Research Report

EEG Correlates of Engagement During Assessment

ETS RR–21-01

Laura K. Halderman
Bridgid Finn
J.R. Lockwood
Nicole M. Long
Michael J. Kahana

December 2021

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

EEG Correlates of Engagement During Assessment

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

1 Educational Testing Service, Princeton, NJ
2 University of Pennsylvania, Philadelphia, PA

In educational assessment, low engagement is problematic when tests are low stakes for students but have significant consequences for teachers or schools. In the current study, we sought to establish the electroencephalographic (EEG) correlates of engagement and to distinguish engagement from mental effort. Forty university students participated in a simulated *GRE*® General Test session while scalp EEG was recorded from 128 channels. Participants completed two verbal and two quantitative GRE test blocks for a total of 40 items each, and after each item, rated either their engagement or mental effort on a scale of 1–6. We computed power for seven frequency bands (delta, theta, alpha, beta, and low, medium, and high gamma) across six regions of interest: left hemisphere (LH) and right hemisphere (RH) frontal, temporal, and parietal. Preliminary results suggested that gamma power (30–150 hertz [Hz]) indexed differences between high- and low-engagement ratings. This pattern was similar but weaker for mental effort. A cumulative logit model with cross-classified random effects determined that high gamma (90–150 Hz) over the LH temporal cortex predicted engagement ratings, while controlling for reaction time and accuracy. However, for effort ratings, reaction time was the sole significant predictor. These results suggest that high gamma may be a correlate of engagement during complex cognitive tasks, but not a correlate of effort. The findings are a promising step toward the goal of objectively measuring engagement during assessment tasks.

Keywords EEG; engagement; effort; *GRE*® General Test; assessment

doi:10.1002/ets2.12312

Background

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 et al., 1988). Engagement, a component of motivation (Skinner & Pitzer, 2012), is a multidimensional construct with various characterizations in the achievement motivation literature (Christenson et al., 2012). Engagement can reference a student's academic engagement over an entire school year or engagement on a particular academic task. The focus of the current manuscript is on the narrower construct of task engagement, in which engagement is exemplified as a person's active involvement in a task or activity (Reeve et al., 2004), in this case a structured educational assessment activity. Our primary goal was to establish preliminary evidence that real-time measures of neural activity measured through electroencephalography (EEG) could provide a valid indicator of engagement in an assessment context. A secondary objective was to determine whether engagement during an assessment could be differentiated from mental effort. Scalp EEG was measured as students took a practice, adaptive version of the *GRE*® General Test. The investigation moved beyond the current behavioral methods typically used to index student motivation and engagement, such as posttest self-reports and item- and test-level response times, with the aim of identifying online, neural correlates of engagement during test taking.

Assessment-related activities are a large part of teaching and learning; however, there is not extensive research measuring students' engagement as they take an assessment (Nichols & Dawson, 2012). Notwithstanding the lack of research on this topic, there are critical reasons that more extensive understanding of student engagement during assessment is needed. Consider when an assessment involves low stakes for the student but significant consequences for a teacher or schools. Low engagement during the assessment results in 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

Corresponding author: L. K. Halderman, E-mail: lhalderman@ets.org

programs and teaching staff. Measuring engagement during a testing session may help identify students who are not engaged during the assessment and improve the overall validity of assessment data.

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 et al., 2006; Wise & DeMars, 2005). In general, self-report instruments ask students to indicate their self-perceived level of motivation and engagement for the test specifically and their academic motivation more generally and are often administered after the test is completed. Test-timing data include the response time to complete a question or time to finish a test, with atypically low values potentially indicating lower student engagement, effort, and motivation. These indirect measures of engagement and motivation could be used to filter scores at the student or item level to improve the validity of the reported scores.

There are several limitations to using such indirect indicators to make inferences about motivation and engagement. For example, while global self-report methods are easy to implement and score, they suffer from the pitfalls of all self-report measures, namely that they require that the student be able to accurately report on their cognitive and affective states (Swerdzewski et al., 2011). People are often inaccurate when making assessments of their own states and knowledge (Kruger & Dunning, 1999; Ziegler et al., 2011). In addition, self-report methods tend to measure engagement at a more global level rather than at the item level, which is where the task-level engagement evoked in the process of problem solving might be better captured. Self-report measures ask the student to remember or to predict their engagement over a test, which may vary in the degree to which specific items engaged the student. People can be insensitive to changes in their engagement over time (Christenson et al., 2012) and can focus on a particular phase of a longer term, extended experience when making their evaluation, such as the beginning or end of an event (Finn, 2010; Finn & Miele, 2015; Kahneman et al., 1999), which may not accurately reflect how engaged the test taker was as they answered specific test items. Furthermore, students' engagement ratings may get contaminated by response biases, such as when engagement and motivation are rated higher than what the student actually experienced in an effort to avoid an anticipated academic penalty (Wise & Kong, 2005; Wise & Ma, 2012). Inferences about engagement are also limited in the case of test-time data. Timing information offers a more fine-grained view of performance throughout a test; however, rapid responses could indicate a range of states for students: that they were not engaged, that they were hurrying to finish given time constraints, or that they knew that they did not know the answer (Schnipke, 1995; Wise & Kong, 2005).

In contrast, online measurement of engagement using physiological data such as EEG could provide an objective view into the test taker's experience and potentially yield stronger evidence of engagement than self-report and test-timing measures. Unlike self-reporting, EEG can be measured without drawing the participant's awareness away from the primary task or asking them to remember or predict their past or future states of engagement. Measuring EEG for the duration of a test additionally provides opportunities to explore item-to-item variation in engagement, which could help to identify problematic items and ultimately lead to improved assessment design. As a first step toward these potential benefits, our study sought to establish initial evidence that EEG measurements contain meaningful information about test-taker engagement during a realistic simulation of an academic testing activity.

Use of EEG to Measure Engagement

There is only modest literature on the use of EEG to measure engagement (e.g., Berka et al., 2007; Chaouachi & Frasson, 2010; Freeman et al., 1999; Pope et al., 1995). Most of the work in this area can be found in the applied human–computer interaction and attention/vigilance literatures. Measures of engagement in these domains have typically involved online cognitive state classification or identification of the extent to which a person is engaged in a task while they are performing it. Typical tasks used 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, pressing a button when a visual stimulus is flashed briefly in a specific location on the screen. Other studies have recorded EEG while participants passively view videos of varying engagement levels. Most of the engagement measures in these domains have analyzed power across a range of frequency bands but have typically focused on the lower bands (i.e., alpha, theta, and beta; Chaouachi & Frasson, 2010, Freeman et al., 1999; Pope et al., 1995).

