Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

A neurocognitive investigation of test methods and gender effects in listening assessment

Vahid Aryadoust[a], Li Ying Ng[a], Stacy Foo[a] and Gianluca Esposito[b,c,d]

[a]National Institute of Education, Nanyang Technological University, Singapore, Singapore;
[b]Psychology Program, School of Social Sciences, Nanyang Technological University, Singapore, Singapore;
[c]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore;
[d]Department of Psychology and Cognitive Science, University of Trento, Rovereto, Italy

FUNDING

This study was supported by a research grant from Paragon Testing Enterprises (Canada), which was administered by the National Institute of Education of Nanyang Technological University (NTU), Singapore. The NTU-Institutional Review Board (IRB) number of the study is IRB-2018-01-052. Authors would like to thank Dr Rohit Tyagi of Aerobe.com (Singapore) for providing technical support.

ORCID

Vahid Aryadoust http://orcid.org/0000-0001-6960-2489
Li Ying Ng http://orcid.org/0000-0002-5352-5180
Stacy Foo http://orcid.org/0000-0001-7377-2800
Gianluca Esposito http://orcid.org/0000-0002-9442-0254

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

Abstract

This is the first study to investigate the effects of test methods (while-listening performance and post-listening performance) and gender on measured listening ability and brain activation under test conditions. Functional near-infra- red spectroscopy (fNIRS) was used to examine three brain regions associated with listening comprehension: the inferior frontal gyrus and posterior middle temporal gyrus, which subserve bottom-up processing in comprehension, and the dorsomedial prefrontal cortex, which mediates top- down processing. A Rasch model reliability analysis showed that listeners were homogeneous in their listening ability. Additionally, there were no significant differences in test scores across test methods and genders. The fNIRS data, however, revealed significantly different activation of the investigated brain regions across test methods, genders, and listening abilities. Together, these findings indicated that the listening test was not sensitive to differences in the neurocognitive processes underlying listening comprehension under test conditions. The implications of these findings for assessing listening and suggestions for future research are discussed.

Keywords: Functional near-infrared spectroscopy; gender; listening test; test methods

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

Introduction

Listening comprehension involves decoding linguistic codes, retrieving literal meaning from memory (i.e., bottom-up processing), and incorporating one's own knowledge to recreate a mental representation of the aural message (i.e., top-down processing) (Kintsch, 1998; Rost, 2016). In general, the arrival of an aural message elicits a series of bottom-up processes, including phonological analysis, word recognition, semantic retrieval, and syntactic decoding that are subserved by the left inferior frontal gyrus (IFG; in particular Broca's area), superior temporal gyrus (STG), and posterior middle temporal gyrus (pMTG) (Friederici, 2011). Through top- down processing mediated by the dorsomedial prefrontal cortex (dmPFC), listeners subsequently make inferences about auditory inputs by integrating bottom-up information with their prior knowledge (Ferstl, Neumann, Bogler, & von Cramon, 2008; Perfetti & Frishkoff, 2008). It has been reported that low-ability listeners rely primarily on word- and sentence- level (i.e., bottom-up) processing, whereas high-ability listeners can execute discourse-level (i.e., top-down) processing more effectively (Rost, 2016).

Traditionally, listening ability is assessed using listening comprehension tests. Previous research has, however, identified several factors that can influence and confound test-takers' performances on such tests. Amongst the factors, test methods (Aryadoust, 2012, 2019; Field, 2012) and gender (Abbott, 2007; Aryadoust, 2012; Pae & Park, 2006; Harding, 2011) are the most under-researched. Listening test methods refer to the manner through which listening texts and test items are presented and comprise while-listening performance (WLP) and post-listening performance (PLP) tests (Aryadoust, 2019). A WLP test presents the listening text and test items simultaneously and thus demands concurrent listening, item reading, and answering (Aryadoust, 2019). In contrast, a PLP test first engages test- takers in listening (often allowing them to take notes), followed by reading and answering of test items (Aryadoust, 2019). Field (2012) noted that the need to multitask can undermine the ecological validity of a WLP test method, as test-takers often over-rely on test-specific strategies and shallow comprehension that differ from real-life listening. The findings from the verbal reports used by Field (2012) are supported by a recent eye-tracking study by Aryadoust (2019). Compared with reading test items at the start of a WLP

https://doi.org/10.1080/09588221.2020.1744667

test, Aryadoust (2019) reported that test-takers gazed at test items more frequently and for longer durations when simultaneously listening to the listening text and reading the test items. Importantly, the increases in gaze behaviors while multitasking reflected increased cognitive processing (Aryadoust, 2019). Kormos, Babuder, and Pizorn (2019) also reported that individual ability to simultaneously process written and spoken language could affect listening test performance. However, it remains unknown whether low-ability and high-ability listeners can be similarly differentiated at the behavioral (i.e., based on test-scores) or neural levels under WLP and PLP test methods.

Apart from test methods, gender can influence listening test performance. Previous research using item response theory (IRT), structural equation modeling, and verbal elicitation methods has reported gender differences in listening test scores (Abbott, 2007; Aryadoust, Goh, & Lee, 2011; Aryadoust, 2012; Harding, 2011; Pae & Park, 2006). Interestingly, several neuroimaging studies have suggested that gender performance differences in linguistic tasks may be associated with the neural substrates involved in bottom-up auditory processing (Burman, Bitan, & Booth, 2008; Kansaku, Yamaura, & Kitazawa, 2000; Phillips, Lowe, Lurito, Dzemidzic, & Mathews, 2001; Shaywitz et al., 1995; Zaidi, 2010). Burman et al. (2008) reported that activity of the IFG and pMTG was associated with auditory spelling and rhyming tasks for both boys and girls, and that girls showed larger hemodynamic responses in the STG during auditory tasks. Gender differences when processing verbal language have also been reported in adults, with women showing bilateral activation in the IFG, STG, and pMTG, and men showing leftward lateralization in the aforementioned areas (Kansaku et al., 2000; Phillips et al., 2001; Shaywitz et al., 1995). However, it is presently unknown if such gender differences extend to the dmPFC, which is involved in top-down auditory processing. Nevertheless, these findings suggest that females and males may utilize different cognitive processes to comprehend aural messages which, in turn, would contribute to differences in performance across both WLP and PLP tests.

The qualitative (i.e., verbal reports) and quantitative methods (i.e., IRT and eye-tracking) previously used for examining test method and gender effects are unable to provide deeper insights into the neurocognitive mechanisms involved in listening under test conditions (Kok & Jarodzka, 2017,

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

Norris, 1990, Mislevy, 2009). Furthermore, the previous neuroimaging studies that investigated brain activation patterns during linguistics tasks primarily used functional magnetic resonance imaging (fMRI) (Burman et al., 2008; Kansaku et al., 2000; Phillips et al., 2001; Shaywitz et al., 1995). During fMRI scanning, individuals' movements are restricted and the scanner generates substantial background noise that can confound experiments involving listening assessments (Gaab, Gabrieli, & Glover, 2007; Lei, Miyoshi, Niwa, Dan, & Sato, 2018).

