corresponding to low or high difficulty predictions. A two-tailed significance criteria of .05 is adopted for all analyses.

4.1 EEG Data Validity

Can EEG yield viable data across a 90-min session? All 12 participants were able to wear the headset and reported no major issues during the learning session (an example of data from one participant is presented in Figure 5). However, simply wearing the headset does not automatically equate to having valid data – the electrodes must remain in contact with the student at all times. This headset (like many others) was originally designed for adults, so it was unclear if fit would be a major issue for some students. About 67% of the participants had head circumferences outside the specified range of the adult size headset used in this study, thereby some sensors did not make good contact on some locations. The sensor locations most affected were occipital (O1, O2) and temporal (T3, T4). Nevertheless, data from the parietal and frontal locations was largely reliable and of high quality.

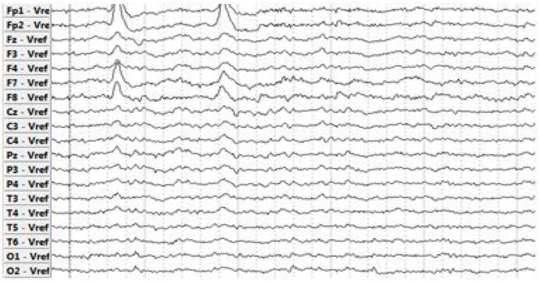

Figure 5. Example of EEG data collected from a student with two eye blinks evident on the top left

Students were required to wear the headset for almost 90 minutes, so it is possible that data quality might decline over time due to excessive movement, etc. Although we were able to collect EEG data from all 12 students, there were some unreliable predictions. For example, a model could predict 0 for an entire session for a number of reasons, including poor sensor connection or excessive movement. To account for this, a session was considered "invalid" if a model made no reliable predictions[2].

Despite some of the practical challenges, the EEG appears to a be a viable source of data during learning with an ITS, with validity rates of 91.7% (EO1); 87.5% (EO2); 79.2% (Task1); & 75% (Task2). Notably, the invalid sessions were not systematically the first or second session, so there is no concern of data lost due to session length.

We also assessed the reliability of model outputs across the two sessions by correlating average cognitive load across the two sessions. The correlation across the four models was .660, ranging from .474 (Task2) to .813 (Task1), suggesting that model output was relatively stable across sessions.

4.2 Subjective Reports

As a manipulation check, we compared the subjective ratings of difficulty across the easy and difficult versions of CGB Instruction and Scaffolded Dialogue (see Table 3).

Paired samples t-tests were conducted to compare participants' subjective difficulty ratings, which were averaged across the two sessions of Guru. There were no significant differences in difficulty ratings between the easy and difficult sections of CGB Instruction, though the trend was in the expected direction $t(11) = 1.48$, $p = .166$, $d = .260$. It is possible that the study was underpowered to detect this small effect. In contrast, easy and difficult sections of the Scaffolded Dialogue were significantly different, consistent with a medium to large effect, $t(11) = 3.63$, $p = .004$, $d = .716$. Thus, we can conclude that the manipulation was more effective for Scaffolded Dialogue than CBB Instruction.

Table 3. Descriptive statistics for perceptions of difficulty

	Easy	**Difficult**		
	M (SD)	*M (SD)*	*d*	*p*
Instruction	2.04 (.722)	2.21 (.582)	.260	.166
Scaffolding	2.75 (.754)	3.33 (.861)	.716	.004

4.3 Model Comparisons

We evaluated the four cognitive load models by comparing them across the easy and difficult sections of Guru (separately for the CGB Instruction and Scaffolding sections). A mixed-effects modeling approach [50] using the *lme4* package in R [6] was adopted for the analyses. This approach is appropriate because there are multiple observations per student with occasional missing data. For all models, *participant* was the random effect. The fixed effects were: *task difficulty* (easy vs. difficult), *time in the Guru session* (to account for fatigue effects, or electrode contact degradation) and *session number* (to account for differences across the two sessions). The results are shown in Table 4.

