International Journal of Artificial Intelligence in Education 22 (2013) 19–38
DOI 10.3233/JAI-130033
IOS Press

Toward Exploiting EEG Input in a Reading Tutor

Kai-min Chang, *Carnegie Mellon University, PA, USA*
kaimin.chang@gmail.com
Jessica Nelson, *Carnegie Mellon University, PA, USA*
Udip Pant, *Brigham Young University, UT, USA*
Jack Mostow, *Carnegie Mellon University, PA, USA*

Abstract. A new type of sensor for students' mental states is a single-channel portable EEG headset simple enough to use in schools. To gauge its potential, we recorded its signal from children and adults reading text and isolated words, both aloud and silently. We used this data to train and test classifiers to detect a) when reading is difficult, b) when comprehension is lacking, and c) lexical status and word difficulty. To avoid exploiting the confound of word and sentence difficulty with length, we truncated signals to a uniform duration. The EEG data discriminated reliably better than chance between reading easy and difficult sentences. We found weak but above-chance performance for using EEG to distinguish among easy words, difficult words, pseudo-words, and unpronounceable strings, or to predict correct versus incorrect responses to a comprehension question about the read text. We also identified which EEG components appear sensitive to which lexical features. We found a strong relationship in children between a word's age-of-acquisition and activity in the Gamma frequency band (30–100 Hz). This pilot study gives hope that a school-deployable EEG device can capture information that might be useful to an intelligent tutor.

Keywords. EEG, Project LISTEN's Reading Tutor, frequency band, machine learning, lexical feature

INTRODUCTION

The ultimate automated tutor could peer directly into students' minds to identify their mental states (knowledge, thoughts, feelings, and so forth) and decide accordingly what and how to teach at each moment. The reality, of course, is that today's automated tutors attempt instead to infer students' mental states from a thin trickle of data, typically in the form of mouse clicks and keyboard input. Some ITS research (e.g., D'Mello et al., 2009; Gluck et al., 2000; Graesser et al., 2006; Mota & Picard, 2003; Woolf et al., 2009) has reported success in using other types of data, such as speech, eye movements, posture, heart rate, skin conductance, and mouse pressure, to detect various cognitive and affective states that may not be feasible to infer from conventional input, but could be of use to intelligent tutors. We report here on a complementary source of input from as close to the brain as non-invasively possible: electroencephalography (EEG).

The EEG signal is a voltage signal that can be measured on the surface of the scalp, arising from large areas of coordinated neural activity manifested as synchronization (groups of neurons firing at the same rate). This neural activity varies as a function of development, mental state, and cognitive activity, and the EEG signal can measurably detect such variation. Rhythmic fluctuations in the EEG signal occur within several particular frequency bands, and the relative level of activity within each frequency band has been associated with brain states such as focused attentional processing, engagement, and frustration (Berka et al., 2007; Lutsyuk et al., 2006; Marosi et al., 2002), which in turn are important for and predictive of learning (Baker et al., 2010). Studies of children's EEG power spectra have also identified specific frequency band differences between readers with different reading proficiencies (Ackerman et al., 1994, 1998; Clarke et al., 2002; Colon et al., 1979; Fein et al., 1983).

The recent availability of simple, low-cost, portable EEG monitoring devices now makes it feasible to take this technology from the lab into schools. The NeuroSky "MindSet," for example, is an audio headset equipped with a single-channel EEG sensor. It measures the voltage between an electrode that rests on the forehead and electrodes in contact with the ear. Unlike the multi-channel electrode nets worn in labs, the sensor requires no gel or saline for recording and therefore requires much less expertise to position. Even with the limitations of recording from only a single sensor and working with untrained users, a previous study (NeuroSky, 2009) found that the MindSet distinguished two fairly similar mental states (neutral and attentive) with 86% accuracy. MindSet has been used in brain-computer interfaces to classify flash patterns (Luo & Sullivan, 2010), human emotional responses (Crowley et al., 2010), and states of human awareness while driving (Yasui, 2009).

The ability to record longitudinal EEG data in authentic school settings is important for several reasons. First, we can analyze learning over intervals longer than a lab experiment, not just short-term memory effects. Second, we can study data generated by children's "*in vivo*" behavior at school, rather than their more constrained behavior under intense adult supervision in unfamiliar lab settings. Third, we can get enough data over a long enough time from enough students to potentially combat the notoriously noisy nature of EEG data with the statistical power of "big data," thereby enabling us to analyze the effects of different forms of instruction and practice on student learning and moment-to-moment engagement. Finally, longitudinal recording of EEG data on a school-based tutor offers the opportunity to obtain enough data over time to develop and train valid student-specific models and apply them on enough occasions to result in better learning.

To assess the feasibility of collecting useful information about cognitive processing and mental states using a portable EEG monitoring device, we conducted a pilot study. Participants wore a NeuroSky Mindset while using Project LISTEN's Reading Tutor (Mostow & Beck, 2007). The Reading Tutor displays text, listens to the student read aloud, and logs detailed moment-by-moment records of its multimodal tutorial dialogue to a database (Mostow & Beck, 2009). We matched this data to our EEG data by user ID and timestamp.

We wanted to know if MindSet data lets us distinguish among mental states relevant to learning to read. If we can do so better than chance, then these data contain relevant information that future work may decode more accurately. Thus we address four questions:

1. Can we use school-deployable EEG to detect when reading is difficult?
2. Can we use such EEG to detect when comprehension is lacking?
3. Can we use such EEG to detect properties of individual read words?
4. Which features of this EEG signal correspond to particular lexical properties?

We first describe the relationship between EEG and cognition. We next describe the study design and analysis methods. In the following sections, we address questions 1–4 above. We then discuss challenges, limitations, and future work, and finally conclude.

EEG AND COGNITION

The EEG signal is a continuous waveform sampled, in our case, 512 times per second. Much like a sound wave, the EEG waveform can be broken down into component frequencies and represented at each time point by the power in each frequency band. The EEG spectrum is typically clustered into 5 frequency bands: delta (1–3 Hz), theta (4–7 Hz), alpha (8–11 Hz), beta (12–29 Hz), and gamma (30–100 Hz). Changes in the synchronization of neural activity at these frequencies, resulting in changes in the measured power, are thought to be important in the control of cognitive processes and can vary as a result of attention,

alertness, memory, mental effort, motor responses, errors, and feedback. The exact nature of the relationship between cognition and EEG frequency can vary across cortical regions and as a function of the specific task requirements. Understanding which types of tasks or cognitive processes modulate activity in each of the EEG bands helps us interpret the meaning of relationships that we find between EEG activity and our tasks. In addition, it helps us identify appropriate features to use in training classifiers.

Generally, oscillations in the alpha and beta bands desynchronize (reducing power in these frequencies) as attention and task demands increase. Alpha and beta suppression have been found for reading words with attention compared to viewing distractor words that are to be ignored (Dalal et al., 2009), and with increasing task demands across a variety of tasks (Fink et al., 2005). Often concurrent with alpha and beta suppression, oscillations become more synchronous in the higher frequency gamma band, as well as the lower frequency delta and theta bands.

Increases in the gamma band are often related to linking perceived stimuli to both short-term and long-term memory, and therefore might be especially relevant during learning. For example, gamma power is higher for attended stimuli that are later recalled (Fell et al., 2001; Mainy et al., 2007), during perceptual learning (e.g. learning pictures which later must be identified in a fragmented form) (Gruber et al., 2002), when reading target words needing attention in a reading task (Dalal et al., 2009), and for viewing images of real objects whose identities are presumably represented in long-term memory rather than novel objects (Herrmann et al., 2004). These increases are often found in occipital brain regions normally associated with the visual-perceptual tasks, but Lachaux et al. (2008) found that in a reading task, when gamma band activity increased in the reading network, it also decreased in other regions. They suggest that gamma band activity can both synchronize and desynchronize during tasks, depending on the cortical region. However, we record from only one location on the scalp, which doesn't measure activity across the whole brain. Consequently we may not always see EEG power change in the expected direction.

Synchronous delta oscillations increase during motor responses where errors are made or in conditions where errors are likely to be made and may reflect error-specific processing (Cavanagh et al., 2012). In addition, delta power increases with increased internal processing during harder mental tasks (Harmony et al., 1996).