The tasks that have been used in the attention/vigilance literature and in the human-computer interaction field are quite different from the assessment context that is typical of academic testing, however. The nature of engagement measured in attention/vigilance paradigms may be too narrow to capture the kind of cognitive engagement that is required when people must reason through complex verbal and quantitative items like those used to assess academic achievement.

Evidence from cognitive neuroscience has established gamma (>30 hertz [Hz]) as a critical frequency band for a variety of cognitive processes including attention (Jensen et al., 2007), word recognition (Jerbi et al., 2009), auditory processing (Crone et al., 2001), and motor movement (Miller et al., 2007). Most notably, gamma increases are indicative of learning mechanisms, as established through both intracranial and scalp EEG (Burke et al., 2013; Gruber et al., 2004; Long et al., 2014; Sederberg et al., 2003, 2006; and see also Fitzgibbon et al., 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 samples of participants. The researchers concluded that scalp EEG was capable of resolving high frequency gamma activity that was indicative of subsequent memory effects. Thus, including gamma band activity, in addition to power across lower frequency bands, may be a worthwhile approach to evaluate the engagement elicited in more complex cognitive tasks that involve active learning and problem solving (e.g., Blumenfeld et al., 2005).

Current Study

In the current study, EEG was measured as students completed a simulated GRE session. A primary goal was to determine whether there was an EEG correlate of engagement that consisted of significant increases and/or decreases in power in specific frequency bands in specific regions of interest (ROIs) on the scalp.

An engagement rating or a mental effort rating was taken after each item. Self-report ratings are used in this study as the initial correlate with EEG. While self-report methods are prone to measurement issues as previously discussed, they are a typical way of assessing engagement and can provide a first link to establishing an online EEG measure of engagement. Consistent with other measures of task engagement (e.g., Csikszentmihalyi, 1997; Reeve et al., 2004), engagement ratings asked participants to evaluate how absorbed, attentive, and involved they had been while answering the item. Mental effort ratings asked participants to evaluate how much mental activity was required to answer the item (e.g., calculating, remembering). A central question was whether patterns of power in specific frequency bands could distinguish between high- and low-engaging items. That is, we asked whether activity changes in particular frequency bands predict whether a person 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. The assessment was an adaptive version of the GRE, in which participants were routed to a second-stage block of low, medium, or high difficulty based on their performance on the first-stage router block. The adaptive version was used with the goal of evaluating an authentic version of the GRE and to keep effort similar across test items, whereas engagement was expected to vary.

To establish a possible EEG correlate of engagement, we conducted a series of behavioral and EEG analyses. For clarity, we preview our three primary analysis approaches below.

First, we evaluated whether engagement and effort could be distinguished behaviorally from one another. Analyses evaluated whether the two types of rating scales were being used distinctly by the participants and whether they were differentially related to important test-taking metrics such as accuracy, response time, and item difficulty. Second, we tested whether engagement and effort could be distinguished neurally. 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. The analyses narrowed in on high gamma, the frequency band that demonstrated the strongest correlation with engagement ratings. Mixed-effects modeling was used to evaluate whether EEG provided a meaningful index of engagement once important behavioral metrics of test-taker engagement, such as accuracy and reaction time, had been controlled for.

To conduct these analyses, we first determined which portion of each trial should be included in the EEG output. Because quantitative and verbal items differed in complexity and difficulty, response times varied considerably across items and between participants. To control for this variability, all EEG measurements were response-locked to the participant's first response on an item and included the 20 seconds (s) prior to the response. Though feelings of engagement and effort likely occur throughout the course of an item, this approach allowed us to keep all trials an equal length (20 s) across items and participants.

Methods

Participants

Forty students (22 female) at the University of Pennsylvania participated in a 2.5-h experimental session. The average age was 22.1 years ($SD = 3.5$ years). Participants were recruited from a pool of students who had previously completed a full neuropsychological battery and a series of 20 experimental sessions in the lab. They were screened for previous practice or experience with the GRE. Participants were included in the study if they practiced or took the GRE prior to 2011 or if they had never practiced or taken the GRE. There was no other criterion for exclusion. Participants were paid $75 for their participation.

Design and Materials

Each participant completed a simulated GRE session while EEG was recorded. Materials were taken from a practice test of the GRE, which contained a routing block and three second-stage blocks (easy, medium, and difficult) for both quantitative and verbal domains. Participants completed two verbal and two quantitative GRE test blocks for a total of 40 items in each domain. The domain of the first test block (verbal or quantitative) was counterbalanced across participants and the second block was matched in domain to the first test block. Table 1 depicts the number of items correct in the routing block required for each second-stage block assignment. Each block contained 20 items. Each domain had three item types, as is standard in the GRE practice test. Quantitative items included constructed-response items (enter the numerical solution to a problem), standard items with only one correct selected-response answer, and multiple-response items, which had multiple correct selected-response answers. Verbal items included items with subquestions (where each subquestion required an answer), a standard item with only one correct answer, and a multiple-response item (which required multiple correct answers). Items were presented in the center of the screen. Responses were made by clicking the correct response(s) or by using the keypad to enter in a constructed response. A sample item is shown in Figure 1.

Self-report prompts were presented after each GRE item. Participants were asked to rate their mental effort and engagement for the preceding item on a Likert scale of 1 – 6 with 1 being low and 6 being high. As can be seen in Table 2, which gives the distribution of each type of rating, there was no evidence of restricted range. 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 or performance?" For engagement, participants were asked to consider, "How absorbed were you while you were answering the item? How attentive and involved were you?"

Procedure

The session began with a brief tutorial. Participants were presented with six practice items to familiarize them with the different item types. After completing the practice, participants were randomly assigned to either a quantitative or verbal routing block. Performance on the routing block determined which second-stage block assignment participants received (see Table 1 for routing assignment specifications). Upon completing the routing block, participants were randomly assigned to either a verbal or quantitative second block. Within each block, item types (described above) but not individual items were randomly presented. Thus, items of the same type were presented together but the different item types appeared in a random order across participants. After each GRE item, a mental effort or engagement self-report prompt was presented with an equal number of items getting an effort or engagement prompt. Participants were given a 10-min break between the second and third blocks.