Given the above-mentioned methodological constraints, functional near- infrared spectroscopy (fNIRS) may be more suitable for shedding light on the neurocognitive mechanisms that underlie test methods and gender effects under test conditions. FNIRS is a non-invasive and quiet optical brain imaging tool that has the potential to "reveal different patterns of cortical responses to those same stimuli, suggesting [sic] that stimuli are actually perceived or processed differently" (Sulpizio et al., 2018, p. 90). This technique contrasts with the behavioral measures discussed above which were based on the correlations of responses across different stimuli (Sulpizio et al., 2018; Watanabe, Yagishita, & Kikyo, 2008). Importantly, research has shown that fNIRS is capable of detecting changes in areas of the brain associated with auditory processing (e.g., the STG) when per- forming functional tasks in speech perception and comprehension where individuals provide correct and incorrect responses (Defenderfer, Kerr- German, Hedrick, & Buss, 2017; Lei et al., 2018). By extension, these works imply that it may be possible to use fNIRS to distinguish low-ability listeners (i.e., those with more incorrect responses) from high-ability listeners (i.e., those with more correct responses) during WLP and PLP tests.

Aims and Hypotheses

Considering the gaps in the literature, this study adopted the fNIRS imaging technique to examine the effects of test methods (i.e., WLP and PLP) and gender on listening ability and brain activation patterns in the dmPFC, IFG, and pMTG under test conditions. The hypotheses for this study were as follows:

Hypothesis One: Both WLP and PLP tests can differentiate low-ability from high- ability test-takers based on test performance scores and brain activation patterns. Based on previous

https://doi.org/10.1080/09588221.2020.1744667

neuroimaging research (Ferstl et al., 2008; Perfetti & Frishkoff, 2008), it was anticipated that high-ability test-takers would exhibit increased dmPFC activation (top-down processing) compared with low-ability test-takers for both the PLP and WLP tests. It was also hypothesized that higher ability test- takers would have lower IFG and pMTG activation (bottom-up processing) than low-ability test-takers during the PLP and WLP tests.

Hypothesis Two: There are significant differences in test performance scores and brain activation patterns between the WLP and PLP test methods. Based on previous neuroimaging studies (Ferstl et al., 2008; Perfetti & Frishkoff, 2008), it was hypothesized that the IFG, pMTG, and dmPFC would be engaged during listening under test conditions. Based on the works of Aryadoust (2019) and Field (2012), it was also hypothesized that the aforementioned neural substrates would be more activated during WLP than PLP tests.

Hypotheses Three: Males and females utilize distinct parts of their brains and score differently on the WLP and PLP listening tests. Based on previous research (Kansaku et al., 2000; Phillips et al., 2001; Shaywitz et al., 1995; Zaidi, 2010), it was hypothesized that females would score higher than males for both the WLP and PLP tests and show greater engagement of the dmPFC, IFG, and pMTG during the PLP test (i.e., when listening to the listening text) and WLP test (i.e., when simultaneously listening to the text and answering the test items).

Methods

Participants

Twenty-five university students without neurological, intelligent, or developmental atypicalities aged between 21 and 28 (22.9 ± 2.0years, females: n = 15, males: n = 10) were recruited for this study. All participants communicated in English either as their first (L1; n = 14) or second language (L2; n = 9) and originated from Singapore (n = 14), China (n = 6), Vietnam (n = 1), India (n = 1), Mauritius (n = 1), Germany (n = 1), and New Zealand (n = 1). This study was approved by the Institutional Review Board at

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667
a local university in Singapore and informed consent was obtained prior to the start of data collection. Upon completing the study, the participants were compensated SGD$10 for their time.

Edinburgh handedness inventory

Only right-handed participants were recruited for this study as previous research has indicated that the language-associated neural networks in right-handers are more activated in the left hemisphere (Kubota et al., 2008). An online version of Oldfield's (1971) Edinburgh Handedness Inventory (EHI) was used to determine participants' handedness. At the end of the questionnaire, a laterality index was automatically generated where a positive index indicates right-handedness and a negative index indicates left-handedness. All participants scored > 60 on the laterality index and were thus considered right-handed.

Listening tests

Two computer-mediated listening tests were used in this study. Each test comprised one long-listening monologue (i.e., lectures on Astronomy and Economics) and 11 comprehension questions. In this study, each test was used as assessment materials for both the WLP (i.e., WLP-Astronomy and WLP-Economics) and PLP (i.e., PLP-Astronomy and PLP-Economics) tests. Each listening text was first divided into 11 segments according to the information that was required for the respective comprehension questions. Both the WLP and PLP tests were presented on a 17-inch laptop (HP Pavilion, Hewlett Packard, CA, USA) using SuperLab 5 (Cedrus Corporation, CA, USA) stimulus presentation software. The WLP test comprised one block of 11 listening segments and the corresponding comprehension questions (Figure 1). The test instructions were displayed for 20 s. A 20 s rest period was given prior to the start of the WLP test and participants were required to fixate on a "þ" presented in the middle of the screen. For the test, all audio segments were presented concurrently with the corresponding test items (i.e., the questions and stems). Each test item was displayed for 5s longer than the audio segment to provide participants with sufficient time to answer the question. Each segment was followed by a 20 s rest period. The end of the test was marked by an "End of section" message presented for 5 s.

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

In contrast, the PLP test comprised two blocks that were presented sequentially, namely the PLP-Audio and PLP-Questions (Figure 1). Prior to the start of PLP-Audio, the instructions were displayed for 20 s

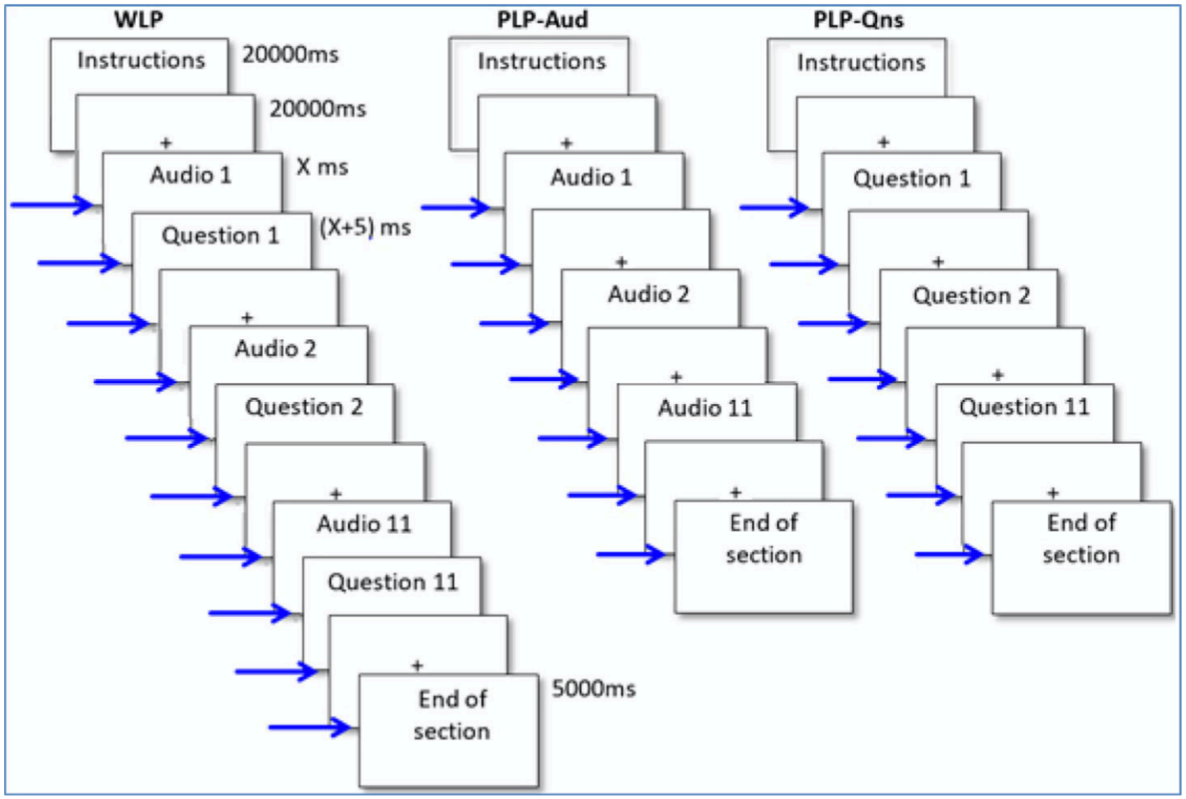

Figure 1. The listening comprehension experiment protocols were administered using SuperLab 5 software (Cedrus Corporation, CA, USA). The arrows represent the event markers sent through to NIRStar software (NIRx Medical Technologies LLC, MN, USA). Abbreviations: Aud = Audio; PLP = post-listening performance; Qns = Questions; WLP = while-listening performance.