Theta oscillations also increase with novel stimuli, conflict, punishment, and error, and may be a good measure of the sustained shift that underlies several well-known transient signals marking responses to errors, correct responses, and feedback – all important elements of learning and performance (Cavanagh et al., 2012). Increased theta in one study was related to decreased behavioral alertness, defined as the times when participants were making more errors (Huang et al., 2001). However, this increase could also be a result of error-related processing rather than alertness. Howells et al. (2010) also found a negative relationship between perceived mental effort and theta and the theta/beta ratio (decreased theta with increased mental effort), but only in certain types of attentional tasks – theta was generally also lower during rest than during tasks requiring attention.

At the finest grain size, the EEG signal can be time-locked to a certain "event," such as the presentation of a word, and the resulting Event-Related Potential (ERP) can be plotted as a raw waveform with recognizable peaks and dips ("components") following certain kinds of "events." These components are named by whether they are a positive deflection (P) or a negative deflection (N), and by how long, in milliseconds, they occur after the event. Studies in carefully controlled laboratory conditions have detected item recognition (P300) (Picton, 1992), surprise/inconsistency (N400) (Kutas & Federmeier, 2000; Kutas & Hillyard, 1980), and violations of expectations (P600), especially in syntax (Coulson et al., 1998; Gouvea et al., 2010). To illustrate, for example, the N400, one can expect to see a negative deflection in the waveform approximately 400 ms after seeing the word "pizza" in the sentence "The girl drank the pizza." However, a reader who does not know the meaning of the word "pizza" should not be sensitive to such inconsistency when the word occurs in an unexpected context. Although it is not clear that such precise components can be detected in less

controlled settings, especially since the waveforms are usually averaged over many trials, the ability to do so might tell an intelligent tutor whether a student recognized a taught word, spotted a new word, or noticed a misspelling or textual inconsistency. Immediate, unobtrusive detection of these cognitive states in place of overt, time-consuming questions to assess student knowledge could potentially enable order-of-magnitude speedups over conventional cycles of teaching, learning, and testing.

In short, previous work has demonstrated that lab-quality EEG can detect many mental states relevant to intelligent tutors. We wanted to know whether a much simpler consumer-quality portable single-channel EEG device that schools could afford and operate could detect any such states.

METHODS: STUDY DESIGN AND ANALYSIS METHODS

We used a within-subjects design to compare the EEG signal during easy vs. difficult reading, at both the passage and single item level, during both oral and silent reading. Our approach is analogous in some ways to some of the earliest efforts to use EEG in intelligent tutors, by Frasson et al. They used EEG to model learners' reactions in ITS (Blanchard et al., 2007), detect learners' emotions (Heraz & Frasson, 2007; Heraz et al., 2007), assess learners' attention (Derbali et al., 2011; Derbali & Frasson, 2011), and more recently to show that subliminal cues were cognitively processed and have positive influence on learners' performance and intuition (Chalfoun & Frasson, 2010, 2011, 2012; Chaouachi et al., 2010; Jraidi et al., 2012). However, it appears that some of this previous work may have inflated classifier accuracy by allowing statistical dependencies between training and test sets. The study reported here also differs in that it focused on detecting the cognitive workload relevant to learning to read and that it used a low-cost, single-channel EEG sensor feasible to deploy in schools.

Study design

We implemented our experimental protocol in the Reading Tutor's language for scripting interactive activities. The Reading Tutor logs a detailed stream of timestamped information that can be linked to EEG records, including word reading times, accuracy (i.e. acceptance by its automatic speech recognizer), clicks for help, answers to comprehension questions, and voice recordings. In this study, the Reading Tutor displayed passage excerpts to read aloud – three easy and three difficult – in alternating order. The "easy" passages were from texts classified at the K-1 level by the Common Core Standards Text Exemplars, Appendix B (www.corestandards.org). The "difficult" passages came from practice materials for the Graduate Record Exam (majortests.com/gre/reading_comprehension.php) and the American Council on Education's General Equivalency Diploma (GED) test (majortests.com/gre/reading_comprehension.php). Each passage was followed by a multiple-choice cloze question (formed by deleting a word from the next sentence in the passage) to ensure that readers were reading for meaning. The protocol then repeated these tasks in a silent reading condition, using different texts. Across the read-aloud and silent reading conditions, passages ranged from 62 to 83 words long.

11 nine- and ten-year-olds participated at their school, and 10 adult readers participated in our laboratory. A few other participants user-tested the protocol, but without EEG data. Figure 1 illustrates the setup for a Reading Tutor session, in which a participant read the text displayed on the screen while wearing an EEG headset. We excluded 2 children and 4 adults due to missing or poor-quality data. We analyzed data for the remaining 15 readers, both overall and separately for the 9 children and 6 adults. Although the study was initially intended as a pilot study, the number of subjects proved sufficient to obtain the results reported here.

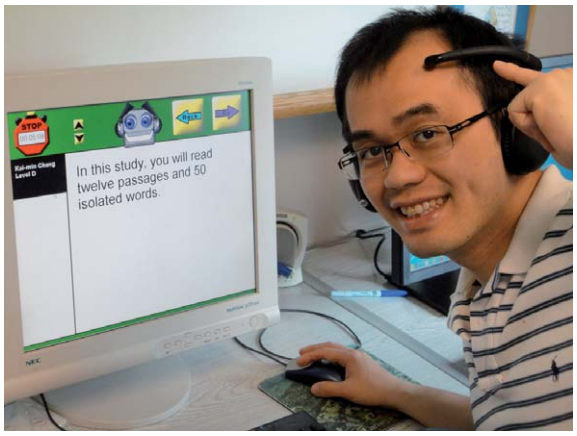

Fig. 1. First author shows setup for using the Reading Tutor while wearing EEG headset.

EEG data collection and artifact removal

The participants interacted with the Reading Tutor while wearing a wireless single-channel MindSet that measured activity over the frontal lobe. The MindSet measures the voltage between an electrode resting on the forehead and two electrodes (one ground and one reference) each in contact with an ear. More precisely, the position on the forehead is Fp_1 (somewhere between left eye brow and the hairline), as defined by the International 10–20 system (Jasper, 1958). We used NeuroSky's NeuroView software to collect the following signal streams:

1. The raw EEG signal, sampled at 512 Hz.
2. A filtered version of the raw signal, also sampled at 512 Hz.
3. An indicator of signal quality, reported at 1 Hz.
4. MindSet's proprietary "attention" and "meditation" signals said to measure the user's level of mental focus and calmness, reported at 1 Hz.
5. A power spectrum, reported at 8 Hz, clustered into the standard named frequency bands.

Figure 2 shows a sample Reading Tutor interaction, where the student is asked to read stories, answer corresponding comprehension questions, and finally read some words. We used the timestamped Reading Tutor logs to split the EEG time series into segments labeled by the type of stimulus (e.g. easy vs. difficult) or response (e.g. correct or incorrect). The Reading Tutor and MindSet log signals at different time scales, so we performed linear interpolation between successive EEG values (using Matlab's built-in "interp1" function) to estimate their values at the points timestamped by the Reading Tutor.

