RUNNING HEAD: NEURAL CORRELATES OF ENGAGEMENT

EEG Correlates of Engagement During Assessment

Laura K. Halderman[1], Bridgid Finn[1], Nicole M. Long[2], J.R. Lockwood[1] & Michael J. Kahana[2]

[1] Educational Testing Service

[2] University of Pennsylvania

Abstract

In educational assessment, low engagement is problematic when tests are low-stakes for students but have significant consequences for teachers or schools. The current study sought to establish the electroencephalographic (EEG) correlates of engagement. Forty university students completed a simulated Graduate Record Exam (GRE) session while EEG was recorded from 128 channels. Participants completed two verbal and two quantitative GRE test blocks for a total of 80 items and rated their engagement or mental effort after each item. High frequency spectral activity (90-150 Hz) over left temporal cortex during item completion predicted subsequent engagement ratings while controlling for reaction time and accuracy. However, reaction time was the sole significant predictor for effort ratings. These results suggest high gamma may be a correlate of engagement during complex cognitive tasks, but not effort. These findings could be used in future studies to measure levels of engagement across different assessment designs.

Introduction

Student Engagement During Assessment

The impact of student motivation and engagement on academic outcomes has been clear to educational stakeholders for decades (e.g. Meece, Blumenfeld, Hoyle, 1988). Engagement, a component of motivation (Skinner & Pitzer, 2012), is a multidimensional construct with various characterizations in the achievement motivation literature (Christenson, Reschly & Wylie, 2012). Engagement can reference a student's academic engagement over an entire school year or on a particular academic task. The current study focused on the narrower construct of task engagement, a person's active involvement in a task or activity, or in this case, an academic assessment (Reeve, Jang, Carrell, Jeon, & Barch, 2004). Our primary goal was to establish preliminary evidence that real-time measures of neural activity measured through EEG can provide a valid indicator of engagement in an academic assessment context.

Assessment related activities are integral to instruction and learning, and there are critical reasons why a more extensive understanding of engagement during assessment is needed. Consider an assessment that has low stakes for students but significant consequences for teachers or schools. Low engagement leads to scores that underestimate the student's actual abilities, which jeopardizes test validity and can leave academic institutions drawing questionable conclusions about the efficacy of their programs and teaching staff (Brown & Finney, 2011).

To date, the two most commonly used indicators of test-taking motivation and engagement are self-report measures and test-timing data (Sundre & Moore, 2002; Wise & DeMars, 2005; Wise, Wise & Bhola, 2006). In general, self-report instruments ask students to indicate their self-perceived level of motivation and engagement for a specific test as a whole

and their academic motivation more generally. Test-timing data include the time to complete a question or to finish a test, with atypically low values potentially indicating lower student engagement, effort and motivation. These measures have been the focus of research efforts aimed at identifying methods of minimizing the error introduced by unengaged test takers and improving the validity of scores.

These indirect measures are low cost and easy to collect, but have a number of identified limitations. Self-report methods typically measure engagement retrospectively and at a global task level rather than measuring it throughout a task. Some researchers have argued that global, retrospective evaluations can be insensitive to the fluctuations in engagement that occur during a task (Christenson et al., 2012) and amplify particular phases of the experience, such as the beginning or end (Kahneman, Diener & Schwarz, 1999; Finn, 2010; Finn & Miele, 2015). Furthermore, research has found that global engagement ratings can be subject to response biases for instance, providing higher ratings to avoid an anticipated academic penalty (Wise & Kong, 2005). Timing information can provide a more fine-grained view of engagement throughout a test, however, timing data alone is not complete enough to draw firm conclusions about a student's engagement. For example, rapid responses which are sometimes used as a marker of low engagement, could indicate a range of states: the student truly was not engaged, they were hurrying to finish given time constraints or they knew they did not know the answer (Metcalfe & Finn, 2008; Schnipke, 1995; Wise & Kong, 2005). To address these limitations, the current study used an item-level self-report method to capture fluctuations in engagement, throughout a testing session. Test-timing data was also collected to provide additional information about participants' experiences with items.

Combining common engagement indicators with physiological data such as EEG could provide a more comprehensive view of test-taker experience. EEG is a useful tool in this context because it is measured passively during a task and provides extremely fine-grained measurement (Luck, 2014). An EEG measure of engagement during the test affords opportunities to explore how cognitive states change over time, at a fine grain size, and without drawing the participant's awareness away from the primary task. Moment-by-moment engagement measures might better capture the item and overall assessment design features that promote or discourage engagement, which may be increasingly important as assessments move toward using more complex tasks (e.g., simulation and scenario-based tasks) that may invoke a wider variety of cognitive states during administration. Academic assessments built upon such designs could provide results that are more valid for supporting inferences about both individual skills and institutional effectiveness. Toward this end, our primary goal was to establish a neural correlate of self-reported engagement at an item level under conditions relevant for academic testing situations.

Use of EEG to Measure Engagement

There is only a modest literature on the use of EEG to measure engagement (e.g. Berka et al., 2007; Pope, Bogart & Bartolome, 1995; Freeman, Mikulka, Prinzel & Screbo, 1999; Chaouachi & Frasson, 2010). The majority of the work can be found in the applied, Human-Computer Interaction and attention/vigilance literatures. Engagement measures in these domains typically involve online cognitive state classification and/or identifying a person's level of engagement during a task. Characteristic tasks include visual search paradigms in which participants search for a specific symbol in an array of visual stimuli or maintain vigilance on a central fixation point. Most of the engagement measures in these domains have typically focused

on lower frequencies (i.e. alpha, theta and beta) (Chaouachi & Frasson, 2010, Freeman et al., 1999; Pope, et al., 1995).