followed by a 20 s rest period. For PLP-Audio, the 11 listening segments were presented in succession with a 20 s rest at the end of each segment. While the audio was playing, participants could take notes on a sheet of paper. However, participants were required to fixate on a "+" during the rest periods. Following the last 20 s rest period of the block, an "End of section" message was displayed for 5 s. Subsequently, the PLP-Questions began with a set of instructions and a rest period lasting for 10s and 20s, respectively. Here, participants were instructed to answer the comprehension questions displayed on the screen using their notes. Like the WLP test, the end of PLP-Questions was indicated by an "End of section" message presented for 5 s.

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

Data collection

All participants completed the abovementioned questionnaire (i.e., EHI) and listening tests (i.e., WLP and PLP) in single testing sessions lasting approximately 75minutes at a computer lab. Following the completion of the EHI, participants undertook the two listening tests as described above. In this study, a counterbalanced measures design was adopted to control for the effects of extraneous factors and test difficulty that could influence the results (Witmer, Aeschlimann, Metz, Troche, & Rammsayer, 2018). Thus, participants were divided into two groups according to the sequence in which they took the listening tests. Twelve participants were assigned to the first paradigm comprising the WLP- Astronomy test then the PLP-Economics test, while 13 participants were assigned to the second paradigm comprising the WLP-Economics test then the PLP-Astronomy test.

To measure the hemodynamic responses of the left IFG, dmPFC, and pMTG during the PLP and WLP tests, participants were fitted with a NIRS cap connected to a portable NIRS system (NIRSport device, NIRx Medical Technologies LLC, MN, USA). This system measured the hemodynamic responses at 7.81Hz and comprised eight pairs of sources and detectors arranged in accordance with the manufacturer's standard topographical montage to form 20 channels (Figure 2).

Prior to the start of the listening tests, an automatic calibration process was conducted using NIRStar 15-0 recording software (NIRx Medical Technologies LLC) to determine the optimum amplification factors for each of the 20 channels. To synchronize the data collected using SuperLab 5 software (Cedrus Corporation) and NIRStar 15-0 software (NIRx Medical Technologies LLC), an additional c-pod (Cedrus Corporation) was used to send event markers (i.e., in the form a square wave jerk) via USB through connections with the NIRSport device (NIRx Medical Technologies LLC). Each event marker was pre- set within SuperLab 5 software (Cedrus Corporation, CA, USA) to mark the start of each segment for the WLP and PLP tasks, as shown in Figure 1.

Data processing

The number-right method was used for scoring the test questions, with scores of 0, 1, and 2 being awarded for incorrect, partially correct, and correct responses, respectively.

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

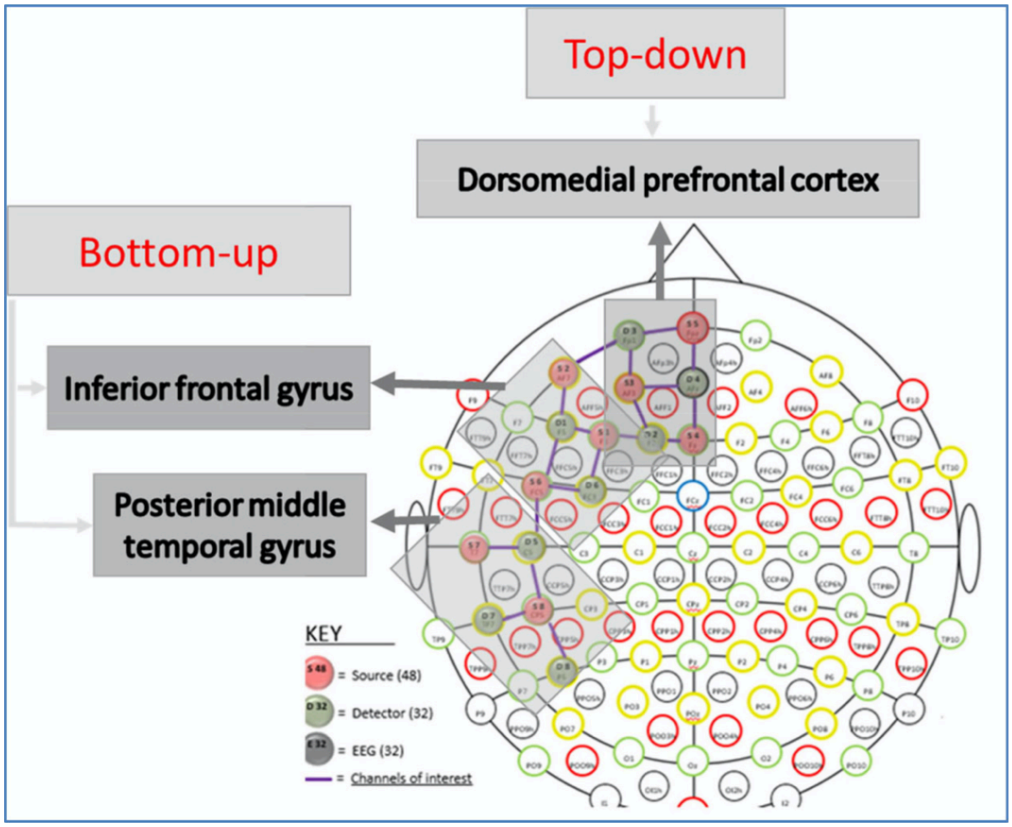

Figure 2. Montage of the functional near-infrared spectroscopy (fNIRS) device. This figure illustrates the locations of eight pairs of sources and detectors.

For brain activity measurements, raw data recorded using the NIRSport device (NIRx Medical Technologies LLC) were firstly exported from NIRStar 15-0 (NIRx Medical Technologies LLC) to NIRSLab v201706 (NIRx Medical Technologies LLC) for pre-processing. Pre-processing included the truncation of unreliable data, detecting possible sources of saturation (i.e., anomalies in the channels), removing discontinuities and spikes, and setting parameters for estimating hemodynamic states. Channels with significant background noise (i.e., gain > 8, coefficient of variation > 7.5) were visually inspected for data quality. Any channels with too many spikes and anomalies following the data pre-processing steps were removed from further analysis.

Two levels of general linear modeling (GLM) are available in NIRSLab v201706 software (NIRx Medical Technologies LLC): within-subject statistical parameter mapping (SPM) and between-subject SPM. To carry out within-subject SPM, the processed hemodynamic-state data file for each participant

https://doi.org/10.1080/09588221.2020.1744667

was uploaded in order to specify the parameters1 for GLM analysis. Using a robust-to-noise SPM-based algorithm (i.e., the restricted maximum likelihood estimation), the GLM coefficients were estimated. Finally, a non-thresholded within-subject SPM analysis was performed to generate the t-statistics of all channels as two-dimensional and three- dimensional color-coded images for easier visualization.