NeuroSky claims to detect frontal lobe EEG activity, but other sources could introduce artifacts or noise into the recorded signals. In particular, facial expressions, eye blinks, and muscle movement elsewhere in the body generate electromyographic (EMG) signals. To remove potential EMG artifacts, Neurosky applies a 3–100 Hz band-pass filter to the raw EEG signal to remove frequencies below 3 Hz and over 100 Hz, which are known to be related to EMG, as well as a Notch filter that eliminates electrical noises from the power source, which varies from 50 to 60 Hz depending on the geographical location. The corresponding filtered signal and power spectrum are based on these denoised raw signals. We used two additional techniques to further mitigate noise. First, we used soft thresholding with wavelets to denoise the signals (Donoho, 1995). The wavelet transform provides a time-frequency decomposition shown to be suitable for EEG/ERP

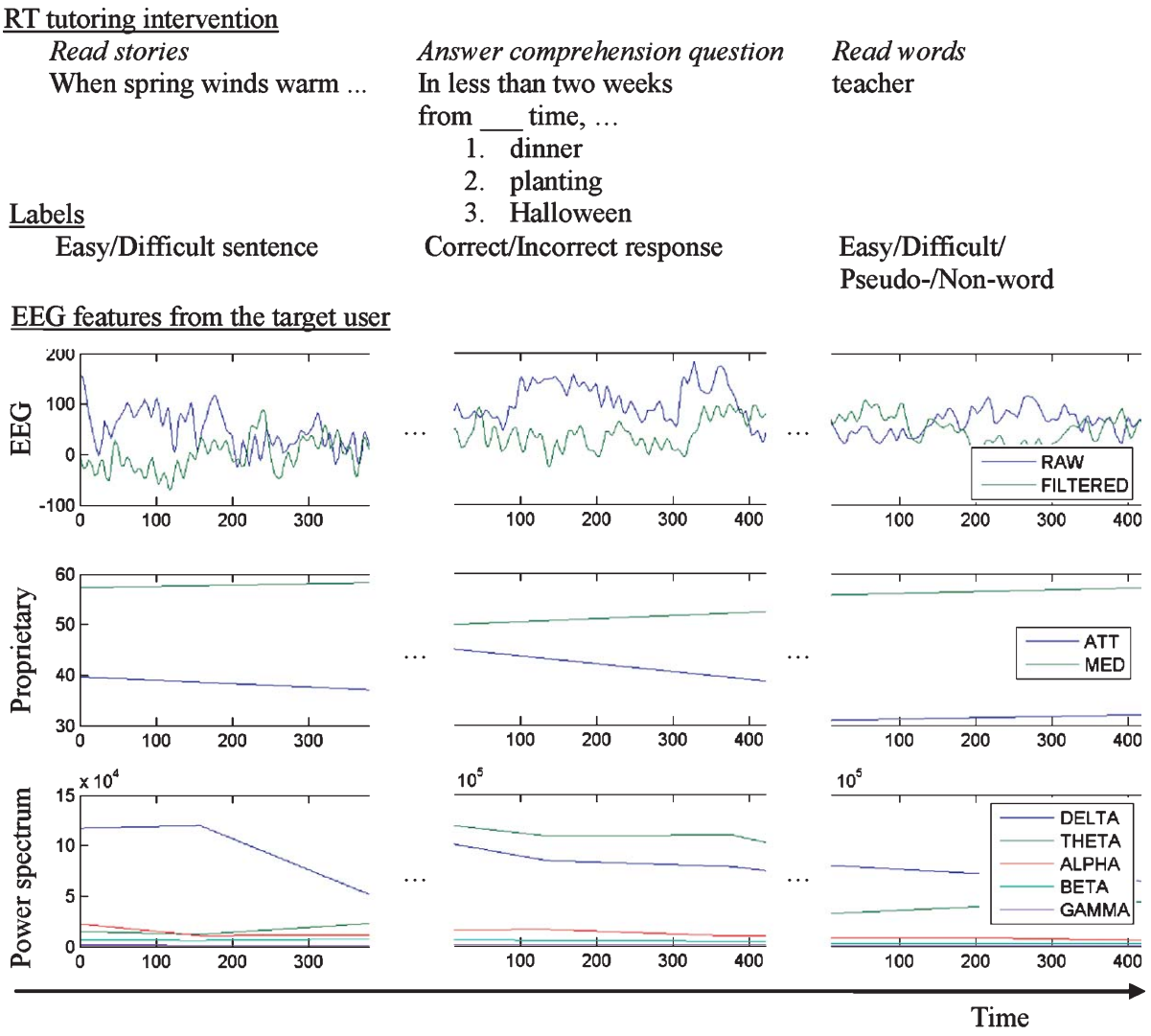

Fig. 2. Sample Reading Tutor interactions (top), classifier labels (middle), and EEG signals (bottom). RAW (or FILTERED) is the average power of the entire raw (or filtered) EEG signal during an utterance; ATT and MED are proprietary "attention" and "meditation" indicators; and DELTA, THETA, ALPHA, BETA, and GAMMA are the average power of the successive named frequency bands.

analysis (Bartnik et al., 1992; Bertrand et al., 1994; Demiralp et al., 1999). Second, we used Neurosky's proprietary signal quality indicator to exclude any signals affected by poor electrode contact due to muscle movement. We analyzed only utterances at least 50% of whose samples had good reported signal quality (represented by a value of 0).

Figure 3 shows the averaged magnitude of various EEG features in a) easy vs. difficult sentences, b) correct vs. incorrect responses to comprehension questions *while they were reading the preceding passage*, c) correct vs. incorrect responses to comprehension questions *while they answered the question*, and d) easy, difficult, pseudo- and non-words. To compare the relative contribution across EEG features, we normalized all features as z-scores. As expected, the difficult sentences and words are associated with higher attention

levels, lower meditation levels, lower magnitudes in low frequency bands and higher magnitudes in high frequency bands. Moreover, correct responses are associated with less attention and higher meditation levels while reading the preceding passage, but higher attention and meditation levels while answering the question. The fact that the different classes exhibit different EEG profiles suggested that it would be feasible to train machine learning classifiers to decode mental states of individual participants.

Training classifiers

We trained Gaussian Naïve Bayes classifiers to estimate based on EEG data the probability that a given stimulus (sentence, word, or comprehension question) was easy rather than difficult, or correct rather than incorrect. We chose this method (rather than, say, logistic regression) because it is generally best for problems with sparse (and noisy) training data (Ng & Jordan, 2002).

The EEG device emits the various signals enumerated earlier, including the raw and filtered EEG signal, proprietary "attention" and "meditation" indicators, and the successive named frequency bands. To characterize their overall values, we computed their means over the interval of each utterance. To characterize the temporal profile of the EEG signal, we computed several features, some of them typically used to measure the shape of statistical distributions rather than of time series:

- The minimum and maximum describe the range of each signal (but are vulnerable to noise).
- The variance measures the amount and magnitude of variation in the signal.
- The linear fit coefficient measures its overall slope.
- The quadratic fit coefficient measures its overall curvature.
- The skewness measures its asymmetry.
- The kurtosis measures its peakiness.

We trained separate classifiers for each condition (oral and silent reading) and group (children and adults), and also classifiers for data pooled across both conditions and groups. We trained and tested two types of classifiers for each classification task.

We trained *reader-specific* classifiers on a single reader's data from all but one stimulus block (e.g. one story), tested on the held-out block (e.g., all other stories), performed this procedure for each block, and averaged the results to cross-validate accuracy within reader. Cross-validating across blocks avoids improperly exploiting statistical dependencies (e.g. temporal continuity) between observations of a reader on successive stimuli within the same block (e.g., sentences in the same story).

We trained *reader-independent* classifiers on the data from all but one reader, tested on the held-out reader, performed this procedure for each reader, and averaged the resulting accuracies to cross-validate across readers. We averaged each feature over the time interval of each stimulus, excluding the 15% of observations with poor signal quality.

We computed *classification accuracy* as the percentage of cases classified correctly, where chance performance is one over the number of categories. For the four-way classification, we evaluated *rank accuracy* as the average percentile rank (normalized between 0 and 100) of the correct category if categories are ordered by the value of the regression formula; chance performance is 50%. Rank accuracy is a more sensitive criterion than classification accuracy for evaluating performance on multi-category tasks such as decoding mental states from brain data (Mitchell et al., 2004).