The nature of engagement measured in attention/vigilance paradigms may be too narrow to capture the kind of engagement that is required when reasoning through complex verbal and quantitative items like those used to assess academic achievement and thus, may not be useful to inform a priori hypotheses about which frequencies might best index engagement in an assessment context. Evidence from cognitive neuroscience has established gamma (> 30 Hz) as a critical frequency band for a variety of cognitive processes including attention (Jensen, Kaiser & Lachaux, 2007), word recognition (Jerbi et al., 2009), auditory processing (Crone, Boatman, Gordon & Hao, 2001), and motor movement (Miller, Shenoy, Miller, Rao & Ojemann, 2007). Most notably, gamma increases are indicative of learning mechanisms, as established through both intracranial and scalp EEG (Sederberg, Kahana, Howard, Donner & Madsen, 2003; Gruber, Tsivilis, Montaldi & Müller, 2004; Sederberg et al., 2006; Long, Burke & Kahana, 2014 and see also Fitzgibbon, Pope, Mackenzie, Clark & Willoughby, 2004). For example, Long et al. (2014) examined the neural correlates of successful memory encoding using intracranial EEG recordings in neurosurgical patients and scalp EEG recordings in healthy controls. In addition to significant theta (3-8 Hz) power modulations, an increase in high gamma power (44 - 100 Hz) was a predictor of successful learning in both participant samples. The researchers concluded that scalp EEG was capable of resolving high frequency gamma activity that was indicative of subsequent memory effects. Thus, gamma may be a useful index of engagement in a task context, which involves higher-level cognitive processes such as problem solving (Blumenfeld, Kempler & Kraicik, 2006).

Current Study

In the current study, EEG was measured as students completed a simulated session of the Graduate Record Exam (GRE). An engagement rating or mental effort rating was taken after each item. Item-level self-report ratings were used in this study as the initial correlate with EEG. A central question was whether changes in particular frequencies during item completion are indicative of whether a person subsequently reports being meaningfully engaged with the task. Mental effort ratings were solicited to evaluate whether the EEG patterns associated with high and low engagement could be distinguished from those associated with high and low effort, a state that has sometimes been conflated with engagement in the literature but many argue is a separable aspect of test-taking motivation (Chapman, 2003; Skinner & Belmont, 1993).

To establish a possible EEG correlate of engagement, we conducted a series of behavioral and EEG analyses. First, we tested the validity of the item level self-report scales. A critical first step was to demonstrate that the two scales were used distinctly by the participants, which we establish by verifying that the two scales were differentially related to important test-taking metrics such as accuracy and response time. This evidence supported the use of the item-level scales as the foundation for the EEG analyses. Second, we tested whether we could find evidence for an EEG correlate of engagement and effort. Separate correlational analyses were conducted for each rating type to evaluate the relationship between the ratings and EEG power. Third, we conducted a more stringent test of EEG as a measure of engagement, focusing on high gamma, which demonstrated the strongest correlation with engagement ratings. Mixed effects ordinal modeling was used to evaluate whether EEG provided a meaningful predictor of self-reported engagement while controlling for important behavioral metrics of test-taker engagement.

Method

Participants

Forty students (22 female; average age 22.1 years (*SD* = 3.5 years)) at the University of Pennsylvania participated in a 2.5-hour experimental session. Participants were recruited from a pool of students who had previously completed experimental sessions in the lab. Participants were included in the study if they had never practiced or taken the GRE, or took it before 2011 when the current, adaptive GRE was introduced. Participants were paid $75 for their participation.

Design and Materials

Participants completed a simulated GRE session while EEG was recorded. Participants completed two verbal and two quantitative GRE test blocks with 20 items in each block. The domain of the first test block was counterbalanced across participants. Materials were taken from a practice GRE test, which contained a routing block and three second-stage blocks (easy, medium and hard). Performance on the routing block determined a participant's second-stage block assignment. Table 1 depicts the number of items correct in the routing block required for each second-stage block assignment. Quantitative item types included constructed responses (numerical solution), standard items with only one correct answer and multiple response items, with multiple correct answers. Verbal item types included items with sub-questions where each sub-question required an answer, standard items and multiple response items.

Table 1

The Number of Items Correct in the Routing Block Required for Each Second-Stage Block Assignment.

Domain	Number of Correct Items	Second Block Assignment
Quantitative	6 or less	Easy
	7 – 11	Medium
	12 or more	Hard
Verbal	6 or less	Easy
	7 – 12	Medium
	13 or more	Hard

Participants rated either their mental effort or engagement after each item on a scale of 1 (Low) – 6 (High). Table 2 shows distributions of each rating type. For mental effort, participants were asked to consider, "How much mental activity was required (e.g. thinking, deciding, calculating, remembering, searching, etc.)? How hard did you have to work to accomplish your level of performance?" For engagement, participants were asked to consider, "How absorbed were you while you were answering the item? How attentive and involved were you?"

Table 2

Distribution of the number of participant responses for Effort (EFF) and Engagement (ENG) ratings.

Rating	EFF	ENG
1	99	88
2	318	260
3	346	313
4	272	268
5	164	221
6	96	135
Total	1295	1285

Procedure

The session began with a brief tutorial with six practice items representing the different item types. Each participant completed a total of four blocks of items consisting of a verbal routing block, a verbal second-stage block, a quantitative routing block, and a quantitative second-stage block. Block order was randomly assigned, however, the first verbal or quantitative block was always a routing block because the second-stage block assignment depended on the routing block performance. Within each block, items of the same type were presented together but the different item types appeared in a random order across participants. After each item, an effort or engagement prompt was randomly but equally presented so that across a testing session, a participant received an equal number of effort and engagement prompts. This procedure was adopted to reduce fatigue effects and predictability which might

have occurred if participants were asked to rate both effort and engagement after every item. Participants could return to previously answered items and change their answers, however, only the first encounter with an item was analyzed. A 10-minute break was given between the second and third blocks.