Table 4. Descriptive statistics for average model output

	Easy	**Diff**	**Linear Mixed Model**		
Model	*M (SD)*	*M (SD)*	*B*	*p*	*95% CI*
EO1					
Instruction	.643 (.237)	.646 (.264)	.013	.110	-.003,.029
Scaffolding	.644 (.198)	.717 (.187)	.055	.000	.040,.070
EO2					
Instruction	.749 (.147)	.734 (.134)	-.021	.791	-.017,.013
Scaffolding	.773 (.217)	.795 (.162)	.011	.080	-.001,.023
Task1					
Instruction	.318 (.176)	.335 (.281)	-.014	.140	-.034,.005
Scaffolding	.371 (.269)	.469 (.315)	.100	.000	.083,.116
Task2					
Instruction	.402 (.220)	.469 (.285)	.027	.012	.006,.049
Scaffolding	.457 (.267)	.537 (.313)	.078	.000	.061,.095

Notes. Descriptives computed at the participant level.

Despite the fact that students did not reliably perceive difficulty differences in the CGB phase of Guru, one of the models (Task2) was able to detect a significant difference. This suggests that students may not have been consciously aware of their increase in cognitive load imposed by the more complex language. The Task2 model notably differs from Task1 and both EO models by including the Easy and Difficult Guru training tasks, which most closely mimic the Instruction phase of Guru. Thus, inclusion of the domain-specific training tasks may have been crucial to detect the subtle difficulty manipulation in CBB Instruction.

Conversely, the main effect of difficulty during the Scaffolding phase was consistent in all four models (albeit marginally

[2] All analyses were re-computed using 100% of the data (including invalid sessions). The pattern of results remained exactly the same.

significant in EO2). Indeed, this corroborates the results with subjective reports of difficulty. There may have been a more distinguishable effect of difficulty in the Scaffolding phase, where true/false questions were juxtaposed with open-ended questions. Even the domain-independent Task1 model was able to pick up on these differences.

4.4 Scaffolding Performance

We examined students' answers to the tutor's questions during the Scaffolding section of Guru. Cognitive load was higher during the difficult scaffolding questions, and students rated them as being more difficult. Thus, we might expect performance to be lower on the difficult compared to easy questions.

Table 5. Linear mixed effects regressions predicting scaffolding performance from cognitive load and difficulty ratings.

Independent Variable	*B*	*p*	*95% CI*
EO1	-.255	.039	-.489, -.021
EO2	-.184	.245	-.487, .118
Task1	-.049	.721	-.034, .024
Task2	-.035	.804	-.311, .240
Difficulty Rating	-.118	.000	-.171, -.061

Easy (true/false) and Difficult (open-ended) questions were scored for correctness (0 = incorrect; 1 = correct). Proportion of correct answers was then computed for easy and difficult sections of Scaffolded Dialogue. After averaging across the two sessions, a paired-samples t-test revealed that students performed worse on the difficult questions (*M* = .589; *SD* = .121) compared to the easy ones (*M* = .771; *SD* = .152), consistent with a large effect size, *t*(11) = 4.89, *p* = .000, *d* = 1.33).

We conducted linear mixed effects regressions to assess if predicted cognitive load during Scaffolded Dialogue related to performance on the questions (see Table 5). Each participant contributed four data points: one for each easy and difficult section, across two topics. The averaged model output in each section was used as the independent variable and the proportion of correct answers in Scaffolded Dialogue was the dependent variable.

In all four models, predicted cognitive load was negatively related to performance on the Scaffolded Dialogue questions, but only the EO1 model was significant. The same mixed-effects approach revealed that participants' subjective difficulty ratings for a Scaffolded Dialogue section were negatively related to performance.

4.5 Examples of Model Output

Example output from the Task1 model on a single participant's sessions is shown in Figure 6. Tutorial modes and their difficulty are indicated by colored blocks (see legend), and white blocks indicate times during self-assessment questions. The order of difficulty is reversed in these two sessions (i.e. top: D→E, bottom: E→D). These two sessions clearly show that the workload model varies throughout the tutorial and between modes. From observations, there appears to be non-trivial activity during each of the tutorial blocks. For example, the workload model shows sharp changes at the onset and end of tutorial modes, as seen for example in Figure 5 (top) at the onset of difficult instruction (128 sec) or at the beginning of the post-test (1366 sec).

