Toward Unobtrusive Measurement of Reading Comprehension Using Low-Cost EEG

Yueran Yuan, Kai-min Chang, Jessica Nelson Taylor, and Jack Mostow

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
1 706 267 7370
yuerany@andrew.cmu.edu

ABSTRACT

Assessment of reading comprehension can be costly and obtrusive. In this paper, we use inexpensive EEG to detect reading comprehension of readers in a school environment. We use EEG signals to produce above-chance predictors of student performance on end-of-sentence cloze questions. We also attempt (unsuccessfully) to distinguish among student mental states evoked by distracters that violate either syntactic, semantic, or contextual constraints. In total, this work investigates the practicality of classroom use of inexpensive EEG devices as an unobtrusive measure of reading comprehension.

Categories and Subject Descriptors

D.3.3 [**Intelligent Tutoring Systems**]: EEG, Machine Learning

General Terms

Human Factors

Keywords

EEG, Intelligent Tutoring Systems, Reading Comprehension

1. INTRODUCTION

Assessments are necessary for tutoring systems that try to make informed and effective interventions. It follows that monitoring comprehension is vital for computerized tutors that teach reading skills. For the scope of this paper, we define comprehension as understanding the meaning of a sentence in the context of a story. Understanding could be hurt by failures in parsing the sentence (syntax), understanding the meaning of the words and phrases (semantics), and understanding how the sentence fits into the larger context of the story (inter-sentential understanding).

Many traditional systems assess students using comprehension questions. Though comprehension questions have strong face validity, they can take time to produce by hand. While computational methods exist to generate these questions, they are imperfect [1]. And though they have shown significant correlation with standardized comprehension tests [2], there remains room for improvement.

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA.

http://dx.doi.org/10.1145/2567574.2567624

Recent work has shown promising first results for the feasibility of using low-cost EEG to generate above-chance predictions of reading difficulty [3]. Ideally, replacing comprehension questions with passive assessment through EEG could (1) save material developers the trouble of creating questions, (2) save students the time to answer the questions, and (3) monitor the various components of understanding (e.g. semantic, syntactic) necessary to more fully diagnose lapses in comprehension.

We hypothesize that through the use of machine learning techniques, we can create an above-chance detector of lapses in reading comprehension. Additionally, we try to create a classifier for distinguishing student mental states as they evaluate 3 types of possible violations (syntactic, semantic, or contextual).

1.1 Related Work

Previous work on methods of unobtrusive reading assessment has shown that fluency measures (e.g. number of words read per second) can produce above-chance predictions of fluctuations in reading comprehension [4]. We hope that EEG can provide additional information about reading comprehension which may be used in conjunction with other unobtrusive measures to boost prediction accuracy.

Much of past work combining EEG and reading comprehension focuses on understanding mechanisms of comprehension [5][6] and not the practical use of EEG as a reading assessment tool. Notably, these experiments use costly EEG devices that are beyond what would be cost-effective in schools.

Studies using low-cost EEG have so far been unable to reliably predict reading comprehension [3]. In the present paper, we expand upon earlier work with low-cost EEG by taking a more principled approach to question design (Section 2.2) and using more sophisticated methods of classification (Section 2.4).

1.2 Overview

We report on two experiments. Our first experiment attempts to predict student performance on end-of-sentence cloze questions using EEG signals over the reading passage. Our second experiment attempts to detect features of an answer choice (e.g. ungrammaticality, semantic non-sensicality) using EEG signals over the short period of time that a student sees that choice. Section 2 describes our methodology. Sections 3 and 4 present respective results of the two experiments. Section 5 concludes.

2. PROCEDURE

We used a computerized reading tutor (Section 2.1) to present 76 multiple-choice cloze questions generated by 4 lab members (Section 2.2). We recorded EEG signals of students who read and completed those cloze questions. We use the correctness of the student responses as our dependent measure of comprehension.

We trained classifiers and evaluated them by testing their predictions against this measure (Section 2.4).