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 fewer	Easy
	7 – 11	Medium
	12 or more	Hard
Verbal	6 or fewer	Easy
	7 – 12	Medium
	13 or more	Hard

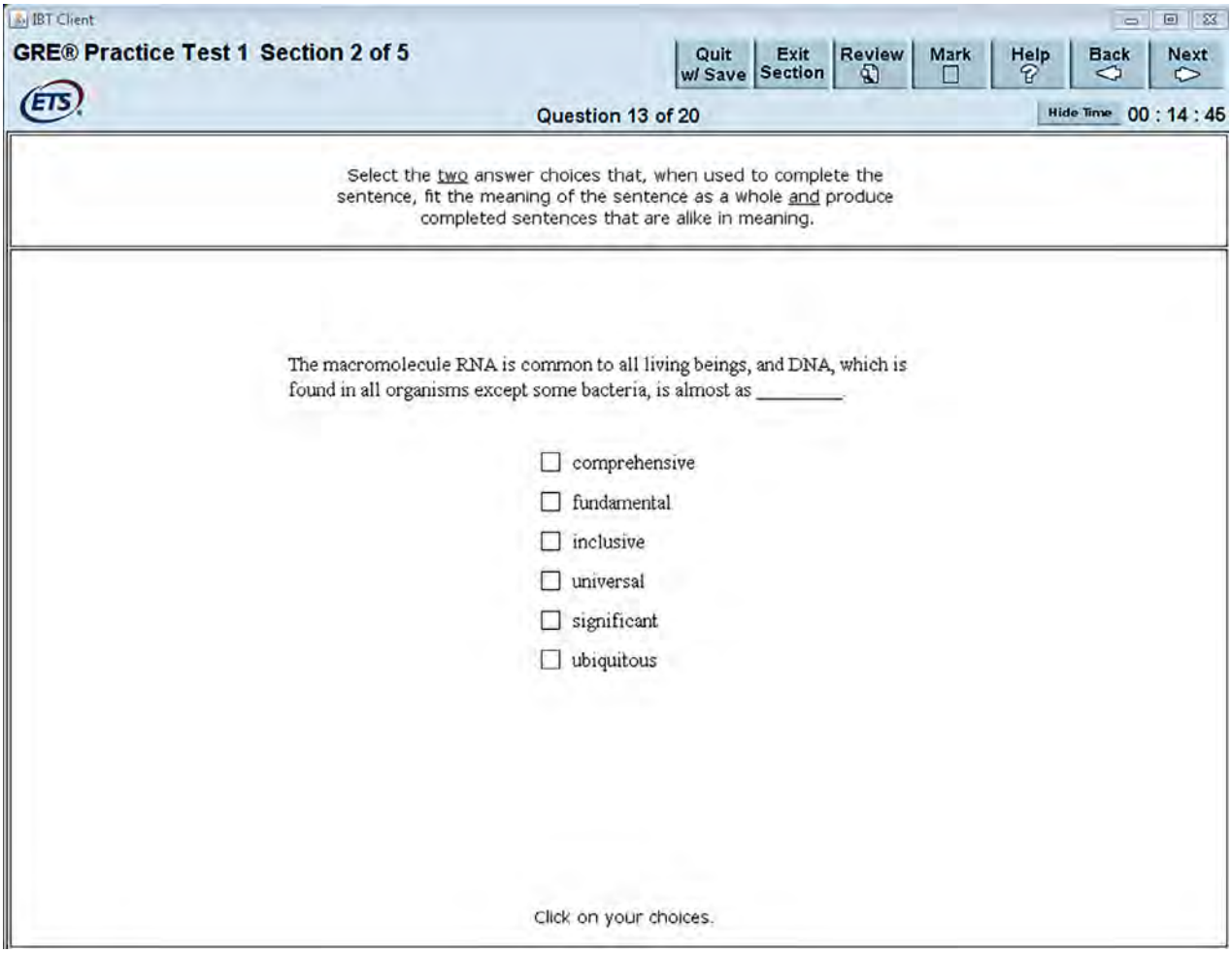

Figure 1 Sample multiple-select GRE practice verbal item.

Table 2 Distribution of the Number of Participant Responses for Effort and Engagement Ratings

Rating	Effort	Engagement
1	99	88
2	318	260
3	346	313
4	272	268
5	164	221
6	96	135
Total	1,295	1,285

EEG Data Collection

EEG was recorded using a Geodesic Sensor Net (GSN; Net Station 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 the Cz electrode and later converted to a bipolar reference. As in Long et al. (2014) bad channels were not removed from the bipolar-referenced data as there were typically zero bad channels marked for a given subject. In addition, ROIs were composed of multiple pairs of electrodes, and so it was highly unlikely that all would be bad simultaneously.

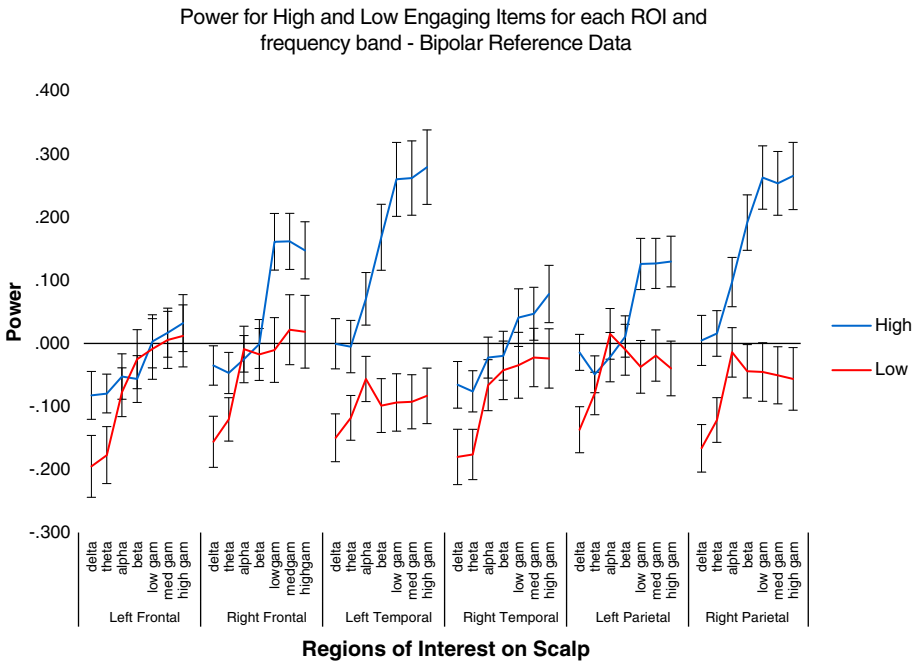

Figure 2 Means for high- and low-engaging items at each region of interest (ROI) for each frequency band using bipolar referenced data.