EEG Data Collection

EEG was recorded using a Geodesic Sensor Net (GSN; Netstation 4.3 acquisition environment, from Electrical Geodesics, Inc.). The GSN provided 129 standardized electrode placements across participants. All channels were digitized at a sampling rate of 500 Hz, and the signal from the caps was amplified via either the Net Amps 200 or 300 amplifier. Recordings were initially referenced to Cz and later converted to a bipolar reference.

EEG Data Processing

To control for the variability in response times across quantitative and verbal items and participants, all EEG measurements were response-locked to the participant's first response on an item and included the 20 seconds prior to the response. 19.4% of the total items were omitted from the analyses because the response time was less than 20 seconds. Though feelings of engagement and effort likely vary throughout the course of an item, this approach allowed us to keep all trials an equal length (20 s) across items and participants.

To minimize confounds resulting from volume conduction and saccades, we analyzed the scalp EEG with bipolar referencing (Nunez & Srinivasan, 2006; Kovach et al., 2011) to minimize contribution of eye blinks to the EEG signal, a method used previously to resolve high gamma signals in scalp EEG (Long et al., 2014). We defined the bipolar montage in our dataset based on the geometry of the scalp EEG electrode arrangements. For each participant and electrode, we isolated pairs of immediately adjacent electrodes and found the difference in

voltage between them (Long et al., 2014). The resulting bipolar signals were treated as new virtual electrodes and are referred to as such in the remainder of the text.

We applied the Morlet wavelet transform (wave number 6) to all bipolar electrode EEG signals during 20 seconds response-locked epochs, across 52 logarithmically spaced frequencies (2-165 Hz). We included a 1000 ms buffer on both sides of the data to minimize edge effects. After log transforming the power, we down-sampled the data by taking a moving average across 100 ms time windows from stimulus onset and sliding the window every 50 ms, resulting in 399 total time windows with 200 non-overlapping time windows. Log-transformed power values were then Z-transformed to normalize power within participants. Power values were Z-transformed by subtracting the mean and dividing by the standard deviation power which were calculated across all events and time points for each frequency. We split the Z-transformed power into seven distinct frequency bands (delta, 2-4 Hz; theta, 4-8 Hz; alpha, 8-12 Hz; beta, 12-28 Hz; low gamma, 28-44 Hz; medium gamma, 44-90 Hz; and high gamma, 90-150 Hz; Long et al., 2014), by taking the mean of the Z-transformed power in each frequency band.

Using regions defined in previous scalp EEG studies of human memory (Long et al., 2014; Long & Kahana, 2017; Weidemann, Mollison & Kahana, 2009), we focused on six a priori regions of interest (ROIs). These particular ROIs encompass most of the scalp while reducing the space to the essential "dimensions", namely left/right, anterior/posterior, superior/inferior. Moreover, by using a priori ROIs that were not defined specifically for this study, we can be confident that our ROI selection is unbiased. Z-power in each frequency band was averaged across our six ROIs: left hemisphere (LH) and right hemisphere (RH) frontal, temporal and parietal (Figure 1). This was done for each GRE item. Items with response times less than 20s on the first encounter were excluded from the analyses. All other items were included.

Results

An average of 14.7 (SD = 3.5) out of 20 questions were answered correctly on the quantitative routing block and 11.0 (SD = 3.2) out of 20 questions were answered correctly on the verbal routing block. Table 3 shows the number of participants routed to the easy, medium and hard second-stage blocks for each domain. Participants completed each quantitative item in an average of 73.2s (SD = 45.9) and each verbal item in 58.2s (SD = 40.9).

Table 3

Number of Participants in Each Second-stage Block Assignment

	Second-Stage Blocks		
Domain	Easy	Medium	Hard
Quantitative	1	5	34
Verbal	2	15	23

Distinguishing Engagement from Effort

While effort (as measured through response time; Wise & DeMars, 2005) is sometimes used as a proxy for engagement, some researchers have argued that engagement and effort are distinct constructs (e.g. Bloom & Turner, 2001). If they are distinct, and our prompts measure these constructs accurately, then we expect that participants would differentiate their responses to the two types of prompts. This should result in the Effort ratings having different properties (e.g., relationships with other variables) than the Engagement ratings. Several analyses were conducted to probe for such differences. Note that although each participant responded to only

one type of prompt for each item, the randomization of the prompt presesentation described previously ensures that neither item nor participant effects confound any observed relationships.

First, a linear model for the ratings was fit with three categorical predictors: the rating type (2 levels, Effort vs Engagement), participant (40 levels, one per participant) and item. The model was then augmented with the interaction of rating type and participant. The model including the interaction was preferred ($F = 3.48$, $p < .001$) suggesting participants were differentiating their responses to the two types of prompts.

The two types of ratings also related differently to performance. Figure 1 plots percent correct as a function of the rating value for each type of rating. Across the rating range, the relationship between Effort ratings and percent correct was effectively flat. Conversely, there was a strong positive relationship between Engagement ratings and percent correct. When participants reported having been highly engaged while responding to an item, they were much more likely to have answered that item correctly.