To test whether a classifier was significantly better than chance, we first computed its overall accuracy for each reader, yielding a distribution of N accuracies, where N is the number of readers. Treating this distribution as a random variable, we performed a one-tailed T-test of whether its mean exceeded chance performance for the classification task in question. Counting N readers rather than observations is con-

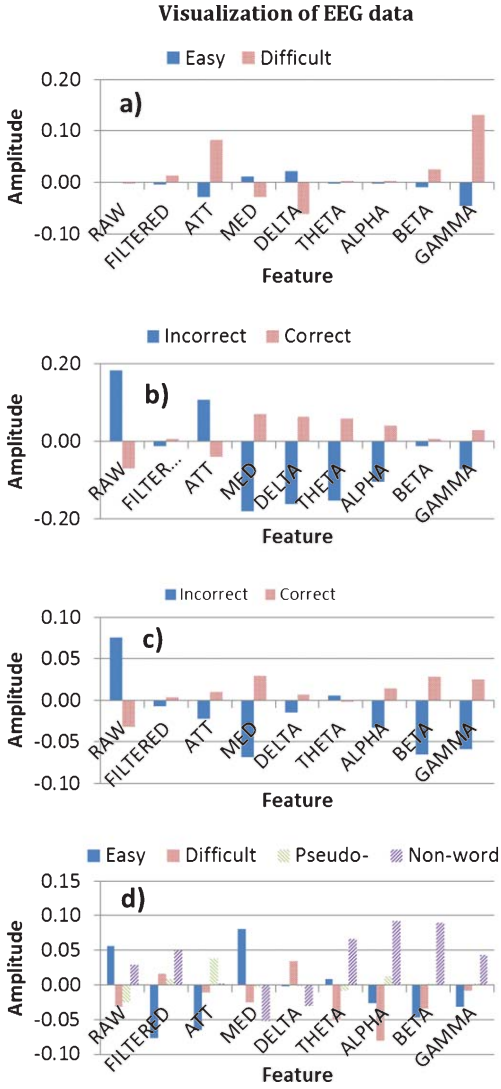

Fig. 3. The magnitude of various EEG features during a) easy vs. difficult sentences, b) correct vs. incorrect responses to comprehension questions *while they read the preceding passage*, c) correct vs. incorrect responses to comprehension questions *while they answer the question*, and d) easy, difficult, pseudo-, and non-words. RAW (or FILTERED) is the average power of the entire raw (or filtered) EEG signal during an utterance; ATT and MED are proprietary "attention" and "meditation" indicators; and DELTA, THETA, ALPHA, BETA, and GAMMA are the average power of the successive named frequency bands.

servative in that it accounts for statistical dependencies among observations from the same reader. Our significance criterion was $p < 0.05$, without correction for multiple comparisons.

Two problems in training classifiers are class size imbalance and the confound between the length and difficulty of the stimuli. We face these issues because easy stories tend to be composed of many short sentences (about 10 on average), whereas difficult stories tend to be composed of a few long sentences (about 4). As a result, we have more easy sentences than difficult ones, and easy sentences are shorter than difficult ones.

Sentence length was not an obvious confound, since we did not include length or duration as an explicit feature. This issue eluded us and the reviewers until we realized late in the process of preparing this article that some of our feature values might correlate with duration. For the goal of detecting cognitive effort or difficulty based solely on EEG, independent of text features such as sentence length or behavioral features such as sentence duration, encoding information about such features in the test data constitutes "cheating." Even innocuous-looking features may encode such information. For instance, the minimum and maximum of a signal as noisy as EEG are likely to correlate with its duration.

Indeed, re-segmenting passage readings into equal-duration units (ignoring sentence boundaries) resulted in worse classification accuracy than using sentences as units. Truncating sentences to equal-duration units fared somewhat better than re-segmenting, despite having less data, presumably because readers' EEG signals have temporal structure related to sentence boundaries. Therefore, to address the sentence duration issue, we truncate all stimuli to a uniform duration (5 seconds for sentence stimuli and 1 second for word stimuli). Thus a difficult sentence that lasts 8 seconds and an easy sentence that lasts 6 seconds are both truncated to 5 seconds, and all isolated words are truncated to 1 second.

Note that this conservative approach differs from our earlier work (Mostow et al., 2011) by reframing the question as "can we classify sentence difficulty *based on the first 5 seconds of reading it*?" We excluded from our data set – both training and test data – the 57% of the sentences shorter than 5 sec (as well as the 9% of the isolated word recordings shorter than 1 sec). Thus the results reported here on the reduced set are not strictly comparable to our original results nor to the results (mentioned above but not detailed) on the larger data set constructed by segmenting the read passages into 5-second segments regardless of sentence boundaries.

As for the problem of class size imbalance, a common solution is to resample the training data to obtain equal-size sets of training data. However, "random undersampling can potentially remove certain important examples [and underestimate performance]; and random oversampling can lead to overfitting [and overestimate performance]" (Chawla et al., 2004). To avoid bias due to class size imbalance, we originally (Mostow et al., 2011) employed three different resampling methods: random oversampling of the smaller class(es), with replacement; random undersampling of the larger class(es), without replacement; and directed undersampling, in our case by truncating the larger class to the temporally earliest k examples. An adaptive tutor would use such temporal truncation to train user-specific models on each user's initial data. For this article we used random undersampling of the larger class(es), which yielded medium performance (a conservative measure) in the original study among the three random resampling methods. We performed the sampling 10 times to limit the influence of particularly good or bad runs and obtain a stable measure of classifier performance.

Resampling addresses class size imbalance in training data but not in test data. Resampling test data would unfairly distort performance, so we report overall cross-validated accuracy on the test data. However, class size imbalance inflated it, yielding a numerically correct but misleading measure of accuracy. We therefore also report within-class accuracy, which is independent of class size.

CAN WE USE EEG TO DETECT WHEN READING IS DIFFICULT?

We trained Gaussian Naïve Bayes classifiers to distinguish between easy and difficult sentences read aloud, silently, or both, for all participants, just for children, and just for adults. Table 1 shows the results; overall accuracy in **bold** here and later is significantly better than chance. These results differ substantially from those in (Mostow et al., 2011) due to bug fixes, wavelet filtering, other improvements, and truncating to uniform duration. Depending on the condition and group, overall accuracy of reader-specific classifiers ranged from 39% to 58% for (2 of 9 significantly above 50%). (The 9 cases are for different but overlapping

Table 1
Accuracy in classifying sentences from easy vs. difficult text. The columns 'Overall', 'Easy', and 'Difficult' respectively show cross-validated accuracy on the entire data set (in **bold** if significantly above chance) and on easy and difficult sentences (highlighted if over 50% for both)

	Condition	Number of cases [Easy, Difficult]	Reader-specific			Reader-independent		
			Overall	*Easy*	*Difficult*	Overall	*Easy*	*Difficult*
Overall	Overall	[314, 271]	0.51	*0.49*	*0.51*	**0.56**	*0.48*	*0.64*
	Oral	[237, 149]	0.51	*0.47*	*0.52*	**0.53**	*0.51*	*0.51*
	Silent	[77, 122]	0.51	*0.56*	*0.45*	0.49	*0.45*	*0.48*
Children	Overall	[235, 157]	0.47	*0.47*	*0.48*	0.50	*0.41*	*0.62*
	Oral	[167, 80]	0.45	*0.44*	*0.45*	0.42	*0.37*	*0.50*
	Silent	[68, 77]	0.50	*0.48*	*0.46*	0.48	*0.45*	*0.45*
Adults	Overall	[79, 114]	**0.58**	*0.54*	*0.57*	**0.58**	*0.54*	*0.59*
	Oral	[70, 69]	**0.56**	*0.50*	*0.56*	**0.60**	*0.67*	*0.48*
	Silent	[9, 45]	0.39	*0.40*	*0.39*	0.50	*0.53*	*0.49*

data sets, so they are not statistically independent tests of our approach.) Accuracy was higher for reader-independent classifiers, ranging from 42% to 60% (4 of 9 significant), suggesting that imperfect transfer across readers was outweighed by the advantage of training on more data, much as in classifying fMRI brain images (Mitchell et al., 2004).

These results seem promising, but are they meaningful, or merely artifacts of the considerable class size imbalance between the number of easy and difficult cases? Within-class accuracy is a more rigorous measure. Accuracy of reader-specific classifiers ranged from 40% to 56% on easy cases and 39% to 57% on difficult cases. Reader-independent classifiers performed better here as well. Their within-class accuracy ranged from 37% to 67% on easy cases and 45% to 64% on difficult cases. However, only one reader-specific classifier and two reader-independent classifiers are over 50% accurate within both classes. Although these results are significantly above chance, EEG-based mental state detectors will presumably need to be much more accurate to help intelligent tutors.

CAN WE USE EEG TO DETECT READING COMPREHENSION?

