RESEARCH Free and Open Access

An eye-tracking investigation of visual search strategies and test performance of L1 and L2 listening test takers

Vahid Aryadoust * and Stacy W. L. Foo

*Correspondence:
vahid.aryadoust@nie.edu.sg
National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616, Singapore
Full list of author information is available at the end of the article

Abstract

Through the use of eye-tracking technology and a while-listening performance (WLP) test, this study examined the differences in gaze behaviors and measured listening performances on test items (across various stages of the test) and compared them between native English-speaking (E-L1) and non-native English-speaking (E-L2) candidates. One hundred students from a public university in Singapore participated in the study. A series of Mann-Whitney U tests indicated that E-L1 candidates outperformed E-L2 candidates in the test with higher test scores. Using stringent data processing cutoffs (presence≥80% gaze data) and a series of non-parametric multivariate analyses, the study further found that the dynamicity of gaze behaviors on the test items across various stages of the test was similar between E-L1 and E-L2 candidates. However, there were distinctive differences in gaze behaviors between the two groups. For E-L1 candidates, none of the gaze behaviors on the test items across the different stages of the test predicted their overall listening test scores. In contrast, the overall listening test scores for E-L2 candidates was predicted by the average proportion of time that they had dwelled on the test items while simultaneously answering them and listening to the auditory text. The study is the first to show that keyword matching on the test items during the while-listening stage significantly contributes to WLP test performance for E-L2 candidates. These results suggest that the focal construct of the listening test is confounded by group-specific reading behaviors on the test items. In line with previous research, the use of the WLP test format for assessing second language listening comprehension is not recommended.

Keywords: Attention, Eye-tracking, Listening test, Construct-irrelevant variance, While-listening performance test

Introduction

Listening comprehension involves an interaction between top-down (i.e., inferential) and bottom-up (i.e., literal) cognitive processes and includes allocation of attention as well as working and long-term memories that allow an individual to deduce meaning from auditory stimuli (Hulstijn, 2003; Imhof, 2010; Vandergrift, 1999, 2004; Vandergrift & Goh, 2012). For university students, these cognitive processes are not only engaged during daily tasks (Rost, 2016), but are also integrated with other linguistic processes such as material-reading and notetaking during academic lectures (Charles & Pecorari, 2016; Graham, 2011; Harding, 2011; Kim, 2019; Song, 2011; Wang, 2018).

Such listening comprehension ability can be examined through while-listening performance (WLP) tests as candidates have to concurrently listen to a lecture, take notes, read and answer the test items (Aryadoust, 2012, 2020). Notably, the listening components of some English language proficiency tests such as the Canadian Academic English Language (CAEL) and the International English Language Testing System (IELTS) are examples of WLP tests. While these high-stakes tests are often undertaken by non-native English-speaking (E-L2) candidates, they are also taken by native English-speaking (E-L1) candidates for career and migration purposes. Understanding how E-L1 and E-L2 candidates performed in WLP tests is important as such information may not only assist academics in designing their teaching processes, syllabus, and classroom activities to support diversity, but also employers and/or immigration authorities in gauging if such blanket assessments are suitable for selecting candidates for career or migration purposes (especially in situations where both E-L1 and E-L2 candidates are equal on other fronts).

However , little is known about how E-L2 candidates fare against E-L1 candidates under WLP academic listening tests as the majority of group performance comparisons across the literature was focused on post-listening performance (PLP) tests (i.e., another form of listening comprehension assessment where the tests items are presented following the listening text) (Babayiğit, 2012; Burgoyne et al., 2009; Conrad, 1985; Dunkel et al., 1989; Hutchinson et al., 2003; Major et al., 2002; Marx et al., 2017). In all these PLP tests, E-L1 candidates outperformed E-L2 candidates. Specifically, Conrad (1985) reported that E-L1 candidates relied more on contextual semantic cues whereas E-L2 candidates paid more attention to syntactic cues from the text when filling in a cloze test after listening to a short lecture. Together, these group differences are expected as Rost (2014) has previously suggested several cognitive challenges (including top-down and bottom-up processing) as to why E-L2 listeners can still possess inadequate second language (L2) listening proficiency despite learning the language over an extended period of time.

