Ruslan Suvorov, Amber Camp

University Of Hawaii At Manoa, Honolulu, USA

rsuvorov@hawaii.edu; acamp@hawaii.edu

An exploratory study of test-takers' cognitive states using electroencephalography (EEG) and retrospective interviews in the context of L2 testing

Bio data

Ruslan Suvorov is Language Technology Specialist at the Center for Language & Technology at the University of Hawai‘ i at Mānoa. His research interests lie at the intersection of applied linguistics and instructional technology and design, with a focus on language assessment, eye tracking, and blended and online learning. He is a co-author of *Blended language program evaluation* (Palgrave Macmillan, 2016).

Amber Camp is a PhD student in the Department of Linguistics at the University of Hawai‘ i at Mānoa. Her primary interests lie in psycholinguistics, first & second language acquisition, and experimental phonetics, with additional focus on using psycholinguistic testing to inform language teaching and assessment.

Abstract

Traditionally, learners' cognitive states and processes in the context of second language (L2) testing have been studied through concurrent and retrospective verbal reports (e.g., Phakiti, 2003; Plakans, 2009). While perceptual data elicited through these methods can provide insight into learners' cognition, their validity is oftentimes questionable due to a potential mismatch between what L2 learners report and what they actually do during specific L2 tasks (Cohen, 2014). More direct evidence can be obtained through electroencephalography (EEG) technology that can directly measure and record cognitive states and processes of L2 learners engaged in language tasks. While EEG research is relatively common in neuroscience and related fields, its application in L2 testing to study learners' cognition seems to be virtually non-existent.

We will report on the results of an exploratory study with 11 non-native speakers of English who were asked to complete six video-based academic listening tasks while wearing a 14-channel EEG headset that measured and recorded their brain signals. Conceptualized as a multiple-case study, this project investigated the relationship between EEG-derived measures of L2 learners' cognitive states of "interest" and "engagement" during these tasks and (a) the learners' self-reported perception of these states and (b) the learners' performance (i.e., scores) on questions associated with the tasks. Correlations calculated to determine these two types of relationships were supplemented with qualitative data from retrospective interviews.

The local context of this study comprised a university setting where students who are non-native speakers of English are required to meet a language proficiency requirement by passing a language proficiency test or enrolling in English language classes. Given the

diversity of the international student population in the local context, this project employed a multiple-case study design that (a) allowed a more in-depth and systematic exploration of the two cognitive states of each individual learner, and (b) identified the aspects of the data that transpire across the majority of cases and can be used to make generalizations to more global contexts. Furthermore, the local context prompted the use of video-based listening assessment tasks (rather than audio-only tasks) in this study because they more genuinely reflect the conditions of the academic environment wherein local learners are exposed to both the visual and the auditory input. Findings of this study can be generalized to all contexts of academic listening assessment, as well as the context of computer-assisted language learning in both local and distance learning scenarios.

Conference paper

Introduction

The overall purpose of this exploratory study was to investigate second language (L2) learners' cognitive processes during computer-based L2 assessment tasks and to explore how these processes relate to L2 learners' performance on such tasks.

Traditionally, learners' cognitive processes during L2 assessment have been studied through concurrent and retrospective verbal reports such as interviews (Phakiti, 2003; Plakans, 2009), questionnaires (Kashkouli & Barati, 2013), and think-aloud protocols (Anderson, Bachman, Perkins, & Cohen, 1991; Cohen & Upton, 2007; Plakans, 2009). While perceptual data elicited through these methods can provide insight into learners' cognition, the validity of such self reports is oftentimes questionable (Bowles, 2010) due to a potential mismatch between what L2 learners *report* they do and what they *actually* do during specific L2 tasks (Cohen, 2014). More direct evidence of L2 learners' mental activity can be obtained through neurophysiological measures such as electroencephalography (EEG).

EEG technology uses electrodes to non-intrusively measure and record the user's brain activity during a specific task. In the context of L2 learning and assessment, EEG data can furnish information about latent (i.e., non-observable) traits related to L2 learners' use of the target language and serve as a direct measure of L2 learners' cognitive responses to a particular stimulus. While EEG as a neurophysiological research method is relatively common in neuroscience and related interdisciplinary fields (e.g., Khushaba et al., 2013; Winslow et al., 2013), it appears to have never been utilized in research on L2 assessment to study learners' cognitive processes.