After each text, our protocol displayed a comprehension question constructed from the next sentence in the story by omitting one of the key words and replacing it with a blank to form a cloze question. The participant had to choose among three choices to complete the sentence – the omitted word as the correct response, and two distractors. Note that this pilot study did not explore the effect of distractors. It is possible that participants could infer the correct answer from common knowledge – especially adult readers on easy passages. A more informative alternative would use different types of distractors to detect different types of comprehension failure (Mostow & Jang, 2012).

We trained Gaussian Naïve Bayes classifiers to predict correct versus incorrect responses to the comprehension questions, based on participants' EEG data collected *while they read the preceding passage*. Table 2 shows the results. Depending on the condition and group, accuracy for reader-specific classifiers averaged from 25% to 53% overall, none of them significantly better than chance, 20% to 67% on correct responses, and 20% to 63% on incorrect responses. We attribute the absence of significant results for

Table 2
Accuracy in predicting correct vs. incorrect responses to comprehension questions using EEG *while they read the preceding passage*. The columns 'Overall', 'Correct', and 'Incorrect' respectively show cross-validated accuracy on the entire data set (in bold if significantly above chance, $p < .05$) and within-class on incorrect and correct responses. "NaN" means "Not a Number," representing undefined results caused by sparsity

	Condition	Number of cases [Incorrect, Correct]	Reader-specific			Reader-independent		
			Overall	*Correct*	*Incorrect*	Overall	*Correct*	*Incorrect*
Overall	Overall	[56, 143]	0.41	*0.40*	*0.41*	0.38	*0.35*	*0.45*
	Oral	[28, 79]	0.35	*0.27*	*0.43*	0.48	*0.62*	*0.33*
	Silent	[28, 64]	0.48	*0.54*	*0.41*	0.46	*0.34*	*0.57*
Children	Overall	[36, 72]	0.39	*0.40*	*0.37*	0.41	*0.40*	*0.44*
	Oral	[20, 35]	0.25	*0.30*	*0.20*	0.35	*0.59*	*0.12*
	Silent	[16, 37]	0.53	*0.67*	*0.38*	0.48	*0.38*	*0.58*
Adults	Overall	[20, 71]	0.51	*0.49*	*0.54*	0.41	*0.32*	*0.64*
	Oral	[8, 44]	NaN	*NaN*	*NaN*	NaN	*NaN*	*NaN*
	Silent	[12, 27]	0.42	*0.20*	*0.63*	**0.65**	*0.33*	*0.97*

reader-specific classifiers to the small sample size caused by asking only one question per text. As before, accuracy was higher for reader-independent classifiers, with overall accuracy ranging from 35% to 65% for adult silent reading, the only one significantly above chance – compared to 4 of 9 for distinguishing easy from difficult sentences, where we had much more training and test data. Moreover, none of the classifiers achieves accuracy above 50% within both classes.

Low accuracy in predicting performance on comprehension questions may well be due to label noise. Consider how we classified text difficulty compared to how we measured comprehension. Our classification of each passage as easy or difficult is reliable in that we chose very easy and much more difficult passages. This classification applies across the passage insofar as sentences in an easy passage tend to be easy, and sentences in a difficult passage tend to be difficult. In contrast, comprehension may fluctuate over the course of a passage, succeeding on some sentences but failing on others. A single question is bound to test comprehension of some sentences better than others. Moreover, a random answer to a 3-choice question has a 33% probability of being correct, which renders our passage comprehension measure even noisier. A more reliable comprehension measure would require asking more questions, with answers that can't be guessed without comprehending the text.

Moreover, we trained Gaussian Naïve Bayes classifiers to predict correct versus incorrect responses to the comprehension questions, based on participants' EEG data collected *while they answer the question*. Table 3 shows the results. Depending on the condition and group, accuracy for reader-specific classifiers averaged from 37% to 55% overall (2 significant), 30% to 57% on correct responses, and 40% to 72% on incorrect responses. We attribute the absence of significant results for reader-specific classifiers to the small sample size caused by asking only one question per text. As before, accuracy was higher for reader-independent classifiers, ranging from 37% to 70% overall (2 significant), 10% to 67% on correct responses, and 49% to 74% on incorrect responses. Two reader-specific and two reader-independent classifiers exceeded 50% accuracy within both classes.

The ability to gauge readers' comprehension in real-time based on their EEG would provide more data and take less student time than interrupting occasionally to ask comprehension questions. Improved performance on this task will likely require not only more accurate classification, but a more accurate

Table 3
Accuracy in predicting correct vs. incorrect responses to comprehension questions using EEG *while they answer the question*. The columns 'Overall', 'Correct', and 'Incorrect' respectively show cross-validated accuracy on the entire data set (in bold if significantly above chance, $p < .05$) and within-class on incorrect and correct responses (highlighted if over 50% for both)

	Condition	Number of cases [Incorrect, Correct]	Reader-specific			Reader-independent		
			Overall	*Correct*	*Incorrect*	Overall	*Correct*	*Incorrect*
Overall	Overall	[60, 140]	**0.53**	*0.51*	*0.60*	0.48	*0.39*	*0.64*
	Oral	[31, 79]	0.43	*0.30*	*0.57*	0.38	*0.27*	*0.50*
	Silent	[29, 61]	0.49	*0.30*	*0.68*	**0.62**	*0.56*	*0.69*
Children	Overall	[38, 69]	**0.55**	*0.57*	*0.52*	0.49	*0.41*	*0.64*
	Oral	[21, 35]	0.38	*0.30*	*0.47*	0.37	*0.10*	*0.63*
	Silent	[17, 34]	0.54	*0.36*	*0.72*	**0.70**	*0.67*	*0.74*
Adults	Overall	[22, 71]	0.52	*0.45*	*0.67*	0.52	*0.53*	*0.49*
	Oral	[10, 44]	NaN	*NaN*	*NaN*	NaN	*NaN*	*NaN*
	Silent	[12, 27]	0.37	*0.33*	*0.40*	0.52	*0.33*	*0.70*

gold standard to validate against than students' performance on a single multiple choice question they can answer correctly with probability 1/3 merely by guessing at random.

CAN WE USE EEG TO DETECT LEXICAL STATUS?

After the read-aloud portion of the passages and comprehension questions, our protocol displayed 10 English words followed by 10 pronounceable pseudo-words to read aloud. After the silent-reading portion of the passages and comprehension questions, it displayed 10 words, 10 pseudo-words, and 10 unpronounceable consonant strings to read silently.

Real words were all 2-syllable 7-letter words; half were easy and half were difficult, and they were presented with alternating difficulty. The easy words occur frequently in text, with a Kucera-Francis (K-F) written frequency count of 30 or more (mean = 84), whereas the difficult words occur infrequently, with K-F written frequencies below 10 (mean = 3.4). The easy words are also learned earlier than the difficult words, with an age-of-acquisition (AOA) below 315 on a scale from 100 -700 (mean = 254.4, corresponding to approximately age 4); the AOA for difficult words was above 450 (mean = 555.5, corresponding to approximately age 10) (Coltheart, 1981).

Just as we used difficult text to see if we could detect when a reader is having difficulty reading, we included non-words to see if we could detect when a reader notices that a word is unfamiliar. Pseudo-words and illegal strings were 3 letters long and chosen to vary in their number of orthographic neighbors (words that differ in spelling by only one letter). These stimuli came from a study by Laszlo & Federmeier (2011) showing that ERPs are sensitive to neighborhood size. Pseudo-words were pronounceable and legal according to English orthography. In contrast, the illegal strings were unpronounceable, and therefore omitted from the read-aloud portion. We varied the orthographic neighborhood size of both types of non-words from 0 neighbors to 22 neighbors, to permit future analysis of its effects.