Neural evidence (fNIRS data)

Beta values representing blood oxygenated hemoglobin (Oxy-Hb) levels were first extracted from the within-subject SPM. Based on the brain region for each of the 20 channels, the data were averaged to create a mean beta value for each brain region of interest (i.e., dmPFC, IFG, and pMTG). The averaged beta values were subjected to further statistical analyses using IBM SPSS software, Version 25 (IBM Corporation, 2017).

For hypothesis one, the participants were split into two groups based on their test scores: (i) above average or (ii) below average. This separation was performed separately for the WLP and PLP tests, resulting in a total of four groups (i.e., low-WLP, high-WLP, low-PLP, and high-PLP). At each brain region (i.e., dmPFC, IFG, and pMTG), brain activation levels were analyzed separately across the WLP (i.e., low-WLP vs. high- WLP) and PLP tests (i.e., low-PLP vs. high-PLP).

Non-parametric statistical analysis was used to investigate the hypotheses. Mann-Whitney U tests were conducted to test for different levels of listening ability (i.e., hypothesis one) and differences in activation levels in the dmPFC, IFG, and pMTG across genders (i.e., hypothesis three). In addition, a Wilcoxon rank-sum test was performed to assess whether there were differences in activation levels in the dmPFC, IFG, and pMTG across the WLP and PLP tests (i.e., hypotheses two). All non-parametric statistics were performed using IBM SPSS software, Version 24 (IBM Corporation, 2017).

Behavioral evidence (test scores)

Psychometric quality of the test

To investigate the psychometric properties of the listening test items, two rounds of Rasch-Andrich rating scale model (RSM) analysis were performed on the WLP and PLP tests (Andrich, 1978). The RSM analysis was used to verify whether the test items functioned properly and were not confounded

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

by sources of construct-irrelevant variance. The RSM is robust for small samples (Linacre, 2018a) and is therefore well-suited for the analysis in this study. Item difficulty, fit statistics, and reliability coefficients were computed using Winsteps, Version 4.4 (Linacre, 2018b). Item difficulty measures the cognitive demand of test items and is expressed in log-odd-units (logits) (Fan & Bond, 2019). Fit statistics are diagnostic measures for investigating whether there are perturbations or anomalies in test data (Fan & Bond, 2019). Two fit measures were generated: the infit mean square (MnSq) and outfit MnSq. The infit MnSq is inlier-sensitive and thus detects anomalies when item difficulty and person ability measures are close, whereas the outfit MnSq is outlier-sensitive and is more useful when item difficulty and person ability measures are far from each other (Linacre, 2018a). While the expected value of both the infit and outfit MnSq is 1, any value between 0.5 and 1.5 reflects acceptable psychometric properties (Linacre, 2018a). Finally, reliability and separation coefficients were computed for persons and items to determine whether items and persons were divisible into separate levels of difficulty and ability, respectively.

Testing the hypotheses

Each hypothesis was further evaluated using the test scores. The RSM computed Rasch-based reliability and separation coefficients for test takers were used to examine the first hypothesis. Reliability and separation coefficients equal to or greater than .80 and 2.00 per WLP and PLP test, respectively, would indicate that the tests discriminated between at least two levels of listening ability.

To test the second hypothesis, a Wilcoxon rank-sum test was per- formed to examine differences between WLP and PLP test scores. To test the third hypothesis, Mann-Whitney U tests were conducted to investigate differences in test scores across genders (i.e., males vs. females). Bonferroni correction was applied for multiple comparisons.

Results

Testing the hypotheses: Neural evidence (fNIRS data)

Hypothesis one (Figure 3)

The omnibus Mann-Whitney U test indicated that the overall Oxy-Hb levels across the three brain regions for high-WLP test-takers (β values: $(3.50 \pm 11.03) * 10^{-5}$) were higher than those for low-

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

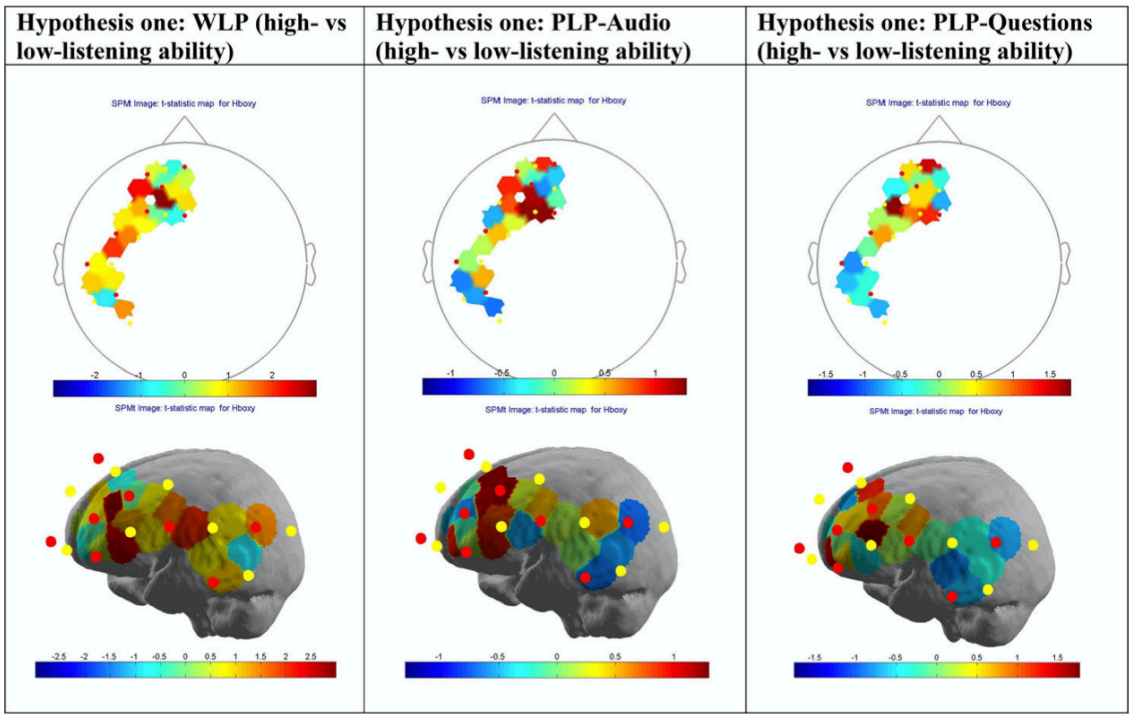

Figure 3. Results of functional near-infrared spectroscopy (fNIRS) for hypothesis one. Positive values (yellow to red) indicate greater activation in individuals with higher listening comprehension ability than individuals with lower listening comprehension ability; negative values (light blue to dark blue) indicate lower activation in individuals with higher listening comprehension ability than individuals with lower listening comprehension ability. Green indicates equal activation across both groups. Abbreviations: WLP = while-listening performance; PLP = post-listening performance.

WLP test-takers (β values: (-1.22±10.13) * 10^{-5}), U = 4016.00, $p = 0.013$. However, the overall Oxy-Hb level was not different between high-PLP test-takers (β values: (0.86 ± 84.77) * 10^{-5}) and low-PLP test-takers (β values: (13.66±14.44) * 10^{-5}) during PLP-Audio, U = 4393.00, $p = 0.169$. In contrast, the overall Oxy-Hb level for high-PLP test-takers (β values: (-1.47±40.86) * 10^{-5}) was lower than that for low-PLP test-takers (β values: (5.45 ± 13.42) * 10^{-5}) during PLP-Questions, U = 4079.00, $p = 0.033$. Further post-hoc tests did not reveal any significant differences in Oxy-Hb levels between high-ability and low-ability test-takers across the dmPFC, IFG, and pMTG for WLP and PLP-Questions ($p > 0.05$) (refer to Table 1 for *p*-values).