In contrast to a PLP test, all candidates will sit through three stages of listening comprehension (SLC) during a WLP test: pre-listening, while-listening, and post-listening (Yildiz et al., 2015). In the pre-listening stage, the candidates are provided time not only to listen/read the instructions but also preview the test items. Subsequently in the while-listening stage, they are permitted to read and answer the test items while listening to the lecture recording. In high-stakes WLP tests like the CAEL and IELTS, the lecture recordings are only played once to better reflect real-life listening scenarios such as academic lectures (Field, 2015). Lastly in the post-listening stage, time is given to the candidates to complete, check and/or revise their answers.

Research in WLP tests has been centered around the cognitive validity of the pre-listening and while-listening stages (Field, 2009, 2013, 2015), where cognitive validity is defined as the degree to which the cognitive processes engaged during the tests reflect that of real-life listening events (Weir, 2005). The previewing of test items in the pre-listening stage has raised concerns as the wordings and order of items presented can serve as visual cues to influence the manner with which E-L2 candidates subsequently source for and utilize the visuo-auditory information that is subsequently presented in the while-listening stage (Field, 2009). Through retrospective verbal reports, studies have shown that both E-L1 and E-L2 candidates utilized a range of test-wise strategies such as lexical search or keyword matching during single-hearing WLP tests and they are indicative of visual cue interference in the listening process (Badger & Yan, 2009; Field, 2009, 2011). Field (2009) reported that such test-wise behaviors can potentially impede performance when E-L2 candidates fail to catch the point that provided the correct answer as their attention is drawn away from the lecture recording towards the test items in the while-listening stage. Such level of focus at the word level, however, facilitated shallow rather than deep comprehension (Field, 2009). Notably, this is supported by Aryadoust et al. (2020)'s recent functional near-infrared spectroscopy (fNIRS, a neuroimaging method) study where in a combined sample of E-L1 and E-L2 candidates exhibited lower activity levels across cortical regions known to support top-down (i.e., dorsomedial prefrontal cortex) and bottom-up processing (i.e., inferior frontal gyrus and posterior middle temporal gyrus) during the while-listening stages in WLP tests compared with notetaking-while-listening during PLP tests. These differences observed between WLP tests and notetaking-while-listening to a lecture recording during PLP tests (Aryadoust et al., 2020) and a simulated academic lecture (Field, 2009) may be attributed to the fact that the written notes are produced by the candidates in the latter activities, whereas the 'notes' (i.e., in the form of test items) are provided by test developers in the former event and may serve as "an unseen text that has to be mastered" by the candidates (Field, 2009, p. 46). Thus, the need to read and master the 'notes' under time pressure in WLP tests will require efficient test-specific reading strategies such as skimming and scanning that are not recognized as components of listening (Vandergrift, 2007; Vandergrift & Goh, 2012). Importantly, the use of such test-specific strategies can introduce sources of construct-irrelevant variance (CIV) and affect test scores, resulting in 'invalidly low' or 'invalidly high' test scores (Messick, 1996; Millman et al., 1965; Rogers & Bateson, 1991).

Although the majority of the abovementioned studies have provided some insights into the E-L1 and E-L2 candidates' processes through retrospective verbal reports (Badger & Yan, 2009; Field, 2009, 2011), the interpretations of these findings may be limited due to interviewers' instructional cue bias and candidates' limited verbal skills and recall abilities (Matsumoto, 1993; Sasaki, 2013). To further understand the cognitive processes associated with E-L2 candidates' test-taking strategies during WLP tests, some studies have used stimulated recalls and incorporated eye-tracking technology to alleviate some of the verbal report limitations associated with memory (Holzknecht et al., 2017; Suvurov, 2018; Winke & Lim, 2014). Alternatively, eye-tracking has also been used as a stand-alone to examine either E-L1 or E-L2 candidates' gaze behaviors on the test items in real-time during WLP or similar tests (Aryadoust, 2020; Holzknecht et al., 2017; Winke & Lim, 2014) as their durational and count/frequency measurement aspects are often regarded as surrogate measures of visual attention and/or cognitive loads (Conklin et al., 2018; Just & Carpenter, 1984; Kruger et al., 2014). Through eye-tracking measures, Winke and Lim (2014) reported that the high-scoring E-L2 candidates were able to read the (fill-in-the-blank) test items more quickly than low-scoring E-L2 candidates and this may allow them to spend more time processing the visual information surrounding the blanks in the items. Furthermore, it has also been suggested that this ability can either provide E-L2 candidates "an advantage on listening test or it may be evidence of their pre-existing advantage in listening" (Winke & Lim, 2014, p. 18). However, no predictive analysis (e.g., linear regression) was performed to substantiate this suggestion. Interestingly, a recent eye-tracking study by Conklin et al. (2020) has shown that E-L1 candidates are faster readers compared with E-L2 candidates when reading story passages, while they showed similar gaze behavior patterns when reading-while-listening to these passages.