By leveraging the ability of EEG technology to measure and record L2 learners' brain signals, this study combined the use of perceptual data from retrospective interviews with EEG-derived neurophysiological data to conduct a more objective, direct, and non-invasive examination of L2 test-takers' cognitive processes during L2 assessment tasks. In particular, the study aimed to answer the following research questions:

1. What is the relationship between the EEG-derived measures of L2 learners' "engagement" and "interest" during L2 assessment tasks and L2 learners' perceptions of these cognitive states, as indicated by self-reported data from retrospective interviews?
2. What is the relationship between the EEG-derived measures of L2 learners' "engagement" and "interest" during L2 assessment tasks and their performance (i.e., scores) on these tasks?

3. To what extent do the participants' levels of "engagement" and "interest" affect their performance on these tasks, as indicated by the self-reported data from retrospective interviews?

Methodology

Conceptualized as a multiple-case study, this research project entailed the collection and analysis of three types of data: (a) performance data comprising participants' scores on L2 assessment tasks, (b) EEG-derived neurophysiological data comprising participants' cognitive states of "interest" and "engagement" while completing L2 assessment tasks, and (c) self-reported data from retrospective interviews comprising participants' perceptions of their "interest" and "engagement" during L2 assessment tasks.

Participants were 11 non-native speakers of English (eight female, three male) who were graduate students from various disciplines at the University of Hawai‘ i at Mānoa. Their age varied from 27 to 35 (M = 31, SD = 3.07). They were native speakers of Mandarin Chinese (n = 4), Indonesian (n = 2), Turkish (n = 2), Japanese (n = 1), Sinhalese (n = 1), and Nsei (n = 1), and had studied English between 10 and 28 years (M = 18, SD = 5.71).

All the data were gathered during individual 2-hour sessions in the Learner & User Xperience (LUX) Lab equipped with Emotiv EPOC+, which is a wireless EEG headset that utilizes 14 channels to record EEG data and is capable of recording six cognitive states, including interest and engagement. Each individual data collection session comprised the following steps: (a) providing instructions to the participant and setting up the EEG headset (15-20 minutes), (b) participant's completion of six computer-based L2 assessment tasks while wearing the EEG headset (40-50 minutes), and (c) a retrospective interview with the participant after each task (40-50 minutes total). Assessment tasks were taken from the Video-based Academic Listening Test (Suvorov, 2015). Each of the six tasks entailed watching a 2-3-minute video clip from an academic lecture and answering five multiple-choice questions about the lecture. Each retrospective interview was audio-recorded using Audacity, which is software for audio recording and editing.

Self-reported data from the interviews were transcribed, coded, and analyzed using NVivo Pro 11, which is software for qualitative data analysis. Data coding was done using a cyclical approach that included First Cycle and Second Cycle coding (Saldaña, 2009). To ensure the quality and consistency of coding, two coders coded ten percent of the same data, which is a common practice in the field (e.g., Brown, 2001; Chandler, 2003; Lee & Winke, 2013). Intercoder reliability was calculated in NVivo using Cohen's kappa coefficient, which was $\kappa = .98$. The high value of agreement provided evidence for the validity of the developed codes and the coding procedure.

Performance data (i.e., scores for the six assessment tasks) were analyzed quantitatively by tallying the number of correctly answered items for each participant. Finally, EEG-derived data recorded via Emotiv Pure•EEG™ software comprised the percentage of "engagement" and "interest" for each assessment task, representing the average value of each metric obtained over the duration of that specific task. It is noteworthy that due to technical issues with software 19 percent of EEG data were lost or corrupted and, therefore, had to be excluded from the analysis.