2.1 Environment

Our experiments are implemented in Project LISTEN's Reading Tutor. The Reading Tutor is a program that listens to children read and provides intervention when it catches mistakes. The program has demonstrated good performance in many aspects of reading instruction [2]. Because of its extensive logs and its support for embedding multiple choice questions within a story, it is a good platform for our experiment.

Table 1: Example Cloze Question

Item	Example	Explanation
Story Context	Then, Kimmie saw a big straw hat with a short red ribbon on it. The bow was not too long. The hat was not too fancy.	Sentences following the previous question and preceding the present question
Question	It would be easy to _____.	The sentence containing the cloze blank
Correct Answer	clean	Actual word in story that we removed to make the cloze blank
Ungrammatical Distracter	car	Makes sentence ungrammatical
Implausible Distracter	eat	Grammatically valid but semantically non-sensical.
Plausible Distracter	win	Grammatically valid and semantically sensible sentence. But the sentence must make no sense in context.

2.2 Task

To generate a question, we hand select an appropriate sentence and remove the last word. Each question has four choices – one correct choice and three distracters – as Table 1 illustrates below:

To probe the mental state of the student given their choices, the three distracters consist of one ungrammatical choice, one grammatically correct but implausible choice, and one plausible but incorrect choice. We hand-wrote these distracters as Table 1 explains. This design for the distracters is based on previous work with comprehension questions in the Reading Tutor [1].

After a student reads a truncated sentence, the tutor displays the 4 choices in random order. The tutor displays one choice at a time for 1 second each, so that we can tell when students are looking at each choice type. This way, we can separate the EEG data for each answer choice type. After displaying each choice individually, the tutor displays them again and the student clicks on one of the choices. Students received instructions on the cloze questions task at the start of each story.

We use multiple-choice cloze questions because they can be scored by a computer. In order to reliably select the one correct answer, the student must isolate the correct answer by recognizing each of the 3 distracters as incorrect. Because each distracter is designed to violate one aspect of understanding (syntax, semantics, inter-sentential context), ruling them out implies an understanding of those 3 aspects of meaning of the sentence.

Due to an implementation error, some questions we deployed had 2 implausible choices and no ungrammatical choice. This error introduces some additional noise for our answer type classifiers in section 4.

2.3 EEG Data

We deployed these cloze questions over the course of 7 weeks to 26 2nd-3rd grade students in classrooms at a local urban elementary school. In total, the students answered 906 questions. The percentages of choices that students selected are as follows: 78.7% correct, 13.0% plausible, 6.1% implausible, and 2.2% ungrammatical.