Table 1. *Hypotheses of the Study.*

Hypothesis	Results of statistical analysis based on test scores (behavioral level)	Brain activity comparisons	Results of statistical analysis based on neural activation (neural level)		Conclusion
Hypothesis one: Both WLP and PLP tests can differentiate low-ability from high-ability test-takers based on test performance scores and brain activation patterns.	*WLP*: Person reliability: 0.50, 1 stratum Item reliability: 0.83, 2 strata *PLP*: Person reliability: 0.70, 1 stratum Item reliability: 0.77, 1.84 strata	High-WLP vs. Low-WLP (Mann-Whitney U tests)	Omnibus: $p = 0.013$ (High-WLP > Low-WLP) dmPFC: $p = 0.385$ (Not significant) IFG: $p = 0.380$ (Not significant) pMTG: $p = 0.208$ (Not significant)	1.	Test scores did not support the hypothesis. Participants were homogenous in their listening ability despite the heterogeneous item difficulty.
		High-PLP vs. Low-PLP (PLP-Audio; Mann-Whitney U test)	Omnibus: $p = 0.169$ (Not significant)	2.	Neural activation partially supported the hypothesis. The results suggested that high ability performers had overall decreased activity, in line with the hypothesis; however, the regional differences were not statistically significant.
		High PLP vs. Low-PLP (PLP-Questions; Mann-Whitney U tests)	Omnibus: $p = 0.033$ (High-PLP < Low-PLP) dmPFC: $p = 0.469$ (Not significant) IFG: $p = 0.668$ (Not significant) pMTG: $p = 0.801$ (Not significant)		
Hypothesis two: There are significant differences in test performance scores and brain activation patterns between WLP and PLP test methods.	WLP vs. PLP scores: N.S.	WLP vs. PLP-Audio (Wilcoxon signed rank tests)	Omnibus: $p < 0.001$ (WLP < PLP-Audio) dmPFC: $p = 0.002$ (WLP < PLP-Audio) IFG: $p = 0.002$ (WLP < PLP-Audio) pMTG: $p = 0.100$ (Not significant)	1.	Test scores did not support the hypothesis.
				2.	Neural activation partially supported the hypothesis. The results suggested that the PLP-Questions resulted in significantly higher brain activation than the WLP.
		WLP vs. PLP-Questions (Wilcoxon signed rank tests)	Omnibus: $p = 0.039$ (WLP > PLP-Questions) dmPFC: $p = 0.050$ (Not significant) IFG: $p = 0.092$ (Not significant) pMTG: $p = 0.584$ (Not significant)		
Hypothesis three: Males and females utilize distinct parts of their brains and score differently on WLP and PLP listening tests.	WLP; Females vs. Males: N.S.	WLP: Females vs. Male (Mann-Whitney U tests)	Omnibus: $p < 0.001$ (Females < Males) dmPFC: $p = 0.023$ (Not significant) IFG: $p = 0.040$ (Not significant) pMTG: $p = 0.038$ (Not significant)	1.	Test scores did not show any differences between genders.
	PLP; Females vs. Males: N.S.	PLP-Audio: Females vs. Males (Mann-Whitney U tests)	Omnibus: $p < 0.001$ (Females < Males) dmPFC: Females < Males, $p = 0.012$ pMTG: Females < Males, $p = 0.005$ IFG: $p = 0.861$ (Not significant)	2.	The neural activation results suggested that females had significantly lower activation relative to males across the PLP and WLP.
		PLP-Questions: Females vs. Males (Mann-Whitney U tests)	Omnibus: $p < 0.0001$ (Females < Males) dmPFC: $p = 0.118$ (Not significant) IFG: $p = 0.026$ (Not significant) pMTG: $p = 0.044$ (Not significant)		

Note: WLP = While-listening performance; PLP = Post-listening performance; dmPFC = dorsomedial prefrontal cortex; IFG = inferior frontal gyrus; pMTG = posterior middle temporal gyrus.

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

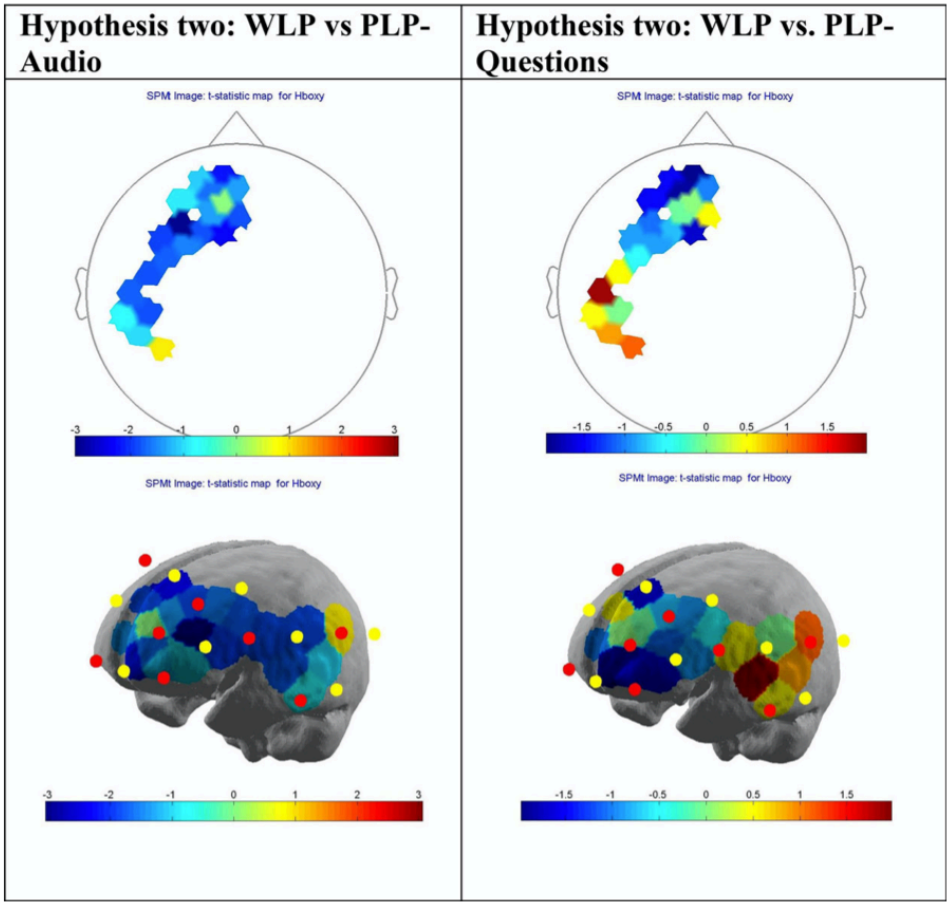

Figure 4. Results of functional near-infrared spectroscopy (fNIRS) for hypothesis two. Positive values (yellow to red) indicate greater activation in the first condition than the second condition; negative values (light blue to dark blue) indicate lower activation in the first condition than in the second condition. Green indicates equal activation across both groups. Abbreviations: WLP = while-listening performance; PLP = post-listening performance.

Hypothesis two (Figure 4).

The omnibus Wilcoxon Signed Ranks test showed that the overall Oxy- Hb level across the three brain regions for the WLP test (β values: (1.63±10.86) * 10^{-5}) was significantly lower compared with PLP-Audio (β values: (5.57±67.96) * 10^{-5}), $Z = -7.26$, $p < 0.001$, but higher compared with PLP-Questions (β values: (1.07 ± 33.50) * 10^{-5}), $Z = -2.07$, $p = 0.039$.