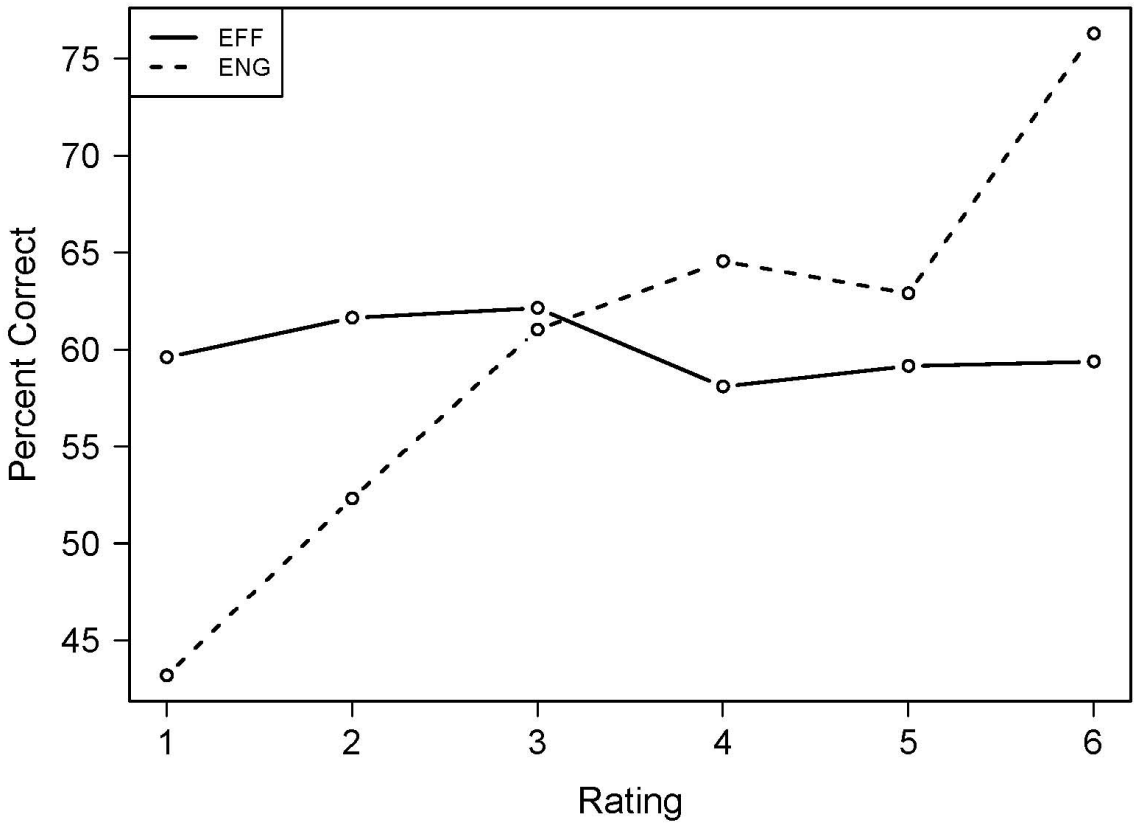

Fig. 1.

Percentage correct by rating level, for Effort and Engagement ratings.

Finally, the two types of ratings also related differently to response time. Figure 2 provides boxplots of the natural log of the response time by each rating and rating type. Both rating types were positively related to response time, but the relationship between Effort ratings and response time was stronger. One explanation is that response time was a more salient cue for evaluating effort than engagement.

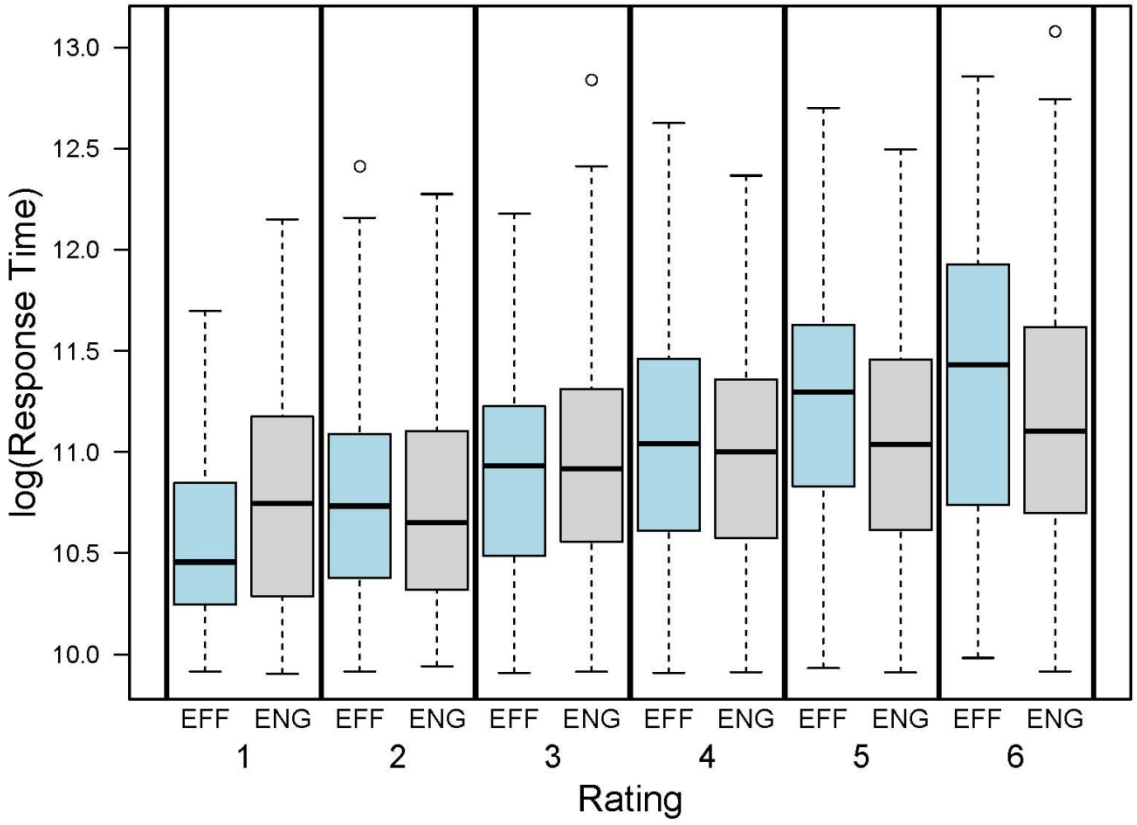

Fig. 2.