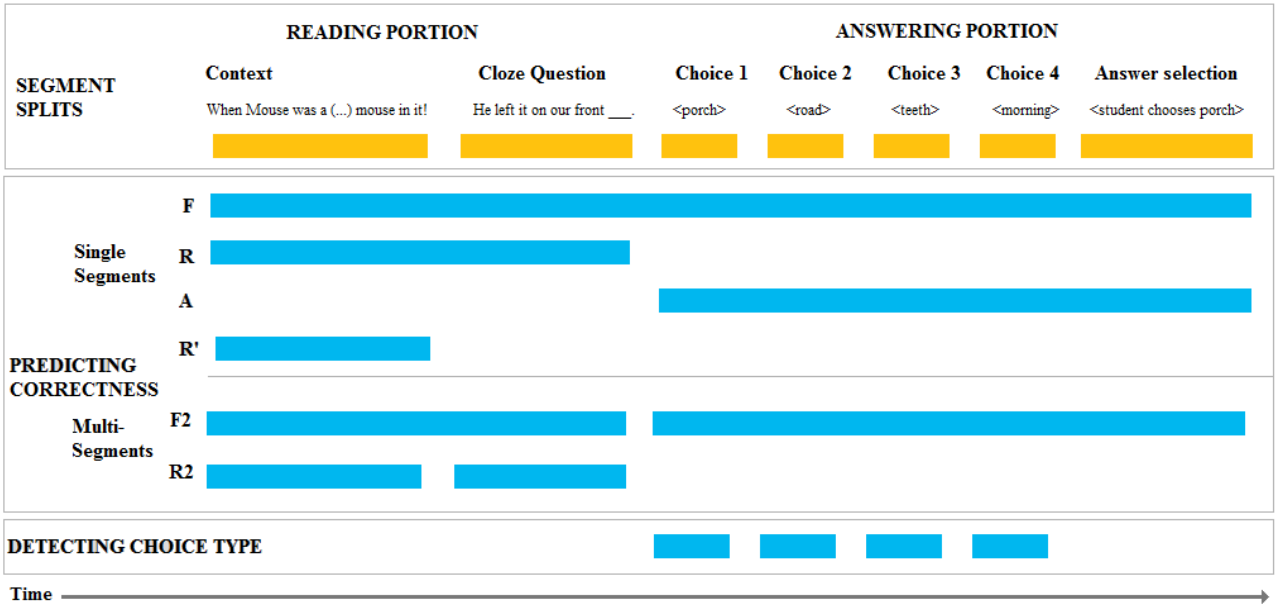

Figure 1. Segments Used in Experiments. F= full, R=reading, A=answering

We used NeuroSky BrainBands to record EEG raw waves at 512 Hz. We used NeuroSky's proprietary algorithms to generate Signal Quality, Attention, and Meditation scores at 1 Hz.

To denoise, NeuroSky's program removed frequencies below 3Hz and above 100Hz. Like Chang et al. [3], we did further denoising by applying a wavelet transform for soft thresholding [7].

2.4 Classification

For each experiment, we broke down EEG data into segments by time (2.4.1) and computed features from the segments (2.4.2). We trained Gaussian Naïve Bayes classifiers with these features and evaluated our classifiers with cross-validation (2.4.4). Figure 3 illustrates the flow of our experiment. This pipeline is based on Chang et al. [3].

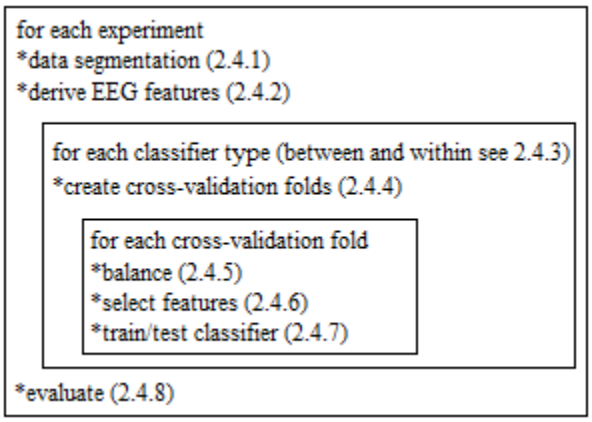

Figure 3: Flow of experiment in the form of pseudo-code

2.4.1 Data Segmentation

The student's task (reading and answering an end-of-sentence cloze question) consists of several segments (see Figure 1). Broadly, there are two overall segments for each question: (1) a reading portion. containing the sentences that form the context of the cloze question and the clozed sentence itself and (2) an answering portion, where the student is reading the 4 answer choices and selecting an answer. Figure 1 illustrates this breakdown.

2.4.2 Features

The NeuroSky program generated 12 data streams [1] – signal quality, attention, meditation, rawwave, delta, theta, alpha1, alpha2, beta1, beta2, gamma1, and gamma2. We removed any trials containing EEG signals with imperfect signal quality scores (for some experiments nearly 50% of data are removed for signal quality). We used the 11 channels other than signal quality to derive features for our classifier. For each channel, we derived 8 features: mean, variance, min, max, skew, kurtosis, first-order-polynomial-fit, and second-order-polynomial-fit. Thus each segment has 88 features in total.

2.4.3 Student-Specific vs. Student-Independent Classifier

For each experiment, we trained one within-subject classifier (trained and tested on data from the same subject) and one between-subject classifier (tested on one subject, trained on all other subjects). The within-subject classifier minimizes the effect of individual differences by custom training a classifier for each subject; the between-classifier simulates the performance of the classifier on unseen subjects. The between-subject classifier has the additional advantage of using a much larger training set (consisting of data from all other subjects) where the within-subject classifier only uses data from a single subject.