EEG Data Processing

To minimize confounds resulting from volume conduction and saccades, we analyzed the scalp EEG with bipolar referencing (Kovach et al., 2011; Nunez & Srinivasan, 2006). We defined the bipolar montage in our data set 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 (Burke et al., 2013). The resulting bipolar signals were treated as new virtual electrodes and are referred to as such in the remainder of the text. Standard EEG processing techniques were used to detect and extract artifacts, including muscle noise (e.g., Long et al., 2014).

We applied the Morlet wavelet transform (Wave Number 6) to all bipolar electrode EEG signals during 20 s response-locked epochs, across 52 logarithmically spaced frequencies (2 – 165 Hz). We included a 1,000 milliseconds (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 nonoverlapping 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 that 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.

We defined six a priori ROIs based on previous scalp EEG studies of human memory (Long et al., 2014; Weidemann et al., 2009). *Z*-power in each frequency band was averaged across our six ROIs: left hemisphere (LH) and right hemisphere (RH) frontal, temporal, and parietal (Figure 2). This was done for each GRE item. Items with response times less than 20 s on the first encounter were excluded from the analyses. All other items were included.

Results

An average of 14.7 questions ($SD = 3.5$) were answered correctly on the quantitative routing block and 11.0 ($SD = 3.2$) 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 the quantitative items in an average of 73.2 s ($SD = 45.9$) and the verbal items in an average of 58.2 s ($SD = 40.9$).

Table 3 Distribution of Second-Stage Block Assignments

	Second-stage blocks		
Domain	Easy	Medium	Hard
Quantitative	1	5	34
Verbal	2	15	23

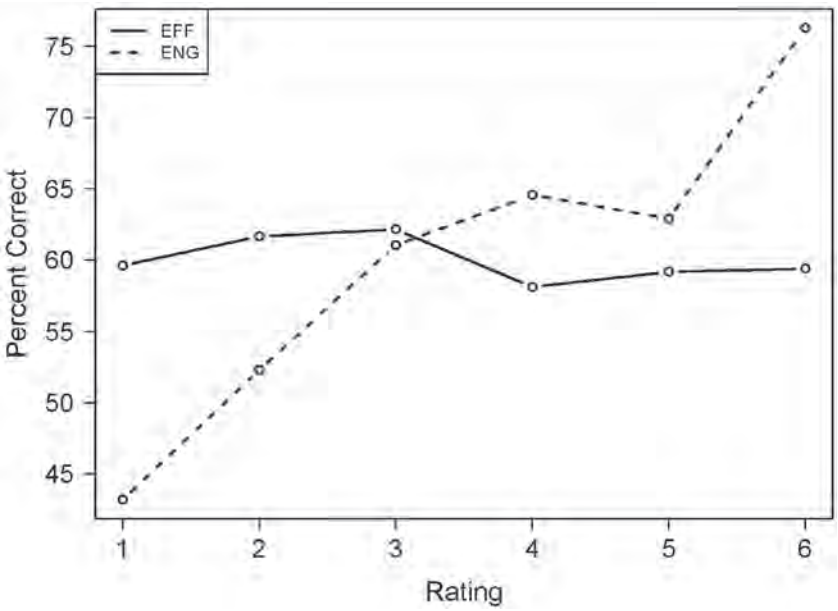

Figure 3 Percentage correct by rating level, for effort and engagement ratings.

Distinguishing Engagement From Effort

As a first step, we sought to determine whether engagement and effort could be distinguished behaviorally from one another. In the educational literature, effort is sometimes used as a proxy for engagement (as measured through response time), with greater effort taken to indicate greater engagement (Wise & DeMars, 2005). However, other developmental and education researchers, such as Bloom and collaborators (Bloom et al., 2001; Bloom & Beckwith, 1989) have argued that engagement and effort are distinct constructs. Our aim was to keep effort relatively constant over the course of the test by using an adaptive version of the GRE in which students were given a second-stage block appropriate for their skill level based on performance on the router block.[1]

Recall that each record in the data set had a participant's response to either an effort or engagement self-report prompt, but not both. That is, for each item that a participant responded to, the participant was asked to report on exclusively either effort or engagment. This allowed us to examine patterns and engagement broadly across the assessment, but not on an individual item basis. Exploratory analyses were conducted to evaluate whether the two types of rating scales were being used distinctly by the participants. Differences would suggest that the ratings were indeed measuring distinctive constructs. The linear mixed-effects model routine *lme* (Pinheiro & Bates, 2000) for the R environment was used to conduct a simple variance decomposition of the ratings, treating them as intervally scaled for this purpose and using random effects for participants. The percentage of the total variance in the engagement ratings that was between participants was about 18%, compared to only about 10% for the effort ratings. That is, there was more variance among participants in their engagement ratings than in their effort ratings. This pattern was in line with the use of an adaptive test design in which effort was expected to vary less over the course of the test than engagement.

Further support for differentiation in participants' responses to the engagement and effort prompts was found with a linear model for the ratings fit with three categorical predictors: the rating type (two levels, effort vs. engagement), participant (40 levels, one per participant), and item (162 levels, one per item). The model was then augmented with

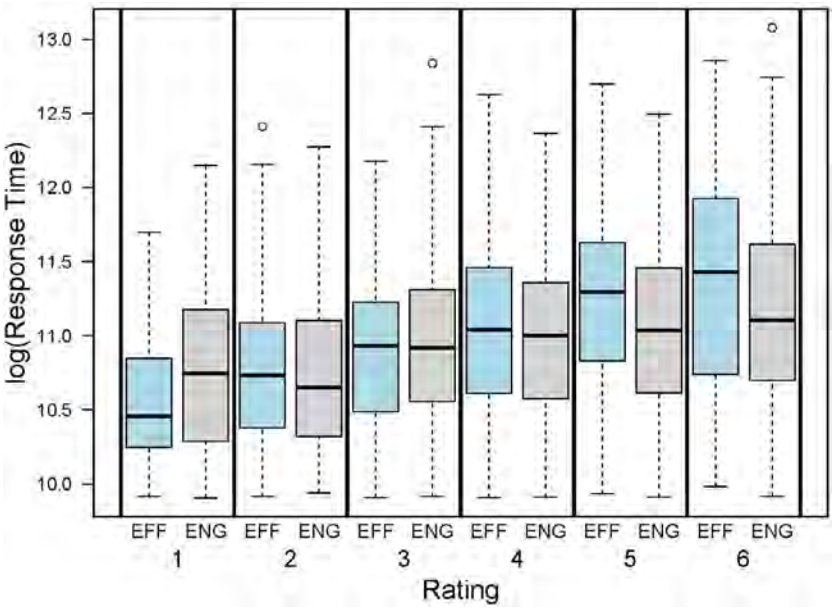

Figure 4 Box plots of log response time by rating level and rating type.

the interaction between rating type and participant. The model with the interactions was preferred ($F = 3.48$, $p < .001$), suggesting that participants were differentiating their responses to the two types of prompts.