Boxplots of log response time by rating level and rating type.

The results collectively support the conclusion that when participants made Engagement and Effort ratings, they were drawing upon different sources of information from their experiences with items to make their ratings.

Relating Ratings to EEG Measurements

We next sought to determine the neural correlates of engagement and effort and to test whether the two cognitive states showed distinct patterns using neural data. To establish an EEG correlate of engagement, we ran correlations between engagement ratings and bipolar Z-

transformed log power in seven frequency bands in six scalp ROIs for each participant. We analyzed a wide range of frequencies so that the results were not pre-emptively constrained. Once these correlations were computed for each participant, one-tailed t-tests were conducted on the entire sample to determine whether the average correlations were significantly different from zero (see Table 4). Bonferroni corrections were used to correct for multiple comparisons. The correlational analyses showed that greater power for low, medium and high gamma frequencies in the left temporal (LT) and right parietal (RP) electrodes was associated with higher engagement ratings. LT and RP electrodes also had significant positive correlations between power in the beta frequency band and engagement ratings, however, these correlations were slightly weaker than those observed in the gamma frequency bands.

Table 4

Average Correlations Between Engagement Ratings and Bipolar Referenced Power in the Seven Frequency Bands in Each of the Six Scalp ROIs

	Left Hemisphere			Right Hemisphere		
Frequency Band	Frontal	Temporal	Parietal	Frontal	Temporal	Parietal
Delta	.054	.074	.089	.081	.051	.102
Theta	.047	.054	.014	.046	.040	.087
Alpha	.012	.071	-.034	-.014	.007	.062
Beta	-.029	.150*	.009	.009	-.009	.140*
Low Gamma	-.009	.191*	.090	.118	.017	.166*
Medium Gamma	-.003	.192*	.087	.123	.013	.169*
High Gamma	-.008	.194*	.080	.086	.019	.161*

Note. Significant correlations were significant after a Bonferroni correction for 42 comparisons

* p <.0012

A 2 (Engagement: High, Low) x 6 (ROI: LH frontal, temporal, parietal; RH frontal, temporal, parietal) x 7 (frequency: delta, theta, alpha, beta, low gamma, medium gamma, high gamma) repeated measures ANOVA was conducted to confirm the pattern established in the correlational analysis. Engagement ratings were combined into two categories, high (ratings of 4-6) and low (ratings of 1-3). The Engagement x ROI x frequency interaction was significant ($F(4.26, 166.11) = 5.05$, $MSE = .10$, $p(GG) = .001$, $\eta^2 = .12$ Greenhouse-Geisser correction applied) indicating that both ROI and frequency modulated the engagement effect. Figure 3 shows the power for each level of engagement at each ROI location and for each frequency. In addition, visual inspection of the confidence intervals for high and low engagement show a similar pattern to the correlational analysis: greater power differences for high compared to low engaging items in the LT and RP ROIs for the gamma frequency bands. Power in the beta frequency band was also greater for high compared to low engaging items in the LT and RP ROIs.

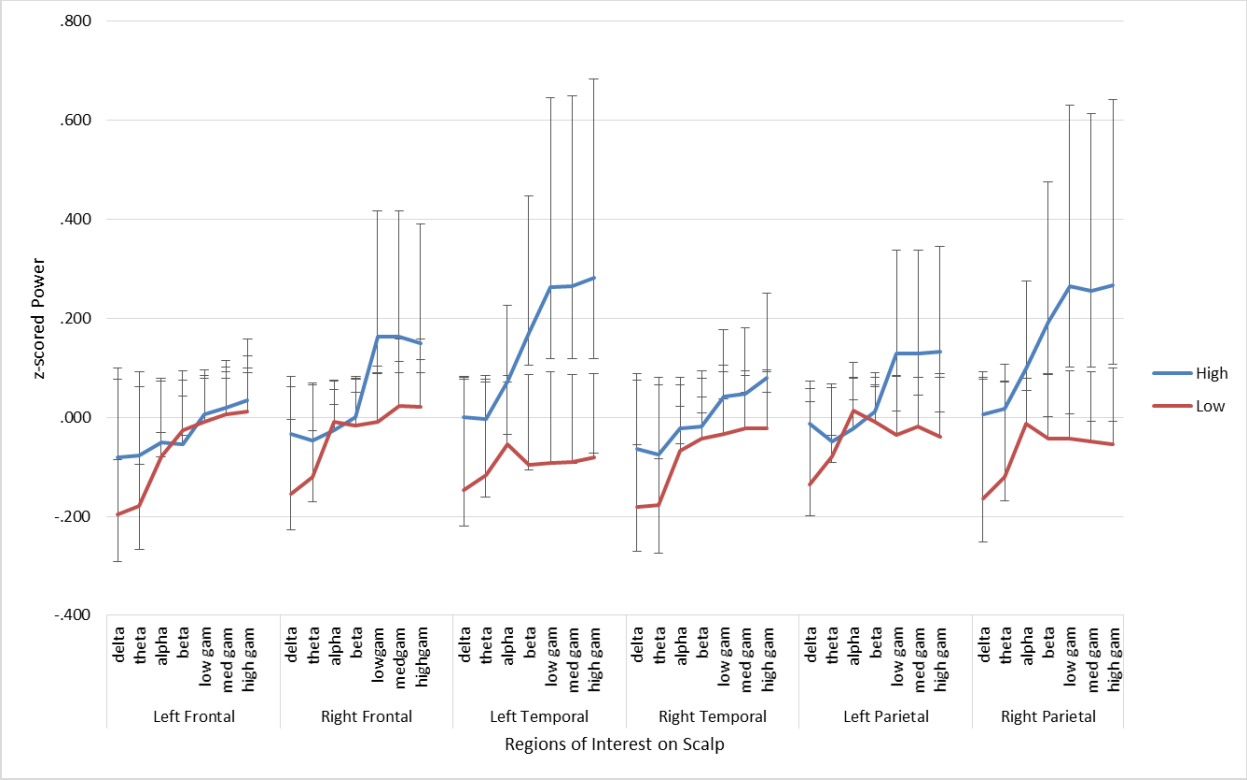

Fig. 3.