For WLP vs. PLP-Audio, post-hoc tests with Bonferroni corrections ($\alpha = 0.01667$) indicated that the Oxy-Hb level measured during WLP was lower than that of PLP-Audio at the dmPFC (β values: for WLP: (0.21 ± 7.63) * 10^{-5}; PLP-Audio: (12.23 ± 14.42) * 10^{-5}, $Z = -3.11$, $p = 0.002$) and IFG ((β values

https://doi.org/10.1080/09588221.2020.1744667

for WLP: (1.51 ± 10.64) * 10^{-5}, PLP-Audio: (14.14 ± 19.37) * 10^{-5}, Z = -3.09, p = 0.002). No difference in Oxy-Hb level was, however, observed at the pMTG between the WLP (β values: (3.04 ± 13.55) * 10^{-5}) and PLP-Audio (β values: (-9.46 ± 114.56) * 10^{-5}), Z = -1.64, p = 0.100.

For WLP vs. PLP-Questions, post-hoc tests with Bonferroni corrections (α=0.01667) indicated that the Oxy-Hb levels at the individual brain regions were not statistically different. The statistics are as follows: dmPFC (β values for WLP: (0.21 ± 7.63) * 10^{-5}, PLP-Questions: (4.35 ± 6.33) * 10^{-5}, Z = -1.96, p = 0.092); and pMTG (β values for WLP: (3.04 ± 13.55) * 10^{-5}, PLP-Questions: (-6.60 ± 56.26) * 10^{-5}, Z = -0.547, p = 0.584).

Hypothesis three (Figure 5)

The omnibus Mann-Whitney U tests indicated that females had lower Oxy-Hb levels than males for the WLP test (β values for females: (-2.18 ± 9.41) * 10^{-5} and males: (7.06 ± 10.60) * 10^{-5}, U = 3304.00, p < 0.001), PLP-Audio (β values for females: (7.11 ± 85.53) * 10^{-5}, males: (23.69 ± 17.71) * 10^{-5}, U = 1809.00, p < 0.001), and PLP-Questions (β values for females: (-2.52 ± 42.23) * 10^{-5}, males: (6.21 ± 12.78) * 10^{-5}, U = 3669.00, p < 0.001).

Further post-hoc tests with Bonferroni corrections (a1⁄40.01667) indicated that females had significantly lower Oxy-Hb levels than males for PLP-Audio at the dmPFC (β values for females: (5.21 ± 13.14) * 10^{-5}, males: (21.58 ± 10.47) * 10^{-5}), U = 19.00, p = 0.012) and pMTG (β values for females (3.28 ± 143.09) * 10^{-5}, males: (26.89 ± 19.50) * 10^{-5}, U = 18.00, p = 0.005. There was, however, no difference in Oxy-Hb level between females (β values: (8.03±15.12) * 10^{-5}) and males for PLP-Audio at the IFG (β values: (22.69 ± 22.13) * 10^{-5}), U = 32.00, p = 0.026.

Testing the hypotheses: Behavioral evidence (test scores)

Hypothesis one

The RSM was used to compute the person and item reliability and separation coefficients of the WLP and PLP tests. The person and item reliability coefficients of the WLP test were 0.50 (1 stratum)

https://doi.org/10.1080/09588221.2020.1744667

and 0.83 (2 strata), respectively. These values suggest that the participants were homogenous in terms of

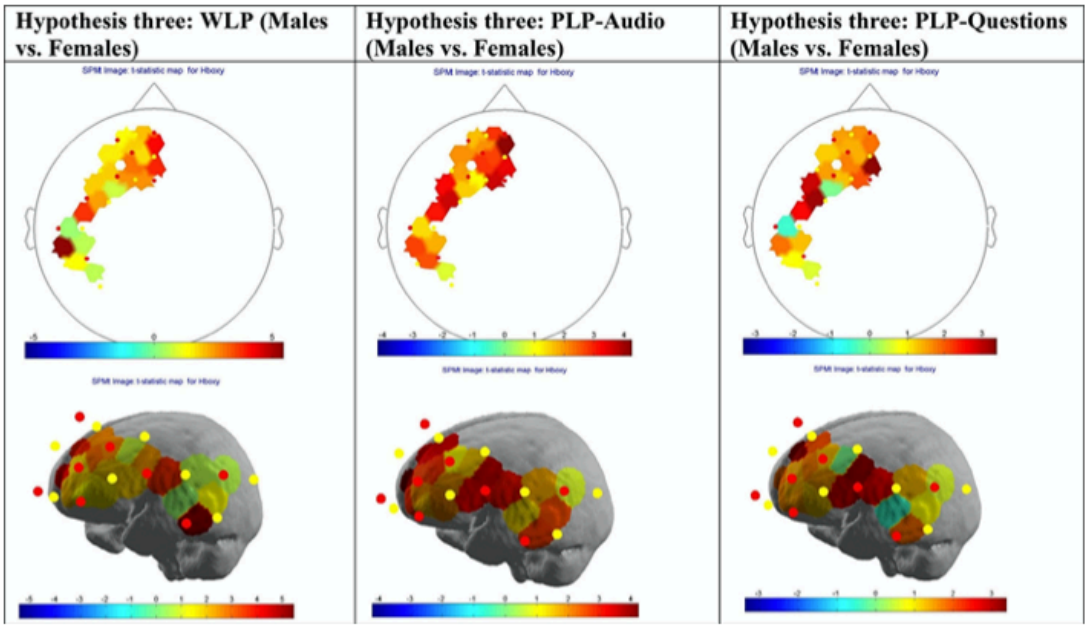

Figure 5. Results of functional near-infrared spectroscopy (fNIRS) for hypothesis three. Positive values (yellow to red) indicate greater activation in males than females; negative values (light blue to dark blue) indicate lower activation in males than females. Green indicates equal activation across both groups. Abbreviations: WLP = while-listening performance; PLP = post-listening performance.

their listening ability but that the test items were heterogeneous in terms of their difficulty level. The person and item reliability coefficients of the PLP test were 0.70 (1 stratum) and 0.77 (1.84 strata), respectively. This finding is indicative of a similar psychometric quality for the WLP and PLP tests, thus rejecting hypothesis one. (In addition, the fit statistics of all test items fell between 0.5 and 1.5, except for item 1 in the PLP-Astronomy (i.e., infit and outfit MnSq values of 1.77 and 1.69, respectively)). Further inspection of this item revealed that this mild deviation from the acceptable fit range was due to three incorrect answers provided by high-PLP participants.)

Hypotheses two

The Wilcoxon rank-sum test indicated no significant differences between the test scores obtained for the WLP (11.92 ± 4.55) and PLP (15.56 ± 4.32) tests, $Z = -1.06$, $p = 0.29$.

Hypothesis three

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

The Mann-Whitney U test indicated no significant differences in test scores for the WLP test between males (13.50 ± 4.15) and females (10.87 ± 4.54), $U = 77.50$, $p = 0.24$. Similarly, there were no significant differences in test scores for the PLP between males (16.50 ± 3.41) and females (14.93 ± 4.76), $U = 65.50$, $p = 0.09$.

Summary

Table 1 provides a summary of the findings of this study. It also presents the hypotheses, comparisons, and statistical analyses carried out.