The two types of ratings also related differently to performance. Figure 3 plots the percent correct as a function of the rating value for each type of rating. For example, for a rating of 3, the effort point indicates the percentage of all questions that were answered correctly for which subjects rated their effort as 3. As can be seen by looking across the rating range, the relationship between effort ratings and percent correct was effectively flat at approximately 60% correct regardless of effort rating. On the other hand, there was a strong positive relationship between engagement ratings and percent correct, ranging from 43% correct for items with the lowest engagement rating to 76% for items with the highest engagement rating. Participants were more likely to get an item correct when they reported having been highly engaged while responding to the item.

While the relationship between engagement and performance was in line with much research showing that test performance is affected by student engagement (e.g., Wise & DeMars, 2005), the result should be interpreted with caution. Although participants were asked to report the engagement they experienced during the process of answering the item, the ratings were taken after the item response had been made. Participants were not given feedback about their performance, but they may have had a sense of whether or not they had answered correctly or not. It is possible that the positive relationship between engagement and performance was an indication that participants were using an evaluation of whether their response had been correct or not as the driver of their engagement ratings rather than it being an indication that higher engagement was related to better performance. The analysis did, however, provide additional evidence that participants were using the engagement and effort ratings distinctly.

The two types of ratings also related differently to the time participants took to respond to the items. Figure 4 provides box plots of the natural log of the response time by each rating and rating type. Both types of ratings were positively related to response time, but the relationship between effort ratings and response time was stronger. Indeed, the R^2 of a regression of log response time on a categorical variable for the rating was .062 for the engagement ratings, but .133 for the effort ratings. Thus, the effort ratings explained about twice as much variation in log response time than did engagement ratings. One explanation is that response time was a cue that was utilized more when making evaluations of the effort expended to answer the question than for evaluating how engaged one had been as they answered the question.

Finally, there is some evidence that the two types of ratings relate differently to item difficulty. The correlation between engagement ratings and item difficulty was .06 while the correlation between effort ratings and item difficulty was .16. The stronger correlations for effort ratings with time spent and item difficulty is consistent with the idea that the items that were more difficult should require both more effort and longer response time.

Table 4 Average Correlations Between Engagement Ratings and Bipolar Referenced Power in the Seven Frequency Bands in Each of the Six Scalp Regions of Interest

	Left hemisphere			Right hemisphere		
Frequency band	Frontal	Temporal	Parietal	Frontal	Temporal	Parietal
Delta	.053	.076	.090	.083	.051	.104
Theta	.044	.055	.014	.046	.040	.090
Alpha	.010	.072	−.036	−.014	.006	.062
Beta	−.031	.152*	.010	.010	−.009	.141*
Low gamma	−.010	.192*	.091	.119	.012	.167*
Medium gamma	−.004	.193*	.087	.124	.013	.169*
High gamma	−.010	.195*	.080	.087	.019	.162*

*Indicates a significant *p* value after a Bonferroni correction for 42 comparisons.

Table 5 Average Correlations Between Effort Ratings and Bipolar Referenced Power in the Seven Frequency Bands in Each of the Six Scalp Regions of Interest

	Left hemisphere			Right hemisphere		
Frequency band	Frontal	Temporal	Parietal	Frontal	Temporal	Parietal
Delta	−.023	−.014	−.006	−.026	.020	−.007
Theta	.015	.012	−.003	.005	.045	−.001
Alpha	.034	.055	.015	.039	.025	.061
Beta	−.011	.068	.023	.027	.005	.082
Low gamma	−.026	.070	.059	.053	.002	.088
Medium gamma	.003	.084	.093	.080	.016	.096
High gamma	.0003	.072	.066	.054	.010	.089

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

Relating Ratings to EEG Measurements

Having argued that effort and engagement are behaviorally distinguishable, we next sought to assess the neural correlates of engagement and effort, as well as test whether the two cognitive states are neurally dissociable.

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 frequency bands so that the results were not preemptively constrained. Data were processed using a bipolar reference to minimize contribution of eye blinks and eye movements to the EEG signal, a method used previously to resolve high gamma signals in scalp EEG (Long et al., 2014).

Once these correlations were computed for each participant, one-tail *t* tests were conducted on the entire sample to determine whether the average correlations were significantly different from zero. See Table 4 for the full list of average correlations and significance information. In this and in the analyses that follow, a Bonferroni correction was used to correct for multiple comparisons. The correlational analyses showed that greater power for low, medium, and high gamma frequencies in the LH temporal and RH parietal electrodes was associated with higher engagement ratings. LH temporal and RH parietal 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.

A 2 (engagement: high, low) × 6 (ROI: LH and RH frontal, temporal, parietal) × 7 (frequency band: delta, theta, alpha, beta, low gamma, medium gamma, high gamma) repeated measures analysis of variance (ANOVA) was conducted to confirm the pattern established in the correlational analysis. Engagement ratings were combined into two categories, high and low. Ratings of 4–6 were categorized as high, and ratings of 1–3 were categorized as low. The engagement x ROI x frequency band interaction was significant ($F(4.23, 166.39) = 4.94$, $MSE = .10$, $p(\text{GG}) < .005$, Greenhouse–Geisser correction applied) indicating that both ROI and frequency modulated the engagement effect. Figure 2 shows the power for each level of engagement at each ROI location and for each frequency band. In addition, visual inspection of the error bars for high and low engagement show a similar pattern to the one determined in the correlation analysis: greater power differences for high compared to low engaging items in the LH temporal and RH parietal ROIs for the gamma frequency

Table 6 Estimated Regression Coefficents From Cumulative Logit Model of Engagement Ratings

Regressor	Estimate	*SE*	*z* value	Pr (>|*z*|)
Item accuracy	0.535	0.119	4.49	<.001
Log response time	0.840	0.102	8.25	<.001
LF high gamma	−0.167	0.108	−1.55	0.12
RF high gamma	−0.080	0.124	−0.65	0.52
LT high gamma	0.364	0.106	3.44	<.001
RT high gamma	0.013	0.101	0.13	0.90
LP high gamma	−0.031	0.120	−0.25	0.80
RP high gamma	0.021	0.107	0.20	0.84

Note. LF = left hemisphere frontal, RF = right hemisphere frontal, LT = left hemisphere temporal, RT = right hemisphere temporal, LP = left hemisphere parietal, RP = right hemisphere parietal.

bands. Power in the beta frequency band was also greater for high compared to low engaging items in the LH temporal and RH parietal ROIs. See Table 4 for the full list of correlations and significance information for effort ratings.