Means and 95% confidence intervals for high and low engaging items for each ROI and for each frequency band.

In contrast to engagement ratings, we found no significant correlations between effort ratings and power in any of the frequency bands in any of the specified ROIs (see Table 5). As no significant correlations were observed, the corresponding ANOVA was not computed. Results across analyses provide confirmation that engagement but not effort ratings were related to increased power in gamma frequency bands and provide strong support for further exploration of factors that modulate these differences.

Table 5

Average Correlations Between Effort Ratings and Bipolar Referenced Power in the Seven Frequency Bands in Each of the Six Scalp ROIs

	Left Hemisphere			Right Hemisphere		
Frequency Band	Frontal	Temporal	Parietal	Frontal	Temporal	Parietal
Delta	-.021	-.015	-.004	-.026	.021	-.005
Theta	.017	.012	-.002	.005	.045	-.001
Alpha	.036	.055	.015	.037	.025	.063
Beta	-.012	.069	.024	.026	.004	.085
Low Gamma	-.027	.072	.062	.054	.001	.091
Medium Gamma	.001	.085	.096	.080	.015	.098
High Gamma	.001	.073	.068	.056	.009	.091

Note. No significant correlations remained after a Bonferroni correction for 42 comparisons

Our next goal was to establish a link between EEG signals and self-report measures of engagement and effort when accounting for behavioral measures typically used to measure these cognitive states. We focused exclusively on high gamma (90-150 Hz) because it had the strongest correlation with engagement ratings and the most robust mean differences in the ANOVA analysis. We used a mixed effects ordinal model to test whether high gamma activity reliably predicted engagement or effort ratings after controlling for reaction time and accuracy.

Modeling engagement and effort ratings. Our model for the ordered categorical Engagement ratings was a function of: 1) response accuracy; 2) log response time (mean-centered across all observations with Engagement ratings); 3) high-gamma activity from the six

ROIs entered as six individual variables; 4) random effects for participants; and 5) random effects for items. Response accuracy and log response time were included in the model because these are traditional proxies for effort and engagement and we were interested in understanding what unique contributions the EEG measurements may have for explaining engagement ratings. The random effects were included to account for unobserved participant and item attributes that are related to the Engagement ratings. We conducted the modeling using the cumulative linked mixed models routine in the ordinal package (Christensen, 2015) for the R environment as this routine can fit cumulative logit models (Agresti, 1990) with cross-classified random effects (Goldstein, 1994; Raudenbush & Bryk, 2002).

The results are summarized in Table 6. The relationship for each predictor is presented in terms of an odds ratio, and the model was parameterized such that odds ratios greater than 1 indicate increased probability of the participant reporting a higher engagement rating on the 1-6 ordered rating scale. The odds ratio for item accuracy refers to a question being answered correctly vs incorrectly. The odds ratio for log response time refers to a one unit change in the log of the response time. The odds ratios for the EEG measurements refer to a one standard deviation unit change in the scale of the EEG measurement.

Table 6

Estimated Odds Ratio from Cumulative Logit Model of Engagement Ratings

Regressor	Odds Ratio	95% C.I.	p-value
Item Accuracy	1.71	(1.35, 2.16)	7.19e-06 ***
Log Response Time	2.32	(1.90, 2.83)	< 2e-16 ***
LF High Gamma	0.85	(0.69, 1.05)	0.121
RF High Gamma	0.92	(0.73, 1.18)	0.519
LT High Gamma	1.44	(1.17, 1.77)	0.00059 ***
RT High Gamma	1.01	(0.83, 1.24)	0.896
LP High Gamma	0.97	(0.77, 1.23)	0.800
RP High Gamma	1.02	(0.83, 1.26)	0.843

Note. *** indicates significant p-values.

The odds ratio for response accuracy and the log response time indicated that each was positively associated with higher engagement ratings, consistent with previous results. The results further demonstrated that high gamma in LT showed a significant positive relationship with the likelihood of reporting higher engagement. Figure 4 provides boxplots of LT high gamma separately for each level of the Engagement rating. The distributions shift upward as the rating increased, consistent with the model results. Finally, the finding of a statistically significant positive association between LT high gamma and the engagement rating was robust to several sensitivity analyses, itemized in the supplementary online materials. It is also worth noting that the model reported in Table 6 was preferred over a simpler alternative that excluded all six high gamma activity terms (LR statistic 28.4 on 6 df, p = 8e-5), providing further

evidence that high gamma activity provides information about engagement above and beyond the traditional proxies.