Taken together, a question may then be asked as to whether E-L1 and E-L2 candidates would exhibit similar or different gaze behavioral patterns during a WLP test. Presently, little is also known about how E-L1 and E-L2 candidates would gaze upon test items across all three SLC in a WLP test, and whether these gaze behaviors can advantage or disadvantage the candidates on their measured listening performances. It is imperative to investigate these questions as they may provide evidence supporting or attenuating the interpretations and uses of WLP test scores.

Present study

Considering the gaps across the literature, this study incorporates eye-tracking technology as a stand-alone to measure the gaze behaviors of E-L1 and E-L2 candidates while they partake in a WLP test. Specifically, this study aims to answer the following research questions:

(1) Can E-L1 candidates outperform E-L2 candidates in a WLP test?

(2) Do E-L2 candidates gaze at test-items differently from E-L1 candidates across the three SLC of a WLP test?

(3) Can the gaze behaviors of E-L1 and E-L2 candidates predict their measured listening performances?

Methods

Participants

One hundred neurotypical participants (mean ± standard error: 26.3±0.6 years; 62 females and 38 males) were recruited for this study, comprising 57 E-L1 and 43 E-L2 candidates. All participants were students from a public university in Singapore and they were compensated SGD$10 upon completing the study. This study was approved by the university's Institution Review Board.

Listening comprehension test

The CAEL test is designed to assess the English proficiency levels of candidates for admission into tertiary institutions in Canada (Malone, 2010). Unlike IELTS, this test is not used as an admission criterion at the university where the experiment is conducted. Provided by the test developer – Paragon Testing Enterprises Inc., two out of three sections of a CAEL Computer Edition (CE) listening comprehension test were used as assessment materials in this study to minimize potential participant test-procedural bias (see Table 1 for details on the test sections).

Table 1 Descriptions of the listening comprehension test sections

	Section 3 (Integrated Listening)	Section 4 (Academic Unit A)
Lecture topics	Economics	World History
Speech rate (words/min)	160	132
Item descriptions	10 MCQ (four-option) 1 MSQ (two from five options)	8 MCQ (four-option) 2 Fill-in-the-blank items 1 Matching item
Number of pages and items per page	5 pages, 1 to 3 items	5 pages, 1 to 4 items
Font type and size of test items	Trebuchet MS, 11-point	

Abbreviations: MCQ=Multiple-choice question or item, and MSQ=Multi-select question or item.

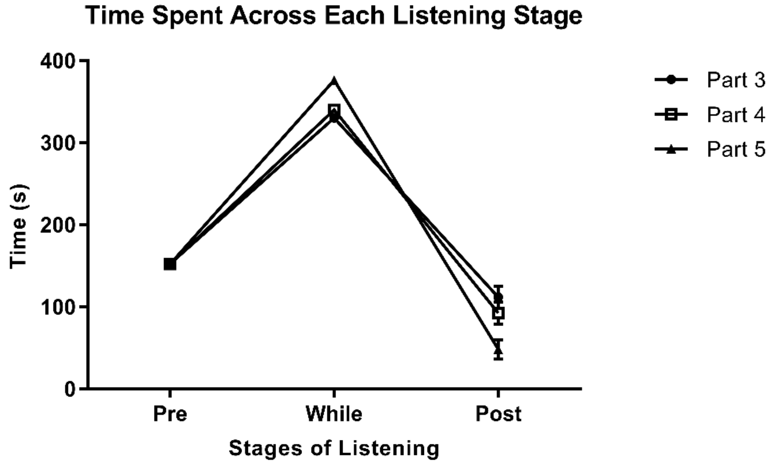


Fig. 1 Time spent across each listening stage. Each set of data point and error bars represent the mean and the 95% confidence intervals respectively

As the three sections were conducted in sequential order, it was necessary to determine if the participants' behaviors were influenced by order effects before further analyses on the participants' test scores and eye movement data were conducted. Thus, the durations of the post-listening stage across the three sections were of interests. We calculated the average amount of time spent by the participants in the post-listening stage, alongside the 95% confidence intervals (CI) for each test. In this study, order effects were considered significant when the 95% CI for the 9 average duration of each of the three post-listening stages did not overlap. Illustrated on Figure 1 is the amount of time spent by the participants across the different SLCs for each test. Evidently, the duration of the post-listening stage for section 5 (Mean (CI): 48.30s (36.31s–60.28s) was shorter than both sections 3 (Mean (CI): 111.75s (98.43s–125.07s) and 4 (Mean (CI): 92.43s (79.11s–105.85s), while no differences was observed between sections 3 and 4. This was indicative that section 5 was influenced by order effects, and thus only data from sections 3 and 4 were analyzed in this study.