Discussion

The aims of this study were to examine (i) whether the WLP and PLP tests could similarly differentiate between low- and high-ability test-takers at the behavioral and neural levels, and (ii) the effects of test methods and gender on brain activation and listening performance during the WLP and PLP tests. At the behavioral level, the three hypotheses were not supported. Based on the Rasch reliability of the test scores obtained, neither the WLP nor PLP test could differentiate low-ability from high- ability test-takers. Furthermore, there were no differences between test methods (i.e., WLP and PLP) and genders according to the test scores. At the neural level, hypotheses one and two were partially supported, while hypothesis three was refuted. Firstly, there were significant differences in the overall hemodynamic responses between low- and high-ability test-takers during the WLP test. No significant differences in hemodynamic responses were, however, observed between high- and low-ability test-takers across specific brain regions (i.e., the dmPFC, IFG, and pMTG). Secondly, test-takers' hemodynamic responses in the dmPFC and IFG during the WLP test were significantly lower than those during the PLP-Audio. Lastly, females exhibited lower hemodynamic responses than males in the dmPFC and pMTG during PLP-Audio. These findings are further discussed below.

Hypothesis one

In relation to the first hypothesis, the differences in overall hemodynamic responses suggest that the left dmPFC, IFG, and pMTG were activated in unison to support listening comprehension during the WLP test. This finding also indicates that none of the aforementioned cortical regions was solely

https://doi.org/10.1080/09588221.2020.1744667

responsible for the differences in test performance between high-ability and low-ability test-takers for the WLP. The differences in WLP test performance may perhaps reflect a general lack of ability to effectively engage the top-down and bottom-up language processing networks that involve the dmPFC, IFG, and pMTG.

In contrast, the absence of overall or specific cortical activation pat- terns during the PLP-Audio is indicative that high-ability and low-ability test-takers engaged similar neural processes when listening to the text. Additionally, high-ability test-takers had lower overall hemodynamic responses than low-ability test-takers across the left dmPFC, IFG, and pMTG during PLP-Questions. This finding suggests that high-ability test-takers used less cognitive effort than low-ability test-takers when responding to questions. This finding contrasts previous research where correct answers corresponded with higher brain activation levels than incorrect answers (Defenderfer et al., 2017; Lei et al., 2018). Alternatively, other studies have shown that less proficient listeners require more extensive brain activation when listening for information in their second language (Hasegawa, Carpenter, & Just, 2002; Lei et al., 2018; Nakai et al., 1999). This finding is further supported by Cannizzaro, Stephens, Breidenstein, and Crovo (2016) who reported that speech that is easier to process is associated with lower brain activation. High-ability test-takers may have found the questions easier to process than low-ability test-takers, thus exhibiting lower brain activity across the left dmPFC, IFG, and pMTG.

The implication of this finding is that low-ability test takers (below average) likely experience a higher cognitive load during listening than high-ability test takers (above average). This assumption raises the possibility that the test content perhaps included input that was unfamiliar or difficult for the low-ability test takers to process due to higher linguistic challenges. This finding is in line with previous studies showing that the demand for processing multiple stimuli taxes the working memory of comprehenders (Chai & Erlam, 2008). Future research is needed to investigate the connection between brain activation, the storage capacity of working memory, and cognitive load in WLP and PLP tests.

Hypothesis two

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

The second hypothesis was partially supported. While higher brain activation was observed during WLP in comparison to PLP-Questions, there was lower activity in WLP in comparison to PLP-Audio. This finding suggests that differences in cognitive demand between the WLP and PLP tests are not one-directional. It is possible that the multitasking nature of WLP (listening and processing new information, reading and processing the question, and answering) was more cognitively demanding overall than the PLP-Questions (reading and processing the question, recalling information, and answering) (Aryadoust, 2019), resulting in the lower overall brain activation observed during the PLP-Questions. On the other hand, listening and note-taking involve more top-down (i.e., involving the dmPFC) and bottom-up (i.e., involving the IFG) cognitive processing, which could account for the higher activation during PLP-Audio.

As previously discussed, previous research using fNIRS has found a possible link between workload and brain activation, where more com- plex sentences elicited greater activation (Lei et al., 2018). In addition, when speech was masked with sounds that increased the workload to perceive the speech, there was also increased brain activation (Defenderfer et al., 2017). Accordingly, the increased brain activation for WLP could be due to the increased cognitive load in engaging both visual and auditory language pathways during WLP compared to PLP- Questions, which only engages the visual language pathway.

For PLP-Audio, previous research has suggested that the left dmPFC is implicated in prospective memory that aids future intentions and planning (Umeda, Kurosaki, Terasawa, Kato, & Miyahara, 2011). Aside from listening for potential cues at the word level (i.e., bottom-up processing; Field, 2012), test-takers must simultaneously (i) remember the listening text content, (ii) predict what information might be relevant, and (iii) take notes in order to answer the comprehension questions during PLP- Questions (i.e., top-down processing). While this finding resonates with Field (2012), who suggested that the WLP engages shallow listening com- pared with the PLP, it contrasts the gaze behavioral findings of Aryadoust (2019). Aryadoust suggested that, because the WLP demands multitasking involving reading, listening, and answering test items, it involves more cognitive processing. However, it should be noted that the

https://doi.org/10.1080/09588221.2020.1744667

work by Aryadoust (2019) measured gaze behaviors during WLP, whereas the present study compared neurocognitive patterns during both WLP and PLP.

Hypothesis three

Lastly, the third hypothesis of this study was refuted as females had lower cortical activity in the dmPFC and pMTG during PLP-Audio. Although this finding adds to the evidence of gender differences in language processing networks (Kansaku et al., 2000; Phillips et al., 2001; Shaywitz et al., 1995; Zaidi, 2010), it partially contradicts the finding of Burman et al. (2008), who reported that girls had higher hemodynamic responses in the STG than boys when performing auditory tasks. This disparity in hemodynamic responses is likely due to differences in task dynamicity as the participants in the Burman et al. (2008) study had to perform an auditory task while lying still in an fMRI scanner. As previously mentioned, fMRI scanners generate substantial background noise that can influence the quality of auditory stimuli (Gaab et al., 2007; Lei et al., 2018).

It is important to note that cortical activation of the right hemisphere was not recorded in this study. Therefore, further investigation is needed to determine whether the gender differences in brain laterality reported in previous studies (Kansaku et al., 2000; Phillips et al., 2001; Shaywitz et al., 1995; Zaidi, 2010) would be observed during the WLP and PLP tests.

Limitations

Considering the sample sizes (and statistical analyses) used in previous research (e.g., Defenderfer et al., 2017; Lei et al., 2018), the sample size in this study is considered small. This may limit the generalizability of our results, especially the impact of test methods on listening ability, as average scores were used to segregate high-ability from low-ability test- takers. Additionally, the sample was heterogeneous in terms of language background, as not all participants communicated in English as L1. While a larger and more homogenous sample is imperative for improv- ing the generalizability of the results, further investigations may also focus on expanding the sample size to examine the impact of language background (e.g., English as L1 vs. L2) on the hemodynamic responses and test scores for both WLP and PLP.

Article Published on: Computer Assisted Language Learning, 2020;1-21.

https://doi.org/10.1080/09588221.2020.1744667

Conclusion

In summary, the findings of this study indicate that the WLP and PLP listening tests were not equally sensitive to differences in the neurocognitive processes underlying listening comprehension under test conditions. Of note, this is the first study to investigate the neural substrates involved in listening comprehension under test conditions. The findings of this study are encouraging and may lead to new possibilities for research in listening assessment.

Firstly, the involvement of different neural substrates across test methods has important implications for test design, validation, and assessment purposes. For a listening test to possess ecological validity, it needs to engage the neurocognitive processes that underlie real-life listening. This is a major gap in knowledge that needs to be investigated in future research.