2.4.4 Cross-Validation

Our experiments used leave-one-out cross-validation to generate train/test splits. For every subject *s*, we trained the within-subject classifier on data from subject *s* and cross-validated on 1 left-out data point. We trained the between-subject classifier on data from all subjects except *s* and cross-validated on data from *s*.

2.4.5 Balancing

For most of our experiments, our data is imbalanced. Because we want to appraise the accuracy of our classifier without allowing it to exploit the class size imbalance, we prebalance the data prior to training. Like Chang et al [2], we chose to undersample from the majority class as a conservative way to prebalance. We lose an appreciable amount of data (~58%) by undersampling, but we are still able to get some significant results (Section 3.1.2)

2.4.6 Feature Selection

Due to the high number of features and our relatively small amount of data, overfitting is a definite concern. To limit our features, we used rank feature selection - ranking on class separability as computed by t-test. We selected the 3 highest ranked features as input for our classifier. We decided on 3 by trying a variety of cutoffs on a single test subject. This subject was not included in any experiments.

2.4.7 Classifier

We trained Gaussian Naïve Bayes classifiers using the selected features. We chose Naïve Bayes because it supports non-linear decision boundaries (by contrast, we tried linear SVM and it did not perform as well) and does not require a lot of data to train. To avoid building classifiers on too few data points, we only used subjects with at least 4 samples in each category.

2.4.8 Evaluation

Because we use leave-one-out classification, we produce exactly two predictions (one between subject and one within subject) for each data point. Thus, we have 2 accuracy measures per experiment – between subject accuracy and within subject accuracy.

We used two tests for significance. One is a Chi-Squared test, comparing the correct-versus-incorrect predictions against a baseline of 50:50. The Chi-Squared test is standard for evaluating the significance of categorical results but it assumes independence of samples. This assumption is a problem for us because we measure our predictor multiple times per subject (once per cross-validation fold) and these measurements of the same subject are not independent of each other.

To deal with these independence assumptions we used Fisher's Method as an additional test of significance. We treat each predictor's performance on each individual subject as a separate experiment and perform a Chi-Squared test on each subject. We then use Fisher's Method to aggregate the test results into a single significance value.

[1] Frequencies corresponding to various channels: delta: (1-3Hz); theta (4-7Hz), alpha1: (8-9Hz); alpha2: (10-12Hz); beta1: (13-17Hz); beta2: (18-30Hz); gamma1: (31-40Hz); gamma2: (41-50Hz) [8]

Fisher's Method is a meta-analysis technique for aggregating distinct experiments. We are aggregating individual subjects from the same experiment so we are uncertain about how well Fisher's Method applies. We noticed that Fisher's Method gave "significance" attributions to several near chance or below chance results. We suspect that we've made further violations of independence assumptions and will explore other significance tests in future work. We recommend caution when interpreting our significance values under Fisher's Method.

3. PREDICTING CORRECTNESS

First, we try to predict whether a student will answer a comprehension question correctly.

3.1 Setup

We want to detect comprehension without using EEG signals collected during the comprehension question. To fully factor out the comprehension question, we need to predict the outcome of our cloze question using exclusively EEG data over time periods before the student sees the question (see Experiment R' in Figure 1).

Each experiment began with 906 questions but many data segments were filtered out due to poor EEG signal quality (see N in Figure 2). The accuracy shown in Figure 2 is the percentage of correct predictions on left out data across all subjects.

3.2 Results and Discussion

Our classifier achieves *significantly above-chance accuracy* ($p < .05$) trained on only the reading portion (experiment *R'*). This result suggests that the reading portion is informative about comprehension. By contrast, the answering portion, without above-chance accuracy, may not be informative.

Significance remains when we limit the data to 4 seconds, the minimum length of the answering portion (see R4sec) indicating that our classifier is not taking advantage of the length of the reading portion. Further, our within-class R' classifier are has an above 50% accuracy for both the correct and incorrect labels, showing that we are not producing degenerate classifiers that only output the majority class.