Data collection procedures

All participants undertook the two aforementioned listening comprehension test sections in a sequential order within a single session in a laboratory. Both test sections were hosted on a .html website via Tobii Pro Studio Version 3.4.8 (Tobii AB, 2017) and were displayed onto a secondary 23-inch monitor (Hewlett Packard EliteDisplay S231d, 1920 x 1080 pixels, Hewlett Packard, CA, USA) that was connected to a primary laptop (HP Pavilion 15, Hewlett Packard). With their heads unrestrained, the participants sat on a chair approximately 0.65m in front of the monitor, with a set of keyboard and mouse placed on the desk in front of them. Their eye movements were recorded binocularly at 300Hz using a screen-based eye-tracker (Tobii TX300, Tobii AB, Stockholm, Sweden) that was affixed on the monitor. Prior to data collection, a five-point calibration procedure was performed in Tobii Pro Studio Version 3.4.8 (Tobii AB, 2017) to establish each participants' gaze in relation to the monitor screen.

Following calibration, the tests began. As per standard CAEL CE procedures, all participants were given 150s to preview the test items at their own pace in the pre-listening stage. Subsequently in the while-listening stage, the lecture (approximately 330s) was automatically played through a set of speakers and the participants were permitted to answer the items as they listened. After the lecture had concluded, the participants were given an additional 180s to complete, check and/or revise their answers in the post-listening stage.

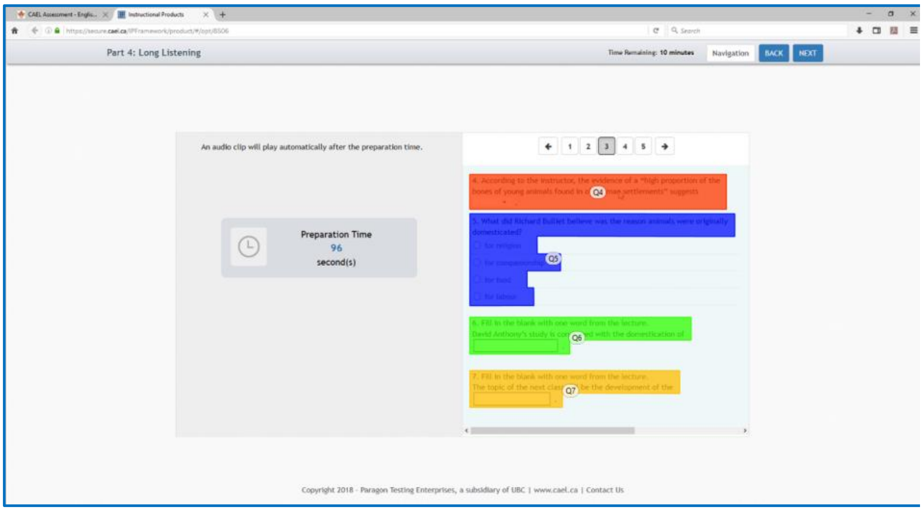
Data processing

Five out of the 100 sets of eye-tracking data were firstly removed due to (i) participants' failure in following the tests instructions or (ii) eye-tracking data not being recorded during the experiment. A further 29 sets of eye-tracking data were excluded from further reduction and analyses as their gaze samples were <80% across the listening comprehension tests (Kruger et al., 2014). Only 66 sets of test scores and eye-tracking data (46 females and 20 males), comprising 43 E-L1 (26.1±0.7years) and 23 E-L2 (25.8±0.7years) candidates, were included in the analyses. Most E-L1 candidates were from Singapore (n=41), while the rest were from Canada (n=1) and Vietnam (n=1). In contrast, most E-L2 candidates were from China (n=18), while the rest were from Indonesia (n=3), India (n= 1), and Myanmar (n=1). The average gaze samples recorded for both groups of participants across both listening comprehension tests were 92±1%.