We trained and evaluated classifiers just as described previously, except that we trained multinomial Gaussian Naïve Bayes classifiers to estimate from EEG data the probability that a word was easy, difficult,

Table 4
Rank accuracy (chance = 50%) in classifying lexical stimuli. Column 'O' shows cross-validated rank accuracy on the entire data set (in bold if significantly above chance, $p < .05$). Columns 'E', 'D', 'P', and 'I' show within-class rank accuracy on easy, difficult, pseudo-, and illegal words (highlighted if over or at 50% for all four)

	Condition	Number of cases [Easy, Difficult, Pseudo-, Illegal]	Reader-specific					Reader-independent				
			O	*E*	*D*	*P*	*I*	O	*E*	*D*	*P*	*I*
Overall	Overall	[134, 137, 269, 114]	**0.55**	*0.56*	*0.57*	*0.49*	*0.66*	0.48	*0.43*	*0.56*	*0.46*	*0.46*
	Oral	[81, 80, 161]	0.51	*0.61*	*0.52*	*0.46*		0.44	*0.54*	*0.49*	*0.36*	
	Silent	[53, 57, 108, 114]	**0.53**	*0.51*	*0.52*	*0.50*	*0.57*	**0.53**	*0.40*	*0.60*	*0.56*	*0.53*
Children	Overall	[71, 72, 145, 60]	**0.55**	*0.45*	*0.59*	*0.51*	*0.74*	0.50	*0.38*	*0.54*	*0.48*	*0.61*
	Oral	[43, 42, 87]	0.51	*0.55*	*0.54*	*0.48*		**0.52**	*0.43*	*0.53*	*0.55*	
	Silent	[28, 30, 58, 60]	0.52	*0.52*	*0.65*	*0.36*	*0.60*	**0.57**	*0.28*	*0.70*	*0.53*	*0.64*
Adults	Overall	[63, 65, 124, 54]	**0.54**	*0.70*	*0.53*	*0.46*	*0.57*	0.49	*0.59*	*0.50*	*0.47*	*0.41*
	Oral	[38, 38, 74]	0.51	*0.69*	*0.50*	*0.42*		0.49	*0.59*	*0.56*	*0.41*	
	Silent	[25, 27, 50, 54]	**0.52**	*0.52*	*0.45*	*0.57*	*0.53*	0.50	*0.47*	*0.47*	*0.58*	*0.47*

a pseudo-word, or (in the silent condition) illegal. We expected it to be harder to distinguish among 3 or 4 kinds of isolated words and non-words than to tell easy from difficult sentences, both because n-way distinctions are harder than binary distinctions, and because reading an isolated word is so brief compared to reading a sentence. Nevertheless, as Table 4 shows, overall accuracy was reliably (albeit barely) better than random for 5 of 9 reader-specific classifiers and 3 of 9 reader-independent classifiers, showing that even with a single noisy channel, few participants, experimental setup less precise than lab studies, and simple analysis methods, there was above-chance performance on single-word stimuli. Moreover, within-class rank accuracy averaged from about 36% to 74% for reader-specific classifiers, and about 28% to 70% for reader-independent classifiers. Often above-chance accuracy on some classes comes at the cost of below-chance accuracy on one or more others. Only one reader-specific classifier was at or above 50% rank accuracy within every class. However, every reader-specific classifier has rank accuracy below 50% within at most one class, which means that it distinguishes among the other three classes with above-chance accuracy within all of them.

WHAT EEG COMPONENTS ARE SENSITIVE TO WHAT LEXICAL FEATURES?

To probe whether the features of the individual word being read affect activity in particular EEG frequency bands, we fit 15 separate linear mixed effects models, one model to predict activity in each of the 5 frequency bands: delta (1–3 Hz), theta (4–7 Hz), alpha (8–11 Hz), beta (12–29 Hz), and gamma (30–100 Hz), for all participants, just for children, and just for adults. Since lexical properties are notoriously highly correlated with each other, we limited the predictors used as fixed factors in the model to four different but relevant lexical properties: age-of-acquisition (AOA; a measure of how long a word has been known and how early it was learned), frequency (a measure of how often a word has been encountered), naming latency (a measure of how difficult the word is to read aloud), and letter bigram frequency (a measure of how visually familiar a word appears). Age-of-acquisition norms (AOA) and word frequencies from the SUBTLEX-US database of subtitles came from Kuperman, Stadthage-Gonzalez, & Brysbaert (2012); naming latencies and the mean

Table 5
Correlations of EEG power spectra to lexical features of words: Age-of-acquisition residuals, SUBTLEX-US word frequency, standardized naming latency, and mean bigram frequency. A single + or – indicates $p < 0.05$; two indicates $p < 0.01$; three indicates $p < 0.001$

		Delta (1–3 Hz)	Theta (4–7 Hz)	Alpha (8–11 Hz)	Beta (12–29 Hz)	Gamma (30–100 Hz)
Overall	AOA					- - -
	Frequency	- -	- -	- -		
	Naming Latency					
	Bigram					
Children	AOA					- - -
	Frequency					
	Naming Latency					
	Bigram					
Adults	AOA	+				
	Frequency	- -	- -	- -	- -	
	Naming Latency				+	-
	Bigram	-				

bigram frequency measure came from the English Lexicon Project (Balota et al., 2007). Even among these four factors, high correlations between age-of-acquisition and word frequency proved problematic for the models, so we used the age-of-acquisition residuals remaining after accounting for the shared variance with frequency (via a linear regression). Since we had no eye tracking, we relied on speech recognition to estimate when a reader read each word, so we excluded silent reading data from this analysis. We computed the start and end times of each read word by manually transcribing each oral reading utterance and using the Sphinx3 speech recognizer (CMU, 2008) in forced alignment mode to time-align the utterance to its transcript. In this analysis, there were variations in the EEG segment durations.

Models predicted the log power in each frequency band from the residual log age-of-acquisition, log frequency, mean letter bigram frequency, and standardized naming latencies (z-scores). All variables were centered by subtracting the mean for the overall sample. We included individual reader identity and word identity as random factors to model the distribution of readers and texts by allowing the intercept to vary by reader and word. Table 4 shows the direction of the correlation between each lexical variable and each EEG frequency band (positive or negative), both overall and separated by age group. The number of pluses or minuses indicates the p-value of the coefficient. Raw coefficient values are difficult to interpret due to the varying scales and log transformations.

Pooled data showed that overall, the lower frequency, more difficult words were associated with an increase in low frequency power across the delta, theta, and alpha bands. The gamma band, on the other hand, decreased for words learned later than would be expected based on frequency alone. To better interpret this finding in relation to AOA itself (rather than AOA residuals), we ran a model containing just log age-of-acquisition, without frequency. This model showed a positive relationship between AOA and gamma band activity ($p < 0.05$); later-learned words are associated with increased gamma activity. A model containing word frequency only (without AOA) shows no relationship in children between word frequency and gamma band activity.

Separate models for children and adults revealed several observations: (1) Word frequency effects on the EEG signal are driven by the adult reader data. (2) Many effects occurred only in adults and did not

reach significance in the pooled data, including delta band sensitivity to age-of-acquisition and bigram frequency, and increased beta power with decreased gamma power for words that take longer to name. (3) The increase in gamma for late AOA words was non-existent in adults, but strong in children.

The adult-only sensitivity to word frequency, naming latency, and bigram frequency can perhaps be accounted for by the fact that word frequencies and bigram frequencies are computed from adult corpora. Children may not have the same relationship with these measures of word frequency or bigram frequency because those frequencies may not reflect the actual frequency of encountered words for this age group. Moreover, naming latencies are based on adult readers, not children.

It stands to reason that age-of-acquisition is a more relevant lexical feature for children, because it is more likely to reflect which words they have learned or not learned, and how robust their word representations are likely to be. The difference in the quality of lexical representations between a word learned at age 2 (e.g., "daddy," "birthday") and age 6 (e.g., "dancer", "camera") might be small for an adult reader, but might be very large for a 7 or 8-year-old reader. Gamma activity has been associated with linking visual stimuli with other knowledge and memory. It is associated generally with binding. The late AOA words are likely to be words that the children are just learning or seeing for the first time, whereas the adult readers should know all of the words in our passages. The gamma band appears to be sensitive to the *learning* process occurring when children see new words. Gamma is not simply sensitive to visual novelty, as it has no relationship to bigram frequency, even in adults. Instead, based on prior studies, we hypothesize that the gamma band is sensitive to when a child is linking the visual form of an unfamiliar word to memory.

CHALLENGES, LIMITATIONS, AND FUTURE WORK

Unlike traditional ERP research conducted under carefully controlled laboratory conditions, our study involved several challenges due to its design and setting.