The same correlations were calculated between effort ratings and the power across frequency bands and ROIs for each participant. In contrast to engagement ratings, we found no significant correlations between ratings and power in any of the frequency bands in any of the specified ROIs. See Table 5 for the full list of correlations and significance information for effort ratings. As no significant correlations were found between ratings and power, 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. The findings provide strong support for further exploration of the difference in patterns between engagement and effort and the factors that modulate these differences.

Our next goal was to establish a link between EEG signals and cognitive states when accounting for behavioral measures that have typically been used to index engagement. Here we focus exclusively on high gamma (90 – 150 Hz) because this frequency band had the strongest correlation with engagement ratings and the most robust mean differences in the ANOVA analysis; analogous results for other frequency bands are provided in the appendix. We used mixed-effects models to test whether high gamma activity reliably predicts engagement or effort ratings after controlling for reaction time and accuracy. The engagement ratings model is presented first, followed by the corresponding results for the Effort ratings.

Modeling Engagement and Effort Ratings

Effort and engagement ratings were modeled separately. Engagement is reported first. Our model for the ordered categorical engagement ratings was a function of: (a) response accuracy, (b) log response time (mean-centered across all observations with engagement ratings), (c) high gamma activity from the six ROIs entered as six individual variables, (d) random effects for participants, and (e) random effects for items. The 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. The random effects for participants account for any persistent differences across participants in the levels of their engagement ratings, which could reflect a combination of persistent differences across participants in their levels of engagement and persistent differences across participants in their use of the ordinal rating scale. 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 estimated regression coefficients from the model are shown in Table 6. The model was parameterized such that a positive coefficient indicates that an increase in a variable was associated with an increased probability of a higher engagement rating, among the ordered 1 – 6 ratings. The coefficients were scaled such that their values indicate the change in log odds of the engagement rating being in a higher rating category (1 – 6 rating) associated with a one unit change in the variable, holding other variables constant.

The coefficients for response accuracy and the log response time indicate that each was positively associated with higher engagement ratings, which is consistent with the descriptive results presented previously. The results further demonstrate

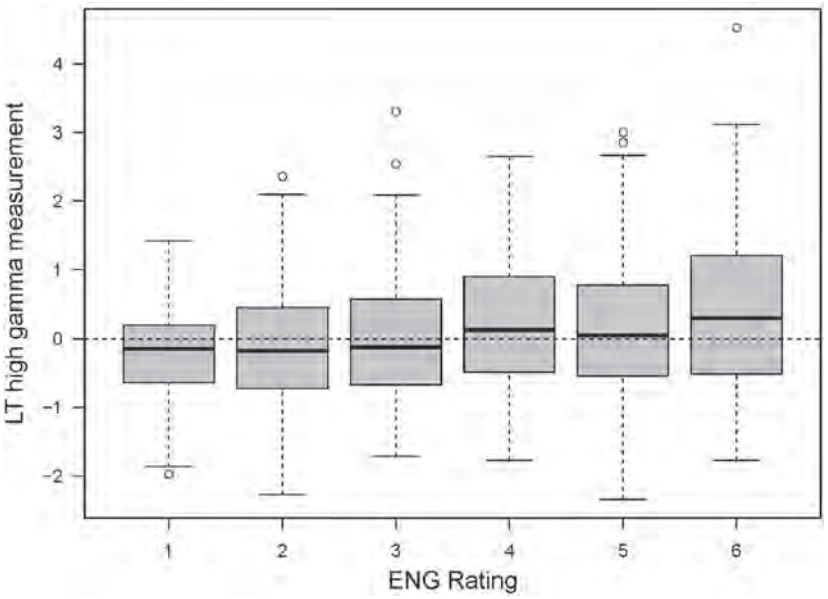

Figure 5 Box plots of the left *hemisphere* temporal high gamma measurements separately for each level of the engagement (ENG) rating.

Table 7 Estimated Regression Coefficents From Cumulative Logit Model of Effort Ratings

Regressor	Estimate	*SE*	*z* value	Pr (>|*z*|)
Item accuracy	−0.021	0.118	−0.18	0.86
Log response time	1.506	0.113	13.31	<.001
LF high gamma	−0.006	0.105	−0.06	0.95
RF high gamma	0.022	0.117	0.19	0.85
LT high gamma	−0.018	0.098	−0.19	0.85
RT high gamma	−0.074	0.105	−0.70	0.48
LP high gamma	0.050	0.119	0.42	0.68
RP high gamma	0.078	0.099	0.79	0.43

Note. LF = left hemisphere frontal, RF = right hemisphere frontal, LT = left hemisphere temporal, RT = right hemisphere temporal, LP = left hemisphere parietal, RP = right hemisphere parietal.

that high gamma in the LH temporal ROI had a significant positive relationship with the likelihood of reporting higher engagement. Figure 5 provides box plots of the LH temporal high gamma separately for each level of the engagement rating. The distributions notably shift upward as the rating increases, consistent with the model results. Finally, the finding of a statistically significant positive association between LH temporal high gamma and the engagement rating was robust to several sensitivity analyses, itemized in the appendix. 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 (likelihood ratio [LR] statistic 28.4 on 6 df, $p = 8e\text{-}5$), providing further evidence that high gamma activity provides information about engagement above and beyond the traditional proxies.

We followed a parallel process for the analysis of the effort ratings. The coefficients reported in Table 7 are analogous to those reported in Table 6. From Table 7, response accuracy was not a significant predictor of the effort rating, 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 reported in Table 7 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 the item response covariates (response accuracy and log response time). The same sets of sensitivity analyses were performed for the effort ratings as

for the engagement ratings. In all cases the EEG measurements demonstrated no statistically significant relationship with the effort ratings.

General Discussion

We found that EEG measurements from specific frequency bands and in specific ROIs, taken while participants were completing items that require reasoning in verbal and quantitative domains in an adaptive version of the GRE, were related to participants' self-reported engagement in those tasks. Greater power was found for high- compared to low-engaging items in the LH temporal areas of the scalp for gamma frequency bands. This pattern was initially found using a correlational analysis approach and confirmed using ANOVA. Power increases in the gamma frequency band have previously been associated with cognitive processes such as learning and subsequent recall for studied items (Long et al., 2014). Thus, the pattern observed in this study is aligned with other high-level cognitive states and may be a component of what is necessary for learning to occur and what supports reasoning with academic content. Future studies would need to empirically tease this apart.