For test scores, the conventional number-right method was used for scoring both fill-in-the-blank items and multiple-choice questions, where a score of one was awarded for correct responses, and zero for incorrect responses (including blanks). For the other test items (see Table 1), the partial-credit method was used where items were marked as correct, partially correct, or incorrect. When the responses to those items were correct, the maximum scores available were awarded. Partial scores were given when the responses to the items were in part correct, while no point was awarded for incorrect responses (including blanks). All item scores were tabulated via scene image visual inspection in Tobii Pro Studio Version 3.4.8 (Tobii AB, 2017) and combined to establish the participants' overall test scores.

To crop the eye-tracking data for analyses, the unique sequences and durations for all scenes viewed by the participants across each SLC for both listening comprehension tests were firstly annotated in Tobii Pro Studio Version 3.4.8 (Tobii AB, 2017). Using the software, a total of 22 customized polygons were subsequently drawn over the area of interests (AOI(s)) where the test item(s) was(were) presented across the scenes (Figure 2). In this study, the gaze behaviors of interests were fixations and dwells. Here, fixations are temporal pauses in eye movements that are involved in the uptake of text information from the test items (Hessels et al., 2018). In contrast, a visit to or a dwell on the AOI is defined as the time interval between the first and last fixations on the test items (Tobii AB, 2016). Thus, a visit/dwell encompasses fixations and saccades for which the latter are defined as rapid eye movements that bring other parts of the text from the test items onto the retina for information uptake (Hessels et al., 2018). In order to quantify the two gaze behaviors of interests, the raw eye-tracking data were interpolated with a maximum gap length set at 75ms (Komogortsev et al., 2010) following the drawing of AOIs. The interpolated data were then averaged across both eyes and smoothed using a moving median noise-reduction filter with a window size of 3 samples (Juhola, 1991). Lastly, fixation eye movements were parsed using a velocity threshold-identification filter (Stuart et al., 2019) with minimum velocity and fixation duration thresholds set at 30°/s (Olsen & Matos, 2012) and 100ms respectively (Rayner, 1998).

Four variables of interests were subsequently extracted for every AOI (n=22) across the three SLC including, (i) visit counts, (ii) total visit duration, (iii) fixation counts, and (iv) total fixation duration (see Tobii AB (2016) for variable definitions). In this study, all 22 AOIs were regarded as a single entity. Thus, the aforementioned dependent variables were summed across each SLC and test section. The average visit and fixation durations on the test items (i.e., AOI) for each SLC across both assessments were calculated when



Part 4: Long Listening

An audio clip will play automatically after the preparation time.

Preparation Time
96
second(s)

4. According to the instructor, the evidence of a "high proportion of the bones of young animals found in c mae settlements" suggests

5. What did Richard Bulliet believe was the reason animals were originally domesticated?

○ for religion

○ for companionship

○ for food

○ for labour

6. Fill in the blank with one word from the lecture.
David Anthony's study is con ed with the domestication of

7. Fill in the blank with one word from the lecture.
The topic of the next clas be the development of the

Copyright 2018 - Paragon Testing Enterprises, a subsidiary of UBC | www.cael.ca | Contact Us

CAEL Assessment - Englis...
Instructional Products
https://secure.cael.ca/IPFramework/product/#/opt/8506
Search
Time Remaining: 10 minutes
Navigation
BACK
NEXT
1
2
3
4
5
Q4
Q5
Q6
Q7

Fig. 2 A sample layout of the listening comprehension test section

Notes: Customized polygons were drawn over each test item in Tobii Pro Studio Version 3.4.8 (Tobii AB, 2017). Only gaze behaviors that fell within the test items were analyzed. The colors of the polygons do not bare any significance in terms of calculations.

the total visit and fixation durations were divided respectively by visit and fixation counts. To account for the intra- and inter-participant differences in the durations across the SLC, the six variables obtained across all participants were subsequently normalized to the amount of time that was spent during each of SLC (Conklin et al., 2018). The normalized total visit duration, normalized average visit duration, normalized total fixation duration, and normalized average fixation duration on the test items were expressed in percentages, while normalized visit and fixation counts were expressed as number of visits per minute (visits/min) and number of fixations per minute (fixations/min) respectively.

Statistical analysis

Using IBM SPSS Statistics for Windows Version 25 (IBM Corporation, 2017), the internal consistency of the listening test used was firstly established using Cronbach's alpha (Cronbach, 1951). The Cronbach's alpha for all 22 test items was .62. According to the recommendations by Nunnally and Bernstein (1994), the listening test used in this study possessed an acceptable level of reliability.