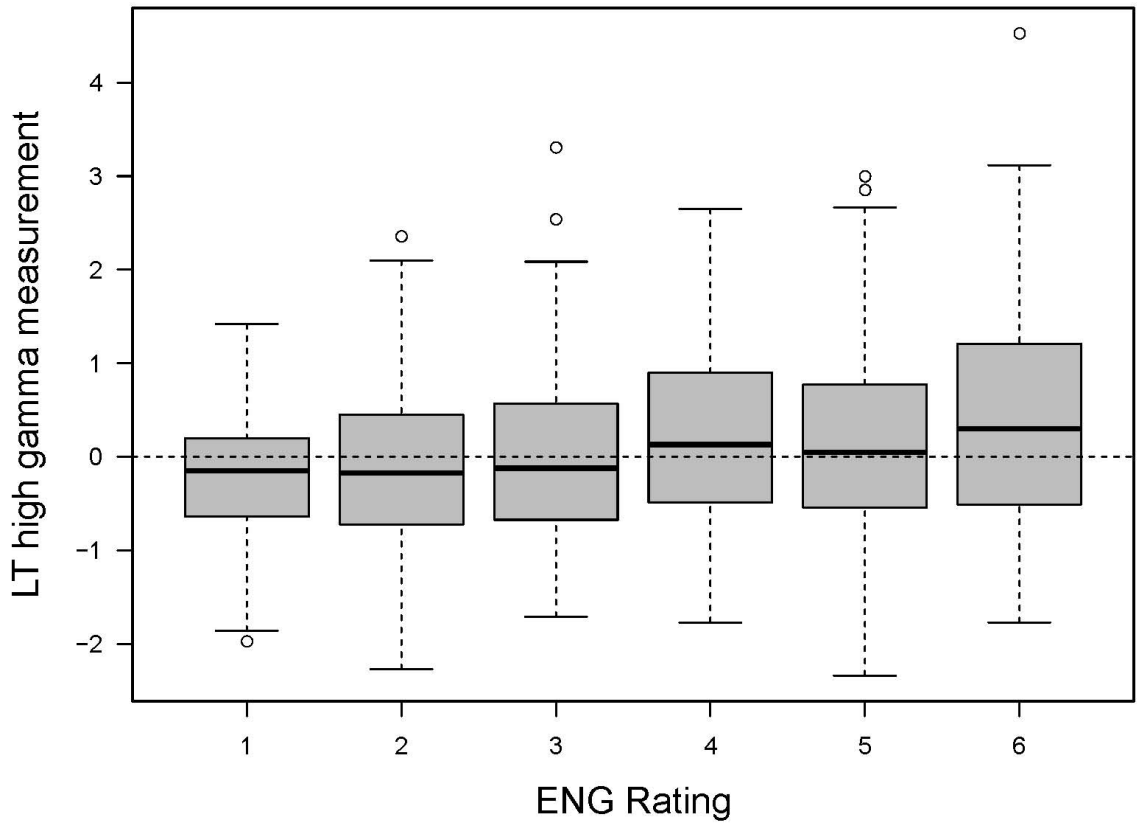

Fig. 4.

Boxplots of the LT high gamma measurements separately for each level of the Engagement (ENG) rating.

We followed a parallel process for the analysis of the Effort ratings to further confirm that these two states can be differentiated in this dataset. As can be seen in Table 7, response accuracy was not a significant predictor of the Effort ratings, whereas the association between the log response time and the Effort rating was extremely large and positive. This is consistent with the descriptive results presented previously. Also in contrast to the Engagement ratings,

high gamma power did not have a statistically significant association with the Effort ratings in any of the ROIs. A likelihood ratio test comparing the model to a simpler model that excludes gamma power in all six ROIs did not reject the simpler model (LR statistic 2.46 on 6 df, p = .87). Thus, there was no evidence that gamma power had any predictive value for the Effort ratings beyond behavioral proxies.

Table 7

Estimated Odds Ratio from Cumulative Logit Model of Effort Ratings

Regressor	Odds Ratio	95% C.I.	p-value
Item Accuracy	0.98	(0.79, 1.23)	0.861
Log Response Time	4.51	(3.61, 5.63)	2e-40***
LF High Gamma	0.99	(0.81, 1.22)	0.951
RF High Gamma	1.02	(0.81, 1.29)	0.851
LT High Gamma	0.98	(0.81, 1.19)	0.851
RT High Gamma	0.93	(0.76, 1.14)	0.483
LP High Gamma	1.05	(0.83, 1.33)	0.676
RP High Gamma	1.08	(0.89, 1.31)	0.431

Note. *** indicates significant p-values.

Discussion

The goal of the current study was to identify neural correlates associated with engagement during assessment, wherein engagement was exemplified as active involvement in a task or activity (Reeve et al., 2004). Behavioral evidence suggested participants used the

engagement and effort scales to index different aspects of their experiences with items. Specifically, response accuracy was more tightly linked with engagement, which corresponds with the educational measurement literature showing a tight link between engagement and performance (Wise et al., 2006), whereas reaction times were more closely associated with effort. Critically, neural signals during GRE questions correlated with post-question engagement ratings. Increased high gamma power (90-150 Hz) was positively correlated with higher engagement ratings, even when accounting for reaction times and response accuracy. Increases in gamma have been observed in studies of learning and memory (Sederberg et al., 2003; 2006; Osipova et al., 2006; Guderian, Schott, Richardson-Klavehn & Duzel, 2009; Burke et al., 2014; Long et al., 2014). Such high frequency activity is taken to be a marker of general cortical activation (e.g. Burke et al., 2014). Thus, the pattern observed is aligned with other high-level cognitive states and may indicate that engagement, as indexed by increases in gamma, is a component of successful learning and reasoning, which can explored through future empirical evaluation. No evidence of neural correlates for effort was found. While we cannot conclude from a null result that there are no neural correlates of effort, it is clear that engagement and effort can be dissociated. Together, these results provide evidence for a neural signal reflecting engagement, which could be used to assess engagement online without explicit self-report.

EEG could inform educational task design by providing novel validity support for item types designed to be more engaging. That is, if items designed to be more engaging demonstrate higher levels of an EEG measure of engagement, it would provide an additional source of validity evidence compared to self-report data. EEG measures of engagement could inform assessment design by determining the optimal spacing of items such that engagement remains consistent while workload varies according to item difficulty. Such improvements would benefit

performance and learning; high levels of task engagement are associated with positive affect, high levels of concentration and interest, and performance gains (Pintrich & De Groot, 1990). Assessment designs that promote higher and/or more consistent levels of engagement could also lead to more valid assessment data.