Additionally, the differences in neural substrates across genders reported in this study may have implications for test design, specifically from a test fairness perspective. One may examine whether differences in language processing can affect test fairness for males and females. Previous research has applied IRT or other latent variable models to examine the role of gender in fairness in computerized and traditional language assessments (Abbott, 2007; Aryadoust et al., 2011; Aryadoust, 2012). Differential test functioning (DTF) analysis is a useful method for examining whether test items function differentially for different groups of test-takers. Based on our findings, we propose the concept of differential brain functioning (DBF) in the present study to refer to the differences in brain activation in different groups of test-takers that result in differences in test performance. Future research may compare brain activation across larger and different groups of test-takers, including (i) L1 vs. L2, (ii) males vs. females, and (iii) high-ability vs. low-ability, to determine whether DTF is related to DBF, and if so, what implications this would have for test fairness and test validity.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(1), 7–36. doi:10.1177/ 0265532207071510

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. doi:10.1007/BF02293814

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening, 26*(1), 40–60. doi:10.1080/10904018.2012. 639649

Aryadoust, V. (2019). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: An eye-tracking study. *Computer Assisted Language Learning,* 1–28. doi:10.1080/09588221.2019.1574267

Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361–385. doi:10.1080/15434303.2011.628632

Burman, D.D., Bitan, T., & Booth, J.R. (2008). Sex differences in neural processing of language among children. *Neuropsychologia, 46*(5), 1349–1362. doi:10.1016/j.neuropsychologia.2007.12.021

Cannizzaro, M. S., Stephens, S. R., Breidenstein, M., & Crovo, C. (2016). Prefrontal Cortical Activity During Discourse Processing: An Observational fNIRS Study. *Topics in Language Disorders, 36*(1), 65–79. doi:10.1097/TLD.0000000000000082

Chai, J., & Erlam, R. (2008). The effect and the influence of the use of video and captions on second language learning. *New Zealand Studies in Applied Linguistics, 14*(2), 25–44.

Defenderfer, J., Kerr-German, A., Hedrick, M., & Buss, A. T. (2017). Investigating the role of temporal lobe activation in speech perception accuracy with normal hearing adults: An event-related fNIRS study. *Neuropsychologia, 106*, 31–41. doi:10.1016/j.neuropsychologia.2017.09.004

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language*

assessment volume I: Fundamental techniques (pp. 83–102). New York, NY: Routledge. doi: 10.4324/9781315187815-5

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping, 29*(5), 581–593. doi:10.1002/hbm.20422

Field, J. (2012). The cognitive validity of the lecture listening section of the IELTS listening paper. In L. Taylor & C. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment.* Cambridge: Cambridge University Press.

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews, 91*(4), 1357–1392. doi:10.1152/physrev.00006.2011

Gaab, N., Gabrieli, J. D. E., & Glover, G.H. (2007). Assessing the influence of scanner background noise on auditory processing. I. An fMRI study comparing three experimental designs with varying degrees of scanner noise. *Human Brain Mapping, 28*(8), 703–720. doi:10.1002/hbm.20298

Harding, L. (2011). Accent and listening assessment: *A validation study of the use of speakers with L2 accents on an academic English listening test. (Language Testing and Evaluation).* Frankfurt: Peter Lang.

Hasegawa, M., Carpenter, P. A., & Just, M. A. (2002). An fMRI Study of Bilingual Sentence Comprehension and Workload*. NeuroImage, 15*(3), 647–660. doi:10.1006/ nimg.2001.1001

IBM Corporation (2017). *IBM SPSS statistics for Windows, Version 25.0 [Computer software].* Armonk, NY: IBM Corporation.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York, NY: Cambridge University Press.

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education, 51*(1), 114–122. doi:10.1111/ medu.13066

Kormos, J., Babuder, M. K., & Pizorn, K. (2019). The Role of Low-level First Language Skills in Second Language Reading, Reading-While-Listening and Listening Performance: A Study of Young

Dyslexic and Non-dyslexic Language Learners. *Applied Linguistics, 40*(5), 834–858. doi:10.1093/applin/amy028

Kubota, M., Inouchi, M., Dan, I., Tsuzuki, D., Ishikawa, A., & Scovel, T. (2008). Fast (100–175ms) components elicited bilaterally by language production as measured by three-wavelength optical imaging. *Brain Research, 1226*, 124–133. doi:10.1016/j. brainres.2008.05.079

Nakai, T., Matsuo, K., Kato, C., Matsuzawa, M., Okada, T., Glover, G. H., ... Inui, T. (1999). A functional magnetic resonance imaging study of listening comprehension of languages in human at 3 tesla-comprehension level and activation of the language areas. *Neuroscience Letters, 263*(1), 33–36. doi:10.1016/S0304-3940(99)00103-2

Kansaku, K., Yamaura, A., & Kitazawa, S. (2000). Sex differences in lateralization revealed in the posterior language areas. *Cerebral Cortex, 10*(9), 866–872. doi:10.1093/cercor/10.9.866

Lei, M., Miyoshi, T., Niwa, Y., Dan, I., & Sato, H. (2018). Comprehension-Dependent Cortical Activation During Speech Comprehension Tasks with Multiple Languages: Functional Near-Infrared Spectroscopy Study. *Japanese Psychological Research, 60*(4), 300–310. doi:10.1111/jpr.12218

Linacre, J. M. (2018b). *Winsteps [Computer program].* Chicago, IL: Winsteps.com.

Linacre, J. M. (2018a). *A user's guide to WINSTEPS.* Chicago, IL: Winsteps.com.

Mislevy, R.J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications.* Charlotte, NC: Information Age Publishing.

Pae, T.-I., & Park, G.-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496. doi:10.1191/0265532206lt338oa

Perfetti, C.M., & Frishkoff, G.A. (2008). Neural Bases of Text and Discourse Processing. In B. Stemmer, and H.A. Whitaker (Eds.), *Handbook of Neuroscience of Language* (pp. 165–174Cambridge, MA: Elsevier.

Phillips, M. D., Lowe, M. J., Lurito, J. T., Dzemidzic, M., & Mathews, V. P. (2001). Temporal Lobe Activation Demonstrates Sex-based Differences during Passive Listening. *Radiology, 220*(1), 202–207. doi:10.1148/radiology.220.1.r01jl34202

Rost, M. (2016). Teaching and researching listening (3rd Ed.). London: Longman. Shaywitz, B. A., Shaywltz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., ... Gore, J. C. (1995). Sex differences in the functional organization of the brain for language. *Nature, 373*(6515), 607–609. doi:10.1038/373607a0

Sulpizio, S., Doi, H., Bornstein, M. H., Cui, J., Esposito, G., & Shinohara, K. (2018). fNIRS reveals enhanced brain activation to female (versus male) infant directed speech (relative to adult directed speech) in Young Human Infants. *Infant Behavior and Development, 52*, 89–96. doi:10.1016/j.infbeh.2018.05.009

Umeda, S., Kurosaki, Y., Terasawa, Y., Kato, M., & Miyahara, Y. (2011). Deficits in prospective memory following damage to the prefrontal cortex. *Neuropsychologia, 49*(8), 2178–2184. doi:10.1016/j.neuropsychologia.2011.03.036

Watanabe, T., Yagishita, S., & Kikyo, H. (2008). Memory of music: Roles of right hippocampus and left inferior frontal gyrus. *NeuroImage, 39*(1), 483–491. doi:10.1016/j.neuroimage.2007.08.024

Zaidi, Z. (2010). Gender Differences in Human Brain: A Review. *The Open Anatomy Journal, 2* 37–55. doi:10.2174/18776094010020100,