5. DISCUSSION

While some advanced learning technologies use generalized learner models to make inferences about student engagement and affect, they do so with peripheral rather than central measures of thought and feeling [4, 19, 32]. We take an initial step towards the goal using brain-based assessments of mental states to optimize instruct for individual students by assessing whether EEG-based estimation of students' cognitive load is sensitive to experimentally manipulated easy and difficult sections of an ITS. Our main findings are summarized below, followed by a discussion of applications, limitations, and future work.

5.1 Main Findings

First, we show that it is feasible to collect EEG data while students learn from an ITS, even for sessions spanning an hour and half. Further, we used a hat-like headset that sits comfortably on students' heads with 24 dry electrodes. Using dry electrodes is much more feasible for collecting EEG in ecological contexts compared to more laborious set up procedures used previously [22, 45, 48]. Additionally, students were relatively unconstrained and could move freely in their chair. Despite the potential problems that could occur (e.g., accidental electrode detachment, headset shifts, hardware issues, etc.), we were able to collect an average of 83% usable data across the two Guru sessions.

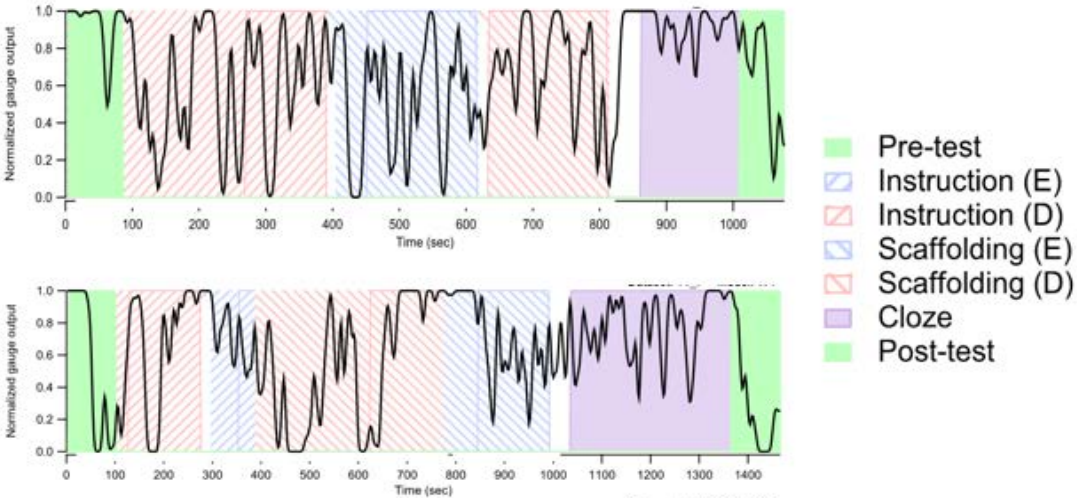

Figure 6. Cognitive load predictions for a single student over the course of two Guru sessions

We also found that manipulated difficulty had an effect on predicted levels of cognitive load in both the CGB Instruction (Task2 model) and Scaffolding (all models) phases of Guru. Moreover, the experimental manipulations mimicked real-life instructional differences in difficulty levels. For example, ITSs have the capability to use different levels of complexity in language, as well as employ a variety of dynamic Scaffolding techniques. Therefore, this study represents a relatively authentic manipulation of difficulty during a complex learning session.

We found that our cognitive load detectors were more highly attuned to the cognitive load differences during the Scaffolding phase (true/false vs. open-ended questions). Unlike the CGB Instruction phase, students also perceived that the open-ended questions were more difficult. Although the domain-independent Task1 was sensitive to manipulations of difficulty in the Scaffolding phase, it failed to do so in the CGB Instruction phase. The only model to successfully pick up on cognitive load differences (Task2) was trained on the easy and difficult Guru Training tasks. Thus, task-specific training data may be needed to detect the subtle differences in cognitive load, especially when humans are not as metacognitively aware of these differences.

Finally, we have shown that output from the cognitive load models might be negatively related to performance on the Scaffolding questions. Although the effect was only significant in one out of four models (EO1), a similar negative relationship was found between subjective ratings of difficulty and performance. Taken together, these findings show that EEG signal holds promise as an online indicator of students' cognitive load and can be used as an input modality to tailor instruction in a manner that is sensitive to load.