Subsequently, the normality across all test scores and gaze behavioral data was tested by assessing the Shapiro-Wilk statistics. As all data violated the assumption of normality, non-parametric statistical analyses were performed in IBM SPSS Statistics for Windows Version 25 (IBM Corporation, 2017) with alpha value set at 0.05. To analyze the differences in overall measured listening performances between E-L1 and E-L2 candidates, the Mann-Whitney *U* test was used on the overall test scores. In terms of the gaze behavioral measurements, the Aligned Rank Transform (ART) procedure was firstly performed using ARTool to align and rank the data prior to non-parametric factorial analyses (Wobbrock, 2011; Wobbrock et al., 2011). To account for the small and unequal sample sizes, the transformed data was analyzed using the linear mixed model (i.e., restricted maximum likelihood) procedures with 'L1' and 'SLC' listed as the fixed effects (Patterson & Thompson, 1971). If the main effect of 'SLC' was statistically significant, post-hoc multiple pairwise comparisons were performed and adjusted using Bonferroni corrections. Any significant post-hoc cross-effect pairwise comparisons with Bonferroni corrections were, however, analyzed using the Mann-Whitney *U* tests (i.e., between-group analysis) or Wilcoxon tests (i.e., within-group analysis following a statistically significant Friedman test) on the original data (Wobbrock, 2011). All descriptive data in the results section were presented in mean ± standard error.

Lastly, the automatic linear modelling procedure (i.e., forward stepwise model) as described by Yang (2013) was also performed in IBM SPSS Statistics for Windows Version 25 (IBM Corporation, 2017) to establish which of the 18 gaze behavior measures across the three SLC was (were) the best predictor(s) of E-L1 and E-L2 candidates' overall measured listening performance.

Results

E-L1 and E-L2 candidates measured listening performance

Based on the combined test scores, E-L1 candidates (18.9±0.4points) performed significantly better than E-L2 candidates (16.8±0.7points) in the listening tests, *Z*=- 2.44, *p*=.015. In addition, Table 2 demonstrates item facility (proportion of correct answers) across E-L1 candidates, E-L2 candidates, and the entire sample. Incorrect responses comprise test items that were either not answered correctly or left blank by the participants. When test items require more than one answer, partially correctly responses indicate that the participants did not fully answered the items correctly. For item 11 in test 1, two answers were required. Thus, partially correct responses denote that the participants only had one answer correct. Similarly, four answers were required for item 1 in test 2. Here, partially correct responses indicate that the participants only had two of the answers correct.

For E-L1 candidates, the easiest items were items 5 and 6 in test 2 (facility = 1.00 & .977, respectively). In contrast, items 10 (test 1) and 6 (test 2) with the facility index of .87 were the easiest for E-L2 candidates. We conducted three related-samples Wilcoxon signed rank tests to compare the facility indices of test 1 and 2 for E-L1 and E-L2 candidates. No significant differences were found, indicating that the test items had no psychometric differences.

Differences in gaze behavioral measures between E-L1 and E-L2 candidates

For normalized total visit duration on the test items, the non-parametric factorial analysis indicated there was no significant main effects for 'L1', *F*(1,389)=2.91, *p*=.089, and 'SLC' , *F*(2,389)=0.68, *p*=.51. Likewise, the interaction effect between 'L1' and 'SLC' was not statistically significant, *F*(2,389)=1.06, *p*=.35 (Figure 3A, see Supplementary Material 1 for specific details). Thus, the proportion of time spent visiting the test items in total was neither different between E-L1 (65±1%) and E-L2 (66±1%) candidates, nor across the three SLC (pre-listening (65±1%), while-listening (66±1%), post-listening (65±2%)). In terms of the normalized total fixation duration on the test items, E-L2 candidates (56±1%) spent significantly larger proportions of their time fixating on the test items in total than E-L1 candidates (54±1%), *F*(1,389)=5.63, *p*=.018. There was, however, no difference in normalized total fixation duration across the pre-listening (55±1%), while-listening (55±1%), and post-listening stages (54±1%), *F*(2,389)=0.09, *p*=.91. Furthermore, the influence of 'L1' on normalized total fixation duration did not depend on the 'SLC' as the interaction effect was not statistically significant, *F*(2,389)=1.06, *p*=.35 (Figure 3B, see Supplementary Material 1 for specific details).