Our study represents an early step toward elucidating the neural correlates of engagement, and exciting avenues of future work remain. There are several limitations of the current study, which should be explored in future studies. For example, the current study restricted the EEG analyses to the 20 seconds prior to a participant's response to an item. This time period of the response might not have captured the fluctuations in power that best indexes effort. In addition, the current study was correlational in nature. Future studies should experimentally manipulate the degree of engagement of the stimuli to provoke predictable levels of high gamma power. These next steps combined with multivariate methods to assess engagement, represent just a few of the potential avenues of research based on the groundwork presented here. As a first step, we have established that an online, potentially objective neural signal of engagement exists – as reflected through modulations of high gamma power.

References

Agresti, A. (1990), *Categorical Data Analysis*, New York, NY: John Wiley & Sons.

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

Bloom, L. & Tinker, E., & Scholnick, E.K. (2001). The intentionality model and language acquisition. *Monograph of the Society for Research in Child Development*, *66,* 1-101.

Blumenfeld, P. C., Kempler, T. M., & Krajcik, J. S. (2006). *Motivation and cognitive engagement in learning environments*. In R. K. Sawyer (Ed.), The Cambridge handbook of learning sciences (pp. 475-488). New York, NY: Cambridge University Press.

Brown, A. R., & Finney, S. J. (2011). Low-stakes testing and psychological reactance: Using the Hong Psychological reactance Scale to better understand compliant and non-compliant examinees. *International Journal of Testing, 11*, 348-270. doi:10.1080/15305058.2011.570884

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *Neuroimage*, *85*, 834-843.

Chaouachi, M., & Frasson, C. (2010, June). Exploring the relationship between learner EEG mental engagement and affect. In *International Conference on Intelligent Tutoring Systems* (pp. 291-293). Springer Berlin Heidelberg.

Chapman, E. (2003). Alternative Approaches to Assessing Student Engagement Rates. *Practical Assessment, Research & Evaluation, 8,* 1-10.

Christensen, R. (2015). *Regression Models for Ordinal Data*. R package version 2015. 6-28.

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012). *Handbook of Research on Student Engagement*. Verlag, NY: Springer US.

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology*, *112*, 565-582.

Finn, B. (2010). Ending on a high note: Adding a better end to difficult study. Journal of Experimental Psychology: Learning, Memory and Cognition, 36, 1548-1553.

Finn, B., & Miele, D. B. (2016). Hitting a high note on math tests: Remembered esuccess influences test preferences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42,* 17-38.

Fitzgibbon, S. P., Pope, K. J., Mackenzie, L., Clark, C. R., & Willoughby, J. O. (2004). Cognitive tasks augment gamma EEG power. *Clinical Neurophysiology*, *115*, 1802-1809.

Freeman, F. G., Mikulka, P. J., Prinzel, L. J., & Scerbo, M. W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology*, *50*, 61-76.

Guderian, S., Schott, B., Richardson-Klavehn, A., Duzel, E., (2009) Medial temporal theta state before an event predicts episodic encoding success in humans. *Proceedings of the National Academy of Sciences*, *106*, 5365-5370.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods and Research*, *22*, 364–375.

Gruber, T., Tsivilis, D., Montaldi, D., & Müller, M. M. (2004). Induced gamma band responses: an early marker of memory encoding and retrieval. *Neuroreport*, *15*, 1837-1841.

Jensen, O., Kaiser, J., & Lachaux, J. P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, *30*, 317-324.

Jerbi, K., Ossandon, T., Hamame, C. M., Senova, S., Dalal, S. S., Jung, J., ... & Lachaux, J. P. (2009). Task-related gamma-band dynamics from an intracerebral perspective: Review and implications for surface EEG and MEG. *Human Brain Mapping*, *30*, 1758-1771.

Kahneman, D., Diener, E., & Schwarz, N. (Eds.). (1999). *Well-being: Foundations of hedonic psychology*. New York, NY: Russell Sage Foundation.

Kovach, C. K., Tsuchiya, N., Kawasaki, H., Oya, H., Howard, M. A., & Adolphs, R. (2011). Manifestation of ocular-muscle EMG contamination in human intracranial recordings. *NeuroImage*, *54*, 213-233.

Long, N. M., Burke, J. F., & Kahana, M. J. (2014). Subsequent memory effect in intracranial and scalp EEG. *NeuroImage*, *84*, 488-494.

Long, N. M. and Kahana, M. J. (2017) Modulation of task demands suggests that semantic processing interferes with the formation of episodic associations. *Journal of Experimental Psychology: Learning, Memory & Cognition. 43*, 167-176

Luck, S. J. (2014). An introduction to the event-related potential technique (2nd ed.). Cambridge, MA: The MIT press.

Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology*, *80*, 514-523.

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1084-1097.

Miller, K. J., Shenoy, P., Miller, J. W., Rao, R. P., & Ojemann, J. G. (2007). Real-time functional brain mapping using electrocorticography. *NeuroImage*, *37*, 504-507.

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of EEG*. Oxford University Press, USA.

Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E., & Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *Journal of Neuroscience, 26*, 7523-7531.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33.

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, *40*, 187-195.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*, 147-169.

Schnipke, D. L. (1995). Assessing Speededness in Computer-Based Tests Using Item Response Times. Paper presented at the annual meeting of the National Council on Measurement in Education April, 1995, San Francisco, CA

Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *The Journal of Neuroscience*, *23*, 10809-10814.

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, *32*, 1422-1431.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*, 571-581.

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In S.L. Christenson, A. L. Reschly & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 21-44). Verlag, NY: Springer US.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, *14,* 8-9.

Weidemann, C. T., Mollison, M. V., and Kahana, M. J. (2009). Electrophysiological correlates of high-level perception during spatial navigation. *Psychonomic Bulletin & Review, 16*, 313–319.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1-17.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*, 65-83.