The design also allowed us to evaluate whether there was a distinction between engagement and effort. Numerous analyses of the two rating types suggested that participants were using the engagement and effort ratings to reference different cognitive states. Engagement showed a positive relationship with answer correctness, whereas effort ratings were more closely aligned with item difficulty and response time. The finding of an increase in the likelihood of answering correctly when the student indicated higher test-taking engagement corresponds with the literature in educational measurement literature showing a tight link between engagement and performance (Wise et al., 2006). Cumulative logit models incorporating the EEG measurements showed that high gamma in LH temporal regions was positively associated with engagement ratings. In contrast none of the EEG measurements for high gamma in LH temporal regions was significantly related to effort ratings. Together these findings confirm that it is possible to differentially target engagement and effort in an authentic assessment context.

There are several implications and opportunities that follow from having a neural measure of student engagement. It is unlikely that EEG will be used in classrooms in the near future to assess engagement. However, future studies could explore how signals from EEG could refine models of behavioral data that can be more easily acquired, especially as testing has been moving to computer formats. In addition, EEG measures of engagement could provide novel methods of informing educational task design by providing validity support for learning and assessment item types that are designed to be more engaging. If items designed to be more engaging demonstrate higher levels of an objective EEG measure of engagement, this could be viewed as stronger validity evidence compared to self-report data. EEG measures of engagement could potentially inform assessment design by determining the optimal spacing of items and section ordering such that engagement remains consistent while workload varies according to item difficulty. These kinds of improvements to test engagement would benefit performance and learning; high levels of task engagement are associated with positive affect, high levels of concentration and interest, and gains to performance (Pintrich & de Groot, 1990).

Future work should confirm the findings discovered here by experimentally manipulating high- and low-engaging contexts and demonstrating that high gamma power is modulated in a predictable way. Future studies could also move toward pattern classification, using data acquired in this study to train a classifier to assess whether people are or are not engaged in a similar task. A key limitation of our study in this regard is that we did not monitor (e.g., video record) participant behavior during the assessment. Such process data collected in a future study could be used to relate observations about changes in examinee behavior during the task, and judgments about changes in examinee engagement during the task, to changes in EEG measurements. These analyses could strengthen the case for an association between EEG measurements and engagement beyond what was possible in this study, which relied on relating summative self-reports of engagement to averages of EEG measurements over a segment of the response period. In addition, EEG analyses and future experimental methods should be broadened to examine other regions and frequencies (e.g., perhaps using finer subdivisions of the frequency spectrum than the seven bands considered here), particularly as they relate to effort during the course of the assessment. Finally, we acknowledge our chosen methods and resultant findings represent just the first step toward a comprehensive understanding of the use of EEG to index engagement involved in higher order cognition. Our goal was to conduct an initial study aimed at establishing the neural correlates of engagement under conditions that are relevant for academic testing situations. We believe these preliminary findings can be used as the foundation for future research on the benefits of using EEG to measure complex task-level engagement.

Acknowledgments

Laura Halderman and Bridgid Finn contributed equally to this work.

Notes

1 A repeated measure of analysis of variance (ANOVA) of effort ratings over the four testing blocks showed that there was not a significant difference in ratings over blocks, confirming that effort was relatively consistent over the course of the test.

References

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Olmstead, R. E., Tremoulet, P. 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*(Supplement 5), B231-B244.

Bloom, L., & Beckwith, R. (1989). Talking with feeling: Integrating affective and linguistic expression in early language development. *Cognition & Emotion*, *3*(4), 313–342. https://doi.org/10.1080/02699938908412711

Bloom, L., Tinker, E., & Scholnick, E. K. (2001). The intentionality model and language acquisition: Engagement, effort, and the essential tension in development. *Monograph of the Society for Research in Child Development*, *66*(4), 1–101. http://www.jstor.org/stable/3181577

Blumenfeld, P. C., Kempler, T. M., & Krajcik, J. S. (2005). Motivation and cognitive engagement in learning environments. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 475–488). Cambridge University Press. https://doi.org/10.1017/CBO9780511816833.029

Burke, J. F., Zaghloul, K. A., Jacobs, J., Williams, R. B., Sperling, M. R., Sharan, A. D., & Kahana, M. J. (2013). Synchronous and asynchronous theta and gamma activity during episodic memory formation. *The Journal of Neuroscience*, *33*(1), 292–304. https://doi.org/10.1523/JNEUROSCI.2057-12.2013

Chaouachi, M., & Frasson, C. (2010). Exploring the relationship between learner EEG mental engagement and affect. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring system: 10th international conference, ITS 2010, Pittsburgh, PA, USA, June 14–18, 2010, proceedings, Part I* (pp. 291–293). Springer. https://doi.org/10.1007/978-3-642-13437-1_48

Christensen, R. (2015). *Regression models for ordinal data* (R package version 2015.6-28). [Computer software]. http://cran.nexr.com/web/packages/ordinal/index.html

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012). *Handbook of research on student engagement*. Springer. https://doi.org/10.1007/978-1-4614-2018-7

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology*, *112*(4), 565–582. https://doi.org/10.1016/S1388-2457(00)00545-9

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1548–1553. https://doi.org/10.1037/a0020605

Finn, B., & Miele, D. B. (2015). Hitting a high note on math tests: Remembered success influence test preferences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *42*(1), 17–38. https://doi.org/10.1037/xlm0000150

Fitzgibbon, S. P., Pope, K. J., Mackenzie, L., Clark, C. R., & Willoughby, J. O. (2004). Cognitive tasks augment gamma EEG power. *Clinical Neurophysiology*, *115*(8), 1802–1809. https://doi.org/10.1016/j.clinph.2004.03.009

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*(1), 61–76. https://doi.org/10.1016/S0301-0511(99)00002-2

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, *22*(3), 364–375. https://doi.org/10.1177/0049124194022003005

Gruber, T., Tsivilis, D., Montaldi, D., & Müller, M. M. (2004). Induced gamma band responses: An early marker of memory encoding and retrieval. *Neuroreport*, *15*(11), 1837–1841. https://doi.org/10.1097/01.wnr.0000137077.26010.12

Jensen, O., Kaiser, J., & Lachaux, J. P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, *30*(7), 317–324. https://doi.org/10.1016/j.tins.2007.05.001

Jerbi, K., Ossandón, T., Hamamé, C. M., Senova, S., Dalal, S. S., Jung, J., Minotti, L., Bertrand, O., Berthoz, A., Kahane, P., & 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*(6), 1758–1771. https://doi.org/10.1002/hbm.20750