5.2 Limitations

There were several limitations of this work. We only collected data from a small number of 9th grade high school biology students (*N* = 12). Testing a larger, more diverse sample would allow for a better evaluation of the cognitive load models. Another limitation is that despite being conducted in a school, the study setup was more similar to a laboratory environment than to a typical classroom setting. Thus, future work should focus on implementing this type of data collection in a more ecological setting, like in the classroom. Finally, we only focused on performance during the Scaffolded Dialogue rather than on performance measured after the learning session. We did this to provide initial evidence that predicted cognitive load is sensitive to changes in real-time task performance, so that instructional strategies may eventually be individually tailored prior to the completion of the tutoring session. However, an important next step is to develop and test specific predictions about how prior knowledge and fluctuations in cognitive load throughout a learning session relate to learning gains measured via a posttest.

5.3 Future Work and Potential Applications

The ultimate goal of this work is to develop a system that can detect and dynamically respond to students' cognitive load with appropriate instructional materials. The present work needs to be extended in many ways to meet this goal. First, we manipulated two levels of difficulty in two sections of Guru. However, cognitive load may have a curvilinear relation with performance [49], where students learn the most at a moderate level of load. Future work should investigate additional levels of difficulty so as to more precisely map out the zone of optimal load.

Our manipulations also increased difficulty in only two ways: changing the complexity of language or changing the question format. Future work needs to address whether the detectors are sensitive to manipulations that induce cognitive load through other means, such as complexity of the materials (intrinsic load) or seductive details (extraneous load).

This work might also be extended by considering the temporal dynamics of cognitive load and how it relates to difficulty. We adopted a relatively course grained approach in analyzing the data, by averaging across easy and difficult sections. It would be valuable to take a more fine-grained approach that focuses on second-by-second changes in detected load. Another extension would be to employ a multisensor, multimodal approach to modeling students' cognitive states. For example, low cost eye-trackers have been used to model attention [32] could be combined with EEG-based cognitive load models. This might be helpful in determining whether low cognitive load is due to decoupling from the learning task because the student has zoned out, or due to some other factor.

Finally, future work could also focus on improving EEG methodology. We used 24 dry electrodes. However, it might be possible to detect cognitive load from fewer electrodes (seven should be sufficient [38, 41]), making the head-set even more scalable in the future. The head-set could also be tailored to fit heads with a smaller circumference.

6. CONCLUSION

Leveraging advances in intelligent learning environments, neuroscience, and in wearable EEG sensing, we provide initial evidence that EEG may be a viable method to track cognitive states during learning, thereby providing unique possibilities of closing the loop between the learning technology and the learner.

7. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) (IIP 1416595; DRL 1108845; IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

8. REFERENCES

[1] Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., & Silva, C. T. 2011. A user study of visualization effectiveness using EEG and cognitive load. *Computer Graphics Forum*, *30*(3), 791–800.

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

[3] Baker, R. S., Corbett, A., & Aleven, V. 2008. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer, 406-415.

[4] Baker, R. S. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York, NY: ACM, 1059–1068.

[5] Basile, L. F. H., Anghinah, R., Ribeiro, P., Ramos, R. T., Piedade, R., Ballester, G., & Brunetti, E. P. 2007. Interindividual variability in EEG correlates of attention and limits of functional mapping. *International Journal of Psychophysiology*, *65*(3) 238–251.

[6] Bates, D., Maechler, M., & Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models using lme4. *Joural of Statistical Software,* 67(1), 1-48.

[7] 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*(5) B231–B244.

[8] Bixler, R., & D'Mello, S. 2014. Toward Fully Automated Person-Independent Detection of Mind Wandering. In *Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization*. Cham, Switzerland: Springer International, 37–48.

[9] Bixler, R., & D'Mello, S. K. 2013. Detecting Boredom and Engagement During Writing with Keystroke Analysis, Task Appraisals, and Stable Traits. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces.* New York, NY:ACM, 225–234.

[10] Chang, K., Nelson, J., Pant, U., & Mostow, J. 2013. Toward Exploiting EEG Input in a Reading Tutor. *International Journal of Artificial Intelligence in Education*, *22*(1–2), 19–38.

[11] Chaouachi, M., & Frasson, C. 2010. Exploring the Relationship between Learner EEG Mental Engagement and Affect. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer, 291–293.

[12] Chaouachi, M., & Frasson, C. 2012. Mental Workload, Engagement and Emotions: An Exploratory Study for Intelligent Tutoring Systems. In *Proceedings of 11th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer, 65–71.