To answer Research Question 1, measures of "engagement" and "interest" derived from the EEG data were correlated with each participant's perceptions of how engaged and interested he/she was during each L2 assessment task by calculating Spearman's rank-order correlation coefficient. To answer Research Question 2, EEG-derived measures of "engagement" and "interest" were correlated with the participants' scores on each task by calculating Spearman's rank-order correlation coefficient. To answer Research

Question 3, the data from retrospective interviews were analyzed qualitatively to determine how L2 learners think their level of "engagement" and "interest" in L2 listening assessment affects their performance on those tasks.

Results

Research Question 1

Spearman's rank-order correlation coefficient for the EEG-derived measure of "engagement" and participants' perceptions of "engagement" was $\rho = .16$, with the two-tailed value of $p < .26$. Consequently, the association between the two variables was not statistically significant.

Spearman's rank-order correlation coefficient for the EEG-derived measure of "interest" and participants' perceptions of "interest" was $\rho = .29$, with the two-tailed value of $p < .04$. Consequently, the association between the two variables was statistically significant.

These results indicate that while there was no relationship between the EEG-derived measure of "engagement" and participants' perceptions of "engagement," we found a small, but statistically significant, positive relationship between the EEG-derived measure of "interest" and participants' perceptions of "interest." In other words, when participants reported higher levels of interest in an L2 listening assessment task, the measure of interest recorded by the EEG also tended to be higher for that specific task.

Research Question 2

Spearman's rank-order correlation coefficient for the EEG-derived measure of "engagement" and participants' scores on each task was $\rho = .14$, with the two-tailed value of $p < .32$. Consequently, the association between the two variables was not statistically significant.

Spearman's rank-order correlation coefficient for the EEG-derived measure of "interest" and participants' scores on each task was $\rho = -.01$, with the two-tailed value of $p < .96$. Consequently, the association between the two variables was not statistically significant.

These results reveal no relationship between the EEG-derived measures of "engagement" and "interest" during the completion of individual L2 listening assessment tasks and participants' scores on those tasks.

Research Question 3

The analysis of participants' responses during the interviews demonstrated that the level of engagement appeared to have a direct impact on the participants' scores in all but one case. Specifically, participants reported that the more engaged they were in the videos, the more confident they felt about the correctness of their responses and the easier it was for them to answer the questions. Lower levels of engagement, in their opinion, always made it difficult for them to answer the questions and, consequently, led to lower scores. The only exception was Participant 5 who claimed that his level of engagement did not have any impact on his responses.

Meanwhile, the participants' opinions about the impact of their interest level on the scores were somewhat more varied. Participants 5 and 8 posited that their level of interest had not impact on their scores whatsoever, whereas for three other participants no effect was found only for individual video-based listening assessment tasks (i.e., Task 3 for Participants 9 and 11 and Task 5 for Participant 6). For other participants the degree of their interest in each video-based L2 assessment task had a direct impact on their scores associated with that task. In other words, these participants tended to perform better on the tasks that they found to be more interesting than on the tasks that did not arouse much interest.

Discussion and Conclusion

The findings revealed a statistically significant relationship between the EEG-derived measure of "interest" and participants' perceptions of "interest" ($\rho = .29$, $p < .04$), which indicates that those participants whose brain activity during the video-based assessment tasks generated "interest" (as measured and recorded by EEG) also perceived those tasks as being interesting. No statistically significant relationship was found between the EEG-derived measure of "engagement" and participants' perceptions of "engagement," as well as between EEG-derived measures of "interest" and "engagement" during video-based L2 assessment tasks and participants' performance on those tasks. Furthermore, while the level of "engagement" in the video-based L2 listening assessment tasks appeared to have a direct impact on the performance of all but one participant, the effect that the participants' level of "interest" in the tasks had on their scores was less pronounced and varied among participants.

These findings suggest that while L2 test-takers' perceptions of "interest" appeared to be associated with "interest" as a neurophysiological signal in the brain measured and recorded by EEG, there was no statistically significant association between the EEG-derived measure of "engagement" and participants' perceptions of "engagement." In other words, how participants perceived their level of "engagement" in the tasks differed from how "engaged" their brains appeared to be as indicated by EEG.