The performance of our multi-segmental classifiers (F2, R2) is comparable to their single-segmental counterparts. It's possible that over-fitting due to the increased number of features cancelled out any benefit those features could have provided.

Notably, most of our significant results are from our between-subjects classifier. This is understandable since nearly all our subjects had fewer than 10 incorrect responses so our within-subject classifiers had very little training data.

We warn that our significance test makes more independence assumptions than warranted (see Section 2.4.8) so significance should be interpreted with caution. We also note that we did not correct for multiple comparisons, although our false discovery rate (expected number of false positives) is less than one, much lower than the number of significant results we found.

Though these results suggest that we can detect how well students understand a sentence, it's possible that we are merely recognizing student preparedness. Specifically, we would see similar results if our between-subject classifier was only picking up on how good of a reader the student is rather than any lapses of understanding in a specific sentence (good readers tend to answer correctly more often). We could check for this hypothesis by looking at whether our predictions for a given student differ across stories. Unfortunately, our current study does not have large enough N to make this analysis feasible.

4. RECOGNIZING ANSWER TYPES

In addition to correctness, we also tried to detect the student's mental state as he/she read each answer choice. Each of the 3 distracters in our task represents a different kind of error (see table 1) and should elicit different mental states in students as the students look at them. Further, these error-elicited mental states should be different from the mental state elicited by the correct choice. We test whether we could use EEG signals to detect these differences in mental state. An above-chance prediction would suggest that these inexpensive EEG devices could pick up some underlying mental state related to semantic, syntactic, or contextual violations.

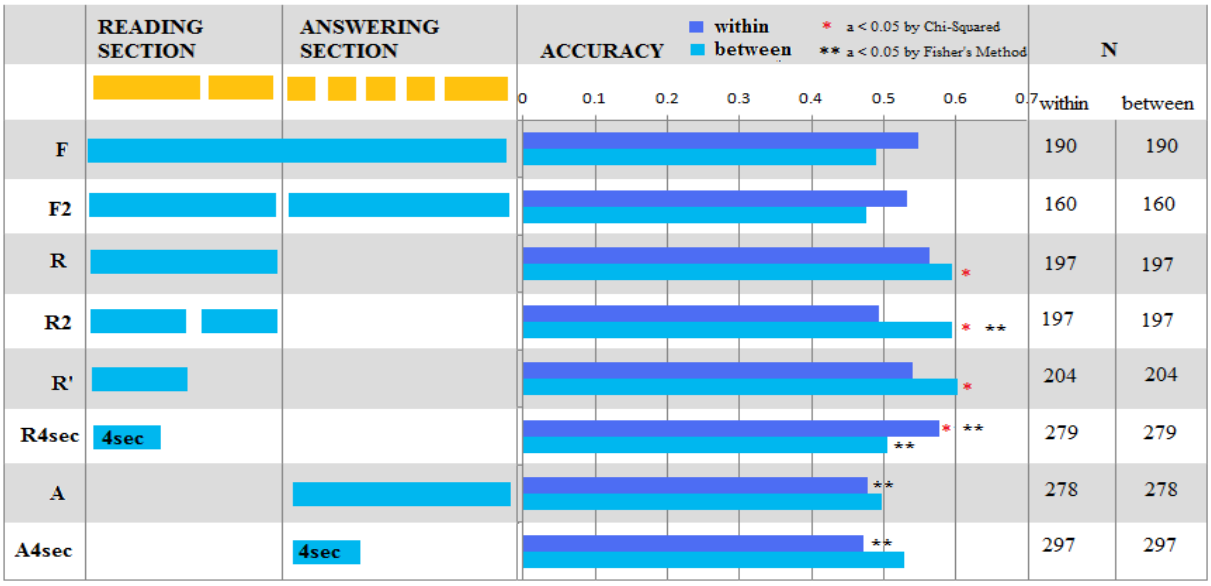

Figure 2: Accuracy of single segment experiments (note: using Fisher's Method produced significant below-chance results)

In these experiments, we use 1-second segments representing the time when a student was reading a particular type of answer choice (see choice 1 through choice 4 in Figure 1). We only used questions where the students gave correct responses, in the hope that they actually understood the answers. We produced binary classifiers for each pair of answer types (e.g. grammatical vs. correct; plausible vs. implausible). There were 6 distinctions in total.