[13] Chaouachi, M., Jraidi, I., & Frasson, C. 2011. Modeling mental workload using EEG features for intelligent systems. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*. Berlin, Heidelberg: Springer, 50–61.

[14] Cirett Galán, F., & Beal, C. R. 2012. EEG Estimates of Engagement and Cognitive Workload Predict Math Problem Solving Outcomes. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*. Berlin, Heidelberg: Springer, 51–62.

[15] DeFalco, J., Baker, R. S., & D'Mello, S. K. 2014. Addressing Behavioral Disengagement in Online Learning. In R. Sottilare, A. C. Graesser, X. Hu, & B. Goldberg (Eds.), *Design Recommendations for Intelligent Tutoring Systems* (Vol. 2). Orlando, FL: U.S. Army Research Laboratory, 49–56.

[16] Derbali, L., & Frasson, C. 2010. Players' Motivation and EEG Waves Patterns in a Serious Game Environment. In *Proceeding of the 10th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer, 297–299.

[17] D'Mello, S.K., Blanchard, N., Baker, R. S., Ocumpaugh, J., & Brawner, K. 2014. I Feel Your Pain: A Selective Review of Affect-Sensitive Instructional Strategies. In R. Sottilare, A. C. Graesser, X. Hu, & B. Goldberg (Eds.), *Design Recommendations for Intelligent Tutoring Systems* (Vol. 2). Orlando, FL: U.S. Army Research Laboratory, 35-48.

[18] D'Mello, S. K., Chipman, P., & Graesser, A. C. 2007. Posture as a predictor of learner's affective engagement. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society.* Red Hook, NY: Curran Associates Inc., 905–910.

[19] D'Mello, S. K., & Graesser, A. C. 2013. AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4), 1–39.

[20] D'Mello, S. K., Olney, A., & Person, N. 2010. Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining*, *2*(1), 2–37.

[21] Estepp, J. R., Christensen, J. C., Monnin, J. W., Davis, I. M., & Wilson, G. F. 2009. Validation of a dry electrode system for EEG. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* New York, NY: SAGE Publications, 1171–1175.

[22] Ferree, T. C., Luu, P., Russell, G. S., & Tucker, D. M. 2001. Scalp electrode impedance, infection risk, and EEG data quality. *Clinical Neurophysiology*, *112*(3), 536–544.

[23] 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), 20–41.

[24] Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., & Rush, G. 1998. 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.

[25] Graesser, A. C., D'Mello, S. K., Craig, S. D., Witherspoon, A., Sullins, J., McDaniel, B., & Gholson, B. 2008. The relationship between affective states and dialog patterns during interactions with AutoTutor. *Journal of Interactive Learning Research*, *19*(2), 293–312.

[26] Graesser, A. C., & McNamara, D. S. 2011. Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*, *3*(2), 371–398.

[27] Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. 2011. Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, *40*(5), 223–234.

[28] Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

[29] Heraz, A., & Frasson, C. 2007. Predicting the three major dimensions of the learner's emotions from brainwaves. *International Journal of Computer Science*, *2*(3), 187–193.

[30] Heraz, A., & Frasson, C. 2009. Predicting Learner Answers Correctness through Brainwaves Assesment and Emotional Dimensions. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*. Amsterdam, Netherlands: IOS Press, 49–56.

[31] Hussain, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. 2011. Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. In *Proceedings of the*

15th International Conference on Artificial Intelligence in Education. Berlin, Heidelberg: Springer, 131–138.

[32] Hutt, S., Mills, C., White, S., Donnelly, P. J., & D'Mello, S. K. 2016. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. In *Proceedings of the 9th International Conference on Educational Data Mining,* US: International Educational Data Mining Society, 86–93.

[33] Jong, T. de. 2009. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, *38*(2), 105–134.

[34] Kalyuga, S. 2011. Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, *23*(1), 1–19.

[35] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (No. RBR-8-75)*. Millington, TN: Naval Technical Training Command.

[36] Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., … Kinces, W. 2007. Improving human performance in a real operating environment through real-time mental workload detection. In G. Dorhage et al. (Eds.), *Toward Brain-Computer Interfacing,* Cambridge, MA: MIT Press, 409–422.