When interpreting these findings, it is important to keep in mind the limitations related to the data that were gathered for this exploratory study. First, EEG measurements are easily affected by non-neural sources such as participants' eye blinks, muscle movements, and nearby electronic devices (Plöchl, Ossandón, & König, 2012), which can easily contaminate EEG data. In this study, the quality of EEG data has not been verified, which undermines the strength of the conclusions that can be drawn from the results of the EEG data analysis. Second, the EEG-derived measures of "engagement" and "interest" were calculated for each entire assessment task (i.e., a stimulus followed by five multiple-choice questions) and presented by Emotiv Pure•EEG™ software as percentages, the meaningfulness of which was limited. Finally, when analyzing the self-reported ratings of "engagement" and "interest," the researchers described differences in participants' interpretations of levels on a 10-point scale (for instance, Participant 2 considered level 6 to be "low," whereas for Participant 3 this level was "moderately high").

We therefore suggest that future studies utilize EEG raw data and move from a macro-level to a micro-level analysis focusing on individual components of an L2 assessment task (i.e., a stimulus and individual test items). Combining EEG and eye tracking in L2 assessment research is another methodological venue worthy of consideration.

CALL in Context

This presentation will address two main topics: language testing and tracking and logging. The two most relevant conference questions that will be discussed in this presentation are "How to determine the role and shape of the most appropriate technologies for our context?" and "How generalizable are the findings from experimental research in our context?" In the context of this exploratory study, we aimed at exploring the relationship between L2 learners' "interest" and "engagement" and their performance on L2 listening assessment tasks. Taking into consideration that self-reported data, which comprised participants' perceptions of their "interest" and "engagement" during the tasks, were subjective and prone to veridicality and reactivity risks (Bowles, 2010),

we utilized electroencephalography (EEG) in the context of this study to complement these indirect measures of "interest" and "engagement" with direct measures from neurophysiological data that were obtained via EEG. Furthermore, because of the diverse

international student population in the local context, we employed a multiple-case study design to explore individual L2 learners' cognitive states of "interest" and "engagement" in greater detail and to uncover aspects of the data that could be used to make generalizations to more global contexts.

Finally, the listening tasks in this study utilized video rather than audio-only prompts because they not only closely simulate the typical classroom learning experience, but also represent distance-learning situations such as MOOCs and telecollaborations that rely heavily on the video rather than the audio channel. The findings from this experimental research are generalizable to all contexts of academic listening, as well as the context of technology-mediated language learning in both local and distance learning scenarios.

References

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41–66.

Bowles, M. A. (2010). *The think-aloud controversy in second language research.* New York, NY: Routledge.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), 267–296.

Cohen, A. D. (2014). *Strategies in learning and using a second language* (2nd ed.). New York, NY: Routledge.

Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing, 24*(2), 209–250.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage Publications.

Kashkouli, Z., & Barati, H. (2013). Type of test-taking strategies and task-based reading assessment: A case in Iranian EFL learners. *Procedia – Social and Behavioral Sciences, 70,* 1580–1589.

Khushaba, R. N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B. E., & Townsend, C. (2013). Consumer neuroscience: Assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. *Expert Systems with Applications, 40,* 3803–3812.

Lee, H., & Winke, P. (2013). The differences among three--, four--, and five--option--item formats in the context of a high--stakes English--language listening test. *Language Testing, 30*(1), 99–123.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(1), 26–56.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*(4), 561–587.

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience, 6,* 1–23.

Saldaña, J. (2009). *The coding manual for qualitative researchers.* London, UK: Sage.

Srinivasan, N. (2007). Cognitive neuroscience of creativity: EEG based approaches. *Methods, 42*(1), 109–116.

Suvorov, R. (2015). Interacting with visuals in L2 listening tests: An eye-tracking study. In V. Berry (Ed.), *ARAGs research reports online (Report #AR-A/2015/1).* Retrieved from British Council http://www.britishcouncil.org/sites/default/files/interacting_with_visuals_in_l2_listening_tests_suvorov.pdf

Winslow, B., Carpenter, A., Flint, J., Wang, X., Tomasetti, D., Johnson, M., & Hale, K. (2013). Combining EEG and eye tracking: Using fixation-locked potentials in visual search. *Journal of Eye Movement Research, 6*(4), 1–11.