We faced a class size imbalance issue because we had more easy sentences than difficult ones and more non-words than real words. We addressed this issue by 1) undersampling to balance the training data, and 2) complementing overall accuracy by reporting accuracy within each class as a class-size-independent measure, unlike precision, which class size imbalance can inflate.

We faced variation in stimulus duration caused by student self-pacing. The Reading Tutor waited to display the next sentence until the student clicked to see it or finished reading the current sentence. Difficult sentences usually have more words, are more complex in syntactic structure, and thus took more time to read than easy sentences. We originally (Mostow et al., 2011) addressed this issue by averaging feature values over the duration of each stimulus. However, average feature values may vary as a result of duration. For example, suppose there is a cyclical peak in raw EEG signal every 10 seconds. Then the maximum raw EEG power will be higher for 10 second long sentences, which will each contain this peak, than for 5 second sentences, only half of which will contain the peak. This phenomenon is due to the method of segmentation, but not to text difficulty. In this paper, we addressed the issue by truncating sentences to equal durations, reframing the question as "can we classify difficulty based on the first 5 seconds of reading a sentence?"

However, this method also poses a potential problem. If there is a systematic signal related to, for example, end-of-sentence processing, it is more likely that a short sentence would include that signal when truncated than a longer sentence (because the shorter sentence will be truncated near the end of the sentence, whereas the longer sentence will be truncated closer to the middle). Taking advantage of this phenomenon is not cheating in the sense of exploiting duration information provided as classifier input, but it ignores the remainder of the sentence. We are currently exploring ways of segmenting the data so as not to take advantage of signals that may be related to the duration and location of the segments, rather than the

cognitive state we intended to induce by varying text difficulty. A general lesson in evaluating classifiers is to beware of clues to class identity based on how the data is segmented into instances.

Another cautionary note when training classifiers on continuous temporal data is to avoid inappropriately exploiting temporal continuity. If a mental state persists or varies slowly over a period of time, lumping all data points together and cross validating on the aggregated data set may overlook temporal dependencies among data points. For instance, when reading sentences in a story, the "reading something difficult" state may persist for several sentences. A naïve 10-fold cross validation that overlooks temporal dependencies and simply splits the data randomly may put sentences from the same story into both the training and test sets, thereby inflating classifier performance. We addressed this issue by cross-validating between stories instead of sentences.

Much work remains. In particular, muscle movements might cause EMG artifacts in the recorded EEG signals. If so, we could be detecting muscle movements as indicators of mental states. For example, perhaps when users realize that a dialogue system has misrecognized what they said, they consistently raise an eyebrow, thereby causing a peak in the EEG signal. We have no evidence of such an effect, and to be safe we mitigated the risk of EMG artifacts by using wavelets to denoise data, and considering only utterances with high reported signal quality. But even just detecting motor proxies for mental states could be useful for a tutor.

Future work includes detecting additional mental states, and improving detection accuracy. We are pursuing multiple complementary approaches. First, we want to add another EEG channel, if necessary by wearing two single-channel devices. Second, we need to collect more data. Besides manipulating stimuli experimentally, we can label training data based on observable events in longitudinal data. For instance, we can label an event as learning a skill if the student performs the skill better at the next opportunity. Third, we are trying more sophisticated training methods.

Finally, we need to figure out how to exploit EEG-based detection of mental states despite imperfect accuracy. Initially it may be useful only by aggregating over many observations, for instance to identify words unfamiliar to many students, or text "hot spots" that cause comprehension difficulty for many students. As the accuracy of mental state detection improves, EEG may contribute additional evidence to aggregate within an individual student model so as to enable more intelligent tutoring. If accuracy improves enough to rely on individual observations, an EEG-enabled tutor may respond in real-time, e.g., by explaining an unfamiliar word or paraphrasing a confusing sentence.

CONCLUSION

This study showed that the EEG data from a single electrode portable recording device can discriminate with significantly above-chance accuracy between reading easy and difficult sentences, on data pooled across populations (children and adults) and modalities (oral and silent reading). We achieved weak but still above-chance accuracy in detecting comprehension failure and lexical status of isolated words. Most interesting, we identified frequency bands sensitive to difficulty and to various lexical properties, suggesting the feasibility of using EEG to detect transient changes in cognitive task demands or specific attributes of lexical access.

Although weak, the statistically reliable relationship between reading difficulty and relatively impoverished EEG data illustrates its potential to detect mental states relevant to tutoring, such as comprehension, engagement, and learning. At the level of longitudinal data aggregated across students, such information could help generate and test hypotheses about learning, elucidate the interplay among emotion, cognition, and learning, and inform specific tutor responses to students. At the level of dynamic data about an individual student, the tutor could adapt to the student, either by responding immediately to a detected mental state, or by adapting more slowly to a cumulative student model updated over time. In summary, this pilot

study gives hope that a school-deployable EEG device may someday capture information that intelligent tutors can use to teach better.

ACKNOWLEDGMENTS

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A080157 and R305A080628 to Carnegie Mellon University, and by the National Science Foundation under Cyberlearning Grant IIS1124240. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute, the U.S. Department of Education, or the National Science Foundation. We thank Sarah Laszlo for stimuli, Tian Xia for wavelet decomposition, Vivian Yun-Nung Chen for forced alignment, Lucas Tan for his help with class size imbalance issues, the AIED2011 and journal reviewers for helpful comments, and the students, educators, and LISTENers who helped generate, collect, and analyze our data.

REFERENCES

Ackerman, P., Dykman, R., Oglesby, D., & Newton, J. (1994). EEG power spectra of children with dyslexia, slow learners, and normally reading children with ADD during verbal processing. *Journal of Learning Disabilities, 27*(10), 619-630.

Ackerman, P., McPherson, W., Oglesby, D., & Dykman, R. (1998). EEG power spectra of adolescent poor readers. *Journal of Learning Disabilities, 31*(1), 83-90.

Baker, R., D'Mello, S., Rodrigo, M., & Graesser, A. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223-241. doi: 10.1016/j.ijhcs.2009.12.003

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445-459.

Bartnik, E. A., Blinowska, K. J., & Durka, P. J. (1992). Single evoked potential reconstruction by means of wavelet transform. *Biological Cybernetics, 67*(2), 175-181.

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

Bertrand, O., Bohorquez, J., & Pernier, J. (1994). Time-frequency digital filtering based on an invertible wavelet transform: An application to evoked potentials. *IEEE Transactions on Biomedical Engineering, 41*(1), 77-88.

Blanchard, E., Chalfoun, P., & Frasson, C. (2007). Towards advanced learner modeling: Discussions on quasi real-time adaptation with physiological data. In *7th IEEE International Conference on Advanced Learning Technologies* (pp. 809-813). Niigata, Japan.

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. B. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology, 49*(2), 220-238.

Chalfoun, P. & Frasson, C. (2010). Showing the Positive Influence of Subliminal Cues on Learner's Performance and Intuition: An ERP Study. In V. Aleven, J. Kay & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (Vol. 6095, pp. 288-290). Pittsburgh, PA: Springer.

Chalfoun, P. & Frasson, C. (2011). Subliminal cues while teaching: HCI technique for enhanced learning. *Advances in Human-Computer Interaction - Special issue on subliminal communication in human-computer interaction*. doi: 10.1155/2011/968753

Chalfoun, P., & Frasson, C. (2012). Cognitive Priming: Assessing the Use of Non-conscious Perception to Enhance Learner's Reasoning Ability. In S. Cerri, W. Clancey, G. Papadourakis & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (Vol. 7315, pp. 84-89). Chania, Crete, Greece: Springer Berlin/Heidelberg.

Chaouachi, M., Chalfoun, P., Jraidi, I., & Frasson, C. (2010). Affect and mental engagement: towards adaptability for intelligent systems. In *Proceedings of the 23rd International FLAIRS Conference*. Daytona Beach, FL.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter - Special Issue on Learning From Imbalanced Datasets, 6*(1), 1-6. doi: 10.1145/1007730.1007733

Clarke, A., Barry, R., McCarthy, R., & Selikowitz, M. (2002). EEG Analysis of Children with Attention-Deficit/Hyperactivity Disorder and Comorbid Reading Disabilities. *Journal of Learning Disabilities, 35*(3), 276-285.