[37] Matthews, R., McDonald, N. J., Anumula, H., Woodward, J., Turner, P. J., Steindorf, M. A., … Pendleton, J. M. 2007. Novel hybrid bioelectrodes for ambulatory zero-prep EEG measurements using multi-channel wireless EEG system. In *Proceedings of the 9th International Conference on Foundations of Augmented Cognition.* Berlin, Heidelberg: Springer, 137–146.

[38] Matthews, R., McDonald, N. J., Fridman, I., Hervieux, P., & Nielsen, T. 2005. The invisible electrode–zero prep time, ultra low capacitive sensing. In *Proceedings of the 11th International Conference on Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 22-27.

[39] Matthews, R., Turner, P. J., McDonald, N. J., Ermolaev, K., Mc Manus, T., Shelby, R. A., & Steindorf, M. 2008. Real time workload classification from an ambulatory wireless EEG system using hybrid EEG electrodes. In *Proceedings of the 2008 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. US: IEEE, 5871–5875.

[40] Mazzoni, G., & Cornoldi, C. 1993. Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General*, *122*(1), 47.

[41] McDonald, N. J., & Soussou, W. 2011. Quasar's qstates cognitive gauge performance in the cognitive state assessment competition 2011. In *Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. US: IEEE, 6542–6546.

[42] Mostow, J., Chang, K., & Nelson, J. 2011. Toward Exploiting EEG Input in a Reading Tutor. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. Berlin, Heidelberg: Springer, 230–237.

[43] Muldner, K., & Burleson, W. 2015. Utilizing sensor data to model students' creativity in a digital environment. *Computers in Human Behavior*, *42*, 127–137.

[44] Nelson, T. O. 1993. Judgments of learning and the allocation of study time. *Journal of Experimental Psychology. General*, *122*(2), 269–273.

[45] Nunez, P. L., Silberstein, R. B., Shi, Z., Carpenter, M. R., Srinivasan, R., Tucker, D. M., … Wijesinghe, R. S. 1999. EEG coherency II: experimental comparisons of multiple measures. *Clinical Neurophysiology*, *110*(3), 469–486.

[46] Olney, A. M., D'Mello, S., Person, N., Cade, W., Hays, P., Williams, C., … Graesser, A. 2012. Guru: A Computer Tutor That Models Expert Human Tutors. In *Proceedings of the 11th International Conference Intelligent Tutoring Systems.* Berlin, Heidelberg: Springer, 256–261.

[47] Olney, A., Person, N. K., & Graesser, A. C. 2012. Guru: Designing a conversational expert intelligent tutoring system. In C. Boonthum-Denecke, P. McCarthy, & T. Lamkin (Eds.), *Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches.* Hershey, PA, USA: IGI Global, 156–171.

[48] O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.

[49] Paas, F., Renkl, A., & Sweller, J. 2003. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4.

[50] Pinheiro, J. C., & Bates, D. M. (Eds.) 2000. *Mixed effects models in S and S-PLUS*. Berlin, Heidelberg: Springer Verlag.

[51] Slater, J. D., Kalamangalam, G. P., & Hope, O. 2012. Quality assessment of electroencephalography obtained from a "dry electrode" system. *Journal of Neuroscience Methods*, *208*(2), 134–137.

[52] Stevens, R., Galloway, T., & Berka, C. 2006. Integrating EEG models of cognitive load with machine learning models of scientific problem solving. In D. Schmorrow, K. Stanney, L. Reeves (Eds.), *Augmented Cognition: Past, Present and Future.* (Vol. 2). Arlington, VA: Strategic Analysis, Inc., 55–65.

[53] Sun, J. C.-Y., & Yeh, K. P.-C. 2017. The effects of attention monitoring with EEG biofeedback on university students' attention and self-efficacy: The case of anti-phishing instructional materials. *Computers & Education*, *106*, 73–82.

[54] Sweller, J., Ayres, P., & Kalyuga, S. (Eds.) 2011. Cognitive load theory (Vol. 55). New York, NY: Springer, 37–76.

[55] VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. 2007. When Are Tutorial Dialogues More Effective Than Reading? *Cognitive Science*, *31*(1), 3–62.

[56] VanLehn, K., Jordan, P. W., Rosé, C. P., … Srivastava, R. 2002. The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer, 158–167.