4.1.1 Results and Discussion

The accuracy apparently varies across the different binary distinctions but because there are *no significant above-chance predictions*, we caution against drawing any conclusions from those apparent differences.

Note that our within-subject plausible vs. ungrammatical accuracy was significantly below chance. With $\alpha = 0.5$, we expect a false positive rate of 1/20. Given that there are 12 answer-type experiments, there is a reasonable (46%) chance of getting at least one false positive. This is in contrast to the correctness experiments where having 4 or more false positives in 16 experiments is highly unlikely (< 0.01% chance).

Table 5: Answer type recognition accuracy (Fischer's significance test not shown)

Experiment	Within		Between	
	N	Accuracy	N	Accuracy
Correct vs. Plausible	402	0.4776	402	0.4776
Correct vs. Implausible	401	0.4713	401	0.4863
Correct vs. Ungrammatical	393	0.4529	393	0.5115
Plausible vs. Implausible	418	0.4569	418	0.5072
Plausible vs. Ungrammatical	401	0.4090*	401	0.4539
Implausible vs. Ungrammatical	400	0.4525	400	0.475

The poor results could mean that these specific mental states are not detectable with EEG but there are a number of alternative explanations: (1) our specific lower-end EEG devices are not sensitive enough to detect these states (2) our classification pipeline does not make good enough use of the data or (3) we didn't collect enough data to overcome over-fitting/noise issues. Any combination of those causes could have led to these poor results. Future work could explore how to resolve these concerns.

5. CONCLUSION

The present work demonstrates that inexpensive EEG devices can generate above-chance predictions of comprehension. Critically, these predictions only use EEG data recorded before the students even saw the cloze question. However, certain independence-assumption violations put the strength of our significance tests in question. We also tried to detect student mental states as they saw various choice types. Thus far, we are unable to produce a significant-above-chance classifier of those mental states.

Our system is certainly not ready to replace comprehension questions. But these results, though modest, do suggest that *some* information about comprehension exists even in inexpensive EEG devices deployed in noisy classroom settings. We believe this work is a necessary first step in evaluating the practical feasibility of EEG as an unobtrusive measure of reading comprehension. Future work will explore ways of increasing prediction accuracy, assessing other dimensions of student knowledge, and applying those assessments to improve learning outcomes.

6. ACKNOWLEDGE

This work was supported 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 National Science Foundation. We thank the reviewers for helpful comments, and the students, educators, and LISTENers who helped generate, collect, and analyze our data.

7. REFERENCES

[1] Mostow, J. and Jang, H. 2012. Generating Diagnostic Multiple Choice Comprehension Cloze Questions. *NAACL-HLT 7th Workshop on Innovative Use of NLP for Building Educational Applications*

[2] Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B. and Tobin, B. 2003. Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29, 1 (December 2003), 61-117.

[3] Chang, K.-m., Nelson, J., Pant, U. and Mostow, J. 2013. Toward Exploiting EEG Input in a Reading Tutor. *International Journal of Artificial Intelligence in Education*, 22, (2013) 19-38

[4] Zhang, X., Mostow, J. and Beck, J. E. Can a computer listen for fluctuations in reading comprehension? *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, Los Angeles, CA, 495-502

[5] Baretta, L., Tomitch, L.M.B., Lim, V.K. & Waldie, K.E. 2012. Investigating reading comprehension through EEG. *Journal Ilha do Desterro*, 63, (2012) 69-100. DOI=http://dx.doi.org/10.5007/2175-8026.2012n63p69

[6] Coulson, S. and C. Petten. 2002. Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition* 30, 6, 958-968: DOI=http://dx.doi.org/10.3758/bf03195780

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

[8] NeuroSky. 2009. What are the different EEG Band Frequencies? (August 2009). Retrieved October 21 2013 from http://support.neurosky.com/kb/science/eeg-band-frequencies