Kahneman, D., Diener, E., & Schwarz, N. (Eds.). (1999). *Well-being: The foundations of hedonic psychology*. 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*(1), 213 – 233. https://doi.org/10.1016/j.neuroimage.2010.08.002
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121 – 1134. https://doi.org/10.1037/0022-3514.77.6.1121
Long, N. M., Burke, J. F., & Kahana, M. J. (2014). Subsequent memory effect in intracranial and scalp EEG. *NeuroImage*, *84*(1), 488 – 494. https://doi.org/10.1016/j.neuroimage.2013.08.052
Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology*, *80*(4), 514 – 523. https://doi.org/10.1037/0022-0663.80.4.514
Miller, K. J., Shenoy, P., Miller, J. W., Rao, R. P., & Ojemann, J. G. (2007). Real-time functional brain mapping using electrocorticography. *NeuroImage*, *37*(2), 504 – 507. https://doi.org/10.1016/j.neuroimage.2007.05.029
Nichols, S. L., & Dawson, H. S. (2012). Assessment as a context for student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 457 – 477). Springer. https://doi.org/10.1007/978-1-4614-2018-7_22
Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of EEG* (2nd ed.). Oxford University Press.
Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer. https://doi.org/10.1007/978-1-4419-0318-1
Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33 – 40. https://doi.org/10.1037/0022-0663.82.1.33
Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, *40*(1 – 2), 187 – 195. https://doi.org/10.1016/0301-0511(95)05116-3
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed., Vol. 1). Sage.
Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*(2), 147 – 169. https://doi.org/10.1023/B:MOEM.0000032312.95499.6f
Schnipke, D. L. (1995, April 19 – 21). *Assessing speededness in computer-based tests using item response times* [Paper presentation]. The annual meeting of the National Council on Measurement in Education, San Francisco, CA, United States.
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*(3), 1422 – 1431. https://doi.org/10.1016/j.neuroimage.2006.04.223
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*(34), 10809 – 10814. https://doi.org/10.1523/JNEUROSCI.23-34-10809.2003
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). Springer. https://doi.org/10.1007/978-1-4614-2018-7_2
Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motivation. *Assessment Update*, *14*(1), 8 – 9.
Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162 – 188. https://doi.org/10.1080/08957347.2011.555217
Weidemann, C. T., Mollison, M. V., & Kahana, M. J. (2009). Electrophysiological correlates of high-level perception during spatial navigation. *Psychonomic Bulletin & Review*, *16*(2), 313 – 319. https://doi.org/10.3758/PBR.16.2.313
Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1 – 17. https://doi.org/10.1207/s15326977ea1001_1
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163 – 183. https://doi.org/10.1207/s15324818ame1802_2
Wise, S. L., & Ma, L. (2012, April 12 – 16). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper presentation]. The annual meeting of the National Council on Measurement in Education, Vancouver, Canada.
Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65 – 83. https://doi.org/10.1207/s15326977ea1101_3
Ziegler, M., MacCann, C., & Roberts, R. (Eds.). (2011). *New perspectives on faking in personality assessment*. Oxford University Press.

Appendix

Sensitivity Analyses Conducted With Engagment Rating Models

We conducted various sensitivity analyses to verify that the association between engagement ratings and high-gamma EEG measurements from the LT region reported in Table 6 was stable. We list the sensitivity analyses here, with the estimated regression coefficient and *p* value for each analysis provided in parentheses.

- Including itemseq (the variable indicating the order in which each subject took the items) in the model and also allowing a random slope on this variable for subjects, to account for possible subject-specific, systematic changes in ENG ratings over the course of the exam period (.36, $p = .001$).
- The removal of a record that has a relatively high value of the LT EEG measurement of 4.52, whereas the next highest value is 3.30 (.35, $p = .0008$).
- Removing all records for the subject in the previous item (.33, $p = .002$).
- Using fixed effects for itemid rather than random effects. This is a stronger control for item attributes than provided by random effects and implicitly accounts for the relationship of any item attributes such as domain, topic, difficulty, discrimination, etc. to the ENG ratings (.38, $p = .0007$).
- The removal of the item response covariates of whether the item was answered correctly, and the log response time (.46, $p < .0001$).
- Using a probit rather than logit link to model the cumulative probabilities of the ENG rating (.21, $p = .0005$); however, note that the scale of the coefficient for the probit model is not directly comparable to that for the logit model).
- Treating the ENG ratings as continuous and intervally scaled rather than ordinal, in which case a standard cross-classified mixed effects model using the lme4 package (Bates et al., 2015) can be fit (.24, $p = .001$); however note that the scale of the coefficient for this model is also not directly comparable to that for the logit model).
- Including records for item responses that took at least 10 s, rather than the 20 s threshold used in the main analysis (.35, $p = .0003$). This analysis included 3,057 records, an 18% increase over the 2,580 records used in the main analysis.

Sensitivity Analyses Conducted With Different Frequency Bands

The results reported in Tables 6 and 7 use EEG measurements from the high-gamma frequency band taken from six regions. Those results demonstrate a strong relationship between LT-region measurements in the high-gamma frequency band, and engagement ratings, but no relationship with effort ratings. This appendix reports results for analogous models fit separately to each of the other six frequency bands (delta, theta, alpha, beta, low-gamma, mid-gamma). For example, for engagement ratings, we fit a model analogous to that reported in Table 6, but using delta-frequency measurements from the six regions in place of the high-gamma-frequency measurements. We replicated this procedure using each of the frequency bands. The coefficients on the LT region measurement for these alternative models are .05, .07, .20, .28, 29, and.33 for the delta, theta, alpha, beta, low-gamma, and mid-gamma frequencies, respectively. The corresponding value for the high gamma region from Table 6 is .36. Estimated coefficients from the beta frequency band upward are statistically significant at level alpha = 0.05 using a two-sided test. There is a clear trend of increasing strength of relationship with increasing frequency.

We conducted the analogous procedure for the effort ratings. The coefficients on the LT region measurement for the alternative models are .09, .12, .03, .02, −.06, and − .02 for the delta, theta, alpha, beta, low-gamma, and mid-gamma frequencies, respectively. The corresponding value for the high gamma region from Table 7 is −.02. There is no clear trend in the estimated coefficients across frequency bands, and none are statistically significant at level alpha = 0.05 using a two-sided test.

Suggested citation:

Halderman, L. K., Finn, B., Lockwood, J. R., Long, N. M., & Kahana, M. J. (2021). *EEG correlates of engagement during assessment* (Research Report No. RR-21-01). Educational Testing Service. https://doi.org/10.1002/ets2.12312

Action Editor: John Sabatini and John Mazzeo

Reviewers: Madeleine Keehner and Blair Lehman