CMU. (2008). The CMU Sphinx Group open source speech recognition engines [software at http://cmusphinx.sourceforge.net].

Colon, E., Notermans, S., Weerd, J., & Kap, J. (1979). The discriminating role of EEG power spectra in dyslexic children. *Journal of Neurology, 221*(4), 257-262.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 33*(4), 497-505.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes, 13*(1), 21-58.

Crowley, K., Sliney, A., Pitt, I., & Murphy, D. (2010). Evaluating a brain-computer interface to categorise human emotional response, In *10th IEEE International Conference on Advanced Learning Technologies* (pp. 276-278). Sousse, Tunisia.

D'Mello, S., Craig, S., Fike, K., & Graesser, A. (2009). Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction (pp. 595-604*. San Diego, CA.

Dalal, S., Baillet, S., Adam, C., Ducorps, A., Schwartz, D., Jerbi, K., Bertrand, O., Garnero, L., Martinerie, J., & Lachaux, J. P. (2009). Simultaneous MEG and intracranial EEG recordings during attentive reading. *NeuroImage, 45*(4), 1289-1304.

Demiralp, T., Ademoglu, A., Schürmann, M., Basar-Eroglu, C., & Basar, E. (1999). Detection of P300 waves in single trials by the wavelet transform (WT). *Brain and Language, 66*(1), 108-128.

Derbali, L., Chalfoun, P., & Frasson, C. (2011). Assessment of learners' attention while overcoming errors and obstacles: An empirical study. In *15th International Conference on Artificial Intelligence in Education* (pp. 39-46). Auckland, New Zealand.

Derbali, L. & Frasson, C. (2011). Physiological evaluation of attention getting strategies during serious game play. In *15th International Conference on Artificial Intelligence in Education* (pp. 447-449). Auckland, New Zealand.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory, 41*(3), 613-627.

Fein, G., Galin, D., Johnstone, J., Yingling, C., Marcus, M., & Kiersch, M. (1983). EEG power spectra in normal and dyslexic children. I. Reliability during passive conditions. *Electroencephalography and Clinical Neurophysiology, 55*(4), 399-405. doi: 10.1016/0013-4694(83)90127-X

Fell, J., Klaver, P., Lehnertz, K., Grunwald, T., Schaller, C., Elger, C., & Fernández, G. (2001). Human memory formation is accompanied by rhinal-hippocampal coupling and decoupling. *Neuroscience, 4*(12), 1259-1264.

Fink, A., Grabner, R. H., Neuper, C., & Neubauer, A. C. (2005). EEG alpha band dissociation with increasing task demands. *Cognitive Brain Research, 24*(2), 252-259.

Gluck, K. A., Anderson, J. R., & Douglass, S. (2000). Broader bandwidth in student modeling: What if ITS were "Eye"TS? In *5th International Conference on Intelligent Tutoring Systems* (pp. 504-513). Montreal, Canada.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes, 25*(2), 149-188.

Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. In *28th Annual Meetings of the Cognitive Science Society* (pp. 285-290). Vancouver, Canada.

Gruber, T., Müller, M. M., & Keil, K. (2002). Modulation of induced gamma band responses in a perceptual learning task in the human EEG. *Cognitive Neuroscience, 14*(5), 732-744.

Harmony, T., Fernández, T., Silva, J., Bernal, J., Díaz-Comas, L., Reyes, A., Marosi, E., Rodríguez, M., & Rodríguez, M. (1996). EEG delta activity: An indicator of attention to internal processing during performance of mental tasks. *International Journal of Psychophysiology, 24*(1-2), 161-171.

Heraz, A. & Frasson, C. (2007). Predicting the three major dimensions of the learner's emotions from brainwaves. *World Academy of Science, Engineering and Technology, 31*, 323-329.

Heraz, A., Razaki, R., & Frasson, C. (2007). Using machine learning to predict learner emotional state from brainwaves. In *7th IEEE International Conference on Advanced Learning Technologies* (pp. 853-857.) Niigata, Japan.

Herrmann, C. S., Lenz, D., Junge, S., Busch, N. A., & Maess, B. (2004). Memory-matches evoke human gamma-responses. *BMC Neuroscience, 5*(13).

Howells, F. M., Stein, D. J., & Russell, V. A. (2010). Perceived mental effort correlates with changes in tonic arousal during attentional tasks. *Behavioral and Brain Functions, 6*(39).

Huang, R. S., Tsai, L. L., & Kuo, C. J. (2001). Selection of valid and reliable EEG features for predicting auditory and visual alertness levels. *Proceedings of the National Science Council, Republic of China. Part B, Life science, 25*(1), 17-25.

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375.

Jraidi, I., Chalfoun, P., & Frasson, C. (2012). Implicit Strategies for Intelligent Tutoring Systems. In S. Cerri, W. Clancey, G. Papadourakis & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (Vol. 7315, pp. 1-10). Chania, Crete, Greece: Springer Berlin/Heidelberg.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978-990.

Kutas, M. & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463-470.

Kutas, M. & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology, 11*(2), 99-116.

Lachaux, J. P., Jung, J., Mainy, N., Dreher, J. C., Bertrand, O., Baciu, M., Minotti, L., Hoffmann, D., & Kahane, P. (2008). Silence is golden: Transient neural deactivation in the prefrontal cortex during attentive reading. *Cerebral Cortex, 18*(2), 443-450.

Laszlo, S. & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology, 48*, 176-186.

Luo, A. & Sullivan, T. J. (2010). A user-friendly SSVEP-based brain-computer interface using a time-domain classifier. *Neural Engineering, 7*(2).

Lutsyuk, N. V., Éismont, E. V., & Pavlenko, V. B. (2006). Correlation of the characteristics of EEG potentials with the indices of attention in 12- to 13-year-old children. *Neurophysiology, 38*(3), 209-216.

Mainy, N., Kahane, P., Minotti, L., Hoffmann, D., Bertrand, O., & Lachaux, J. P. (2007). Neural correlates of consolidation in working memory. *Human Brain Mapping, 28*(3), 183-193.

Marosi, E., Bazán, O., Yañez, G., Bernal, J., Fernández, T., Rodríguez, M., Silva, J., & Reyes, A. (2002). Narrow-band spectral measurements of EEG during emotional tasks. *International Journal of Neuroscience, 112*(7), 871-891.

Mitchell, T., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. A., & Newman, S. D. (2004). Learning to decode cognitive states from brain images. *Machine Learning, 57*(1-2), 145-175.

Mostow, J. & Beck, J. E. (2007). When the rubber meets the road: Lessons from the in-school adventures of an automated Reading Tutor that listens. In B. Schneider & S. K. McDonald (Eds.), *Scale-Up in Education* (Vol. 2, pp. 183-200). Lanham, MD: Rowman & Littlefield Publishers.

Mostow, J. & Beck, J. E. (2009). Why, what, and how to log? Lessons from LISTEN. In *2nd International Conference on Educational Data Mining* (pp. 269-278). Córdoba, Spain.

Mostow, J., Chang, K. M., & Nelson, J. (2011). Toward exploiting EEG input in a Reading Tutor. In *15th International Conference on Artificial Intelligence in Education* (pp. 230-237). Auckland, NZ.

Mostow, J. & Jang, H. (2012). Generating diagnostic multiple choice comprehension cloze questions. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 7th Workshop on Innovative Use of NLP for Building Educational Applications*, Montréal, Canada.

Mota, S. & Picard, R. W. (2003). Automated Posture Analysis for Detecting Learner's Interest Level. In *Computer Vision and Pattern Recognition Workshop* (p. 49).

NeuroSky. (2009). NeuroSky's eSense™ meters and dtection of mntal sate: Neurosky, Inc.

Ng, A. Y. & Jordan, M. I. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*.

Picton, T. W. (1992). The P300 wave of the human event-related potential. *Clinical Neurophysiology, 9*(4), 456-479.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognising and responding to student affect. from Inderscience Enterprises Ltd.

Yasui, Y. (2009). A brainwave signal measurement and data processing technique for daily life applications. *Physiological Anthropology, 28*(3), 145-150.