Table 2 Item facility and frequency of responses for each of the questions in the listening comprehension tests

		Overall (*N* = 66)			E-L1 (*N* = 43)			E-L2 (*N* = 23)		
Test	Question	Incorrect (N)	Correct (N)	Item facility	Incorrect (N)	Correct (N)	Item facility	Incorrect (N)	Correct (N)	Item facility
1 (section 3)	1	27	39	0.591	17	26	0.605	10	13	0.565
	2	14	52	0.788	9	34	0.791	5	18	0.783
	3	17	49	0.742	11	32	0.744	6	17	0.739
	4	10	56	0.848	5	38	0.884	5	18	0.783
	5	27	39	0.591	19	24	0.558	8	15	0.652
	6	21	45	0.682	13	30	0.698	8	15	0.652
	7	8	58	0.879	3	40	0.930	5	18	0.783
	8	25	41	0.621	11	32	0.744	14	9	0.391
	9	26	40	0.606	12	31	0.721	14	9	0.391
	10	9	57	0.864	6	37	0.860	3	20	0.870
2 (section 4)	2	17	49	0.742	9	34	0.791	8	15	0.652
	3	10	56	0.848	4	39	0.907	6	17	0.739
	4	61	5	0.076	41	2	0.047	20	3	0.130
	5	4	62	0.939	0	43	1.000	4	19	0.826
	6	4	62	0.939	1	42	0.977	3	20	0.870
	7	27	39	0.591	14	29	0.674	13	10	0.435
	8	55	11	0.167	36	7	0.163	19	4	0.174
	9	34	32	0.485	25	18	0.419	9	14	0.609
	10	34	32	0.485	23	20	0.465	11	12	0.522
	11	23	43	0.652	8	35	0.814	15	8	0.348

Test	Question	Incorrect (N)	Partially Correct (N)	Correct (N)	Incorrect (N)	Partially correct (N)	Correct (N)	Incorrect (N)	Partially correct (N)	Correct (N)
1 (section 3)	11	9	39	18	4	25	14	5	14	4
2 (section 4)	1	0	3	63	0	2	41	0	1	22

As for the normalized average visit duration on the test items, the main effects for ‘L1’ (*F*(1,389)=10.02, *p*=.002) and ‘SLC’ (*F*(2,389)=426.26, *p*<.0001) were statistically significant. However, the interaction effect between ‘L1’ and ‘SLC’ was also statistically significant, *F*(2,389)=17.85, *p*<.0001. This indicated that the influence of ‘L1’ on this dependent variable depended on the ‘SLC’ (Figure 3C). Further post-hoc analysis with Bonferroni corrections (alpha value (α)=0.017) indicated E-L1 candidates spent significantly lower proportions of their time visiting the test items on average than E-L2 candidates during the pre-listening (E-L1=0.94±0.04%, E-L2=1.16±0.06%, *Z*=-3.19, *p*=.0014) and while-listening stages (E-L1=0.35±0.01%, E-L2=0.43±0.02%, *Z*=-3.78, *p*<.001), but not the post-listening stage (E-L1=2.71±0.23%, E-L2=2.15±0.31%, *Z*=-2.12, *p*=.034). Additionally, Friedman tests indicated that the normalized average visit duration on the test items were significantly different across the three SLC for both E-L1 (χ^2(2,N=86)=150.58, *p*<.0001) and E-L2 candidates (χ^2(2,N=45)=60.40, *p*<.0001). For E-L1 candidates, they spent significantly lower proportions of their time visiting the test items on average during the while-listening stage (0.35±0.01%) than in the pre-listening (0.94±0.04%, *Z*=-7.07, *p*<.0001) and post-listening stages (2.71±0.23%, *Z*=-8.05, *p*<.0001). Furthermore, E-L1 candidates also spent lower proportions of their time visiting the test items on average during the pre-listening stage (0.94±0.04%) than in the post-listening stage (2.71±0.23%, *Z*=-8.05, *p*<.0001). For E-L2 candidates, the normalized average visit duration on the test items during the while-listening stage (0.43± 0.02%) was also shorter than both the pre-listening (1.16±0.06%, *Z*=-5.91, *p*<.0001) and post-listening stages (2.14±0.31%, *Z*=- 2.93, *p*=.003). The proportion of time that E-L2 candidates spent visiting the test items on average during the pre-listening stage (1.16±0.06%) was also shorter than in the post-listening stage (2.14±0.31%, *Z*=- 5.75, *p*<.0001).

Similarly, the main effects for ‘L1’ (*F*(1,389)=95.66, *p*<.0001) and ‘SLC’ (*F*(2,389)=363.25, *p*<.0001) were statistically significant for normalized average fixation duration on the test items. As the interaction effect between ‘L1’ and ‘SLC’ was also statistically significant (*F*(2,389)=41.35, *p*<.0001), this suggested that influence of ‘L1’ on the normalized average fixation duration depended on the ‘SLC’ (Figure 3D). Further post-hoc analysis with Bonferroni corrections (α=0.017) indicated that E-L1 candidates spent significantly lower proportions of their time fixating on the test items on average compared with E-L2 candidates during the pre-listening (E-L1=0.152±0.002%, E-L2=0.171±0.002%, *Z*=-6.00, *p*<.0001) and while-listening stages (E-L1=0.071±0.001%, E-L2=0.080±0.001%, *Z*=-3.19, *p*<.0001). In contrast, the average proportion of time that E-L1 candidates (0.444±0.056%) spent fixating on the test items was significantly higher than E-L2 candidates (0.254±0.023%) during the post-listening stage, *Z*=-2.91, *p*=.004. The Friedman tests also indicated that the normalized average fixation duration on the test items were significantly different across the three SLC for both E-L1 (χ^2(2,N=86)=148.56, *p*<.0001) and E-L2 candidates (χ^2(2,N=45)=67.60, *p*<.0001). On average, E-L1 candidates spent significantly lower proportions of their time fixating on the items during the while-listening stage (0.071±0.001%) than in the pre-listening (0.152±0.002%, *Z*=-8.06, *p*<.0001) or post-listening stages (0.444±0.056%, *Z*=-8.05, *p*<.0001). Furthermore, the normalized average fixation duration on the test items for E-L1 candidates was also significantly shorter during the pre-listening stage (0.152±0.002%) than in the post-listening stage (0.444±0.056%, *Z*=-7.20, *p*<.0001). On average, E-L2 candidates also spent lower proportions of their time fixating on the items during the while-listening stage (0.080±0.001%) than in both pre-listening (0.171±0.002%, *Z*=-5.91, *p*<.0001) and post-listening stages (0.254±0.023%, *Z*=-5.84, *p*<.0001). Additionally, the normalized average fixation duration on the test items for E-L2 candidates was also significantly shorter during the pre-listening stage (0.171±0.002%) than in post-listening stage (0.254±0.023%, *Z*=- 2.61, *p*=.009).

In terms of the normalized visit counts, both the main effects of ‘L1’ (*F*(1,390)=47.33, *p*<.0001) and ‘SLC’ (*F*(2,390)=41.37, *p*<.0001) were statistically significant. The interaction effect between ‘L1’ and ‘SLC’ for normalized visits counts on the test items was, however, not statistically significant, *F*(2,390)=1.23, *p*=.30 (Figure 3E, see Supplementary Material 1 for specific details). Thus, E-L1 candidates (30±1visits/min) made more visits to the test items per unit time compared with E-L2 candidates (25±1visits/min). Further post-hoc analysis with Bonferroni corrections (α=0.017) indicated that all participants made significantly more visits to the test items per unit time during the while-listening stage (27±1visits/min) than both the pre-listening (33±1visits/min, *p*<.0001) and post-listening stages (24±1visits/min, *p*=.008). Also, all participants made significantly more visits to the test items per unit time during the pre-listening stage (33±1visits/min) than in the post-listening stage, (24±1visits/min, *p*<.0001). Lastly for normalized fixation counts on the test items, only the main effect for ‘L1’ was statistically significant, *F*(1,390)=9.11, *p*=.003. This indicated that E-L1 candidates (137±2 visits/min) made more fixations on the test items per unit time compared with E-L2 candidates (128±3visits/min). There was, however, no difference in normalized fixation counts on the test items across pre-listening (137±3fixations/min), while-listening (133±2fixations/min) and post-listening stages (131±3fixations/min) as the main effect of SLC was not statistically significant, *F*(2,390)=0.90, *p*=.41. Similarly, the interaction between ‘L1’ and ‘SLC’ for normalized fixation counts on the test items was also not statistically significant, *F*(2,390)=1.07, *p*=.35 (Figure 3F, see Supplementary Material 1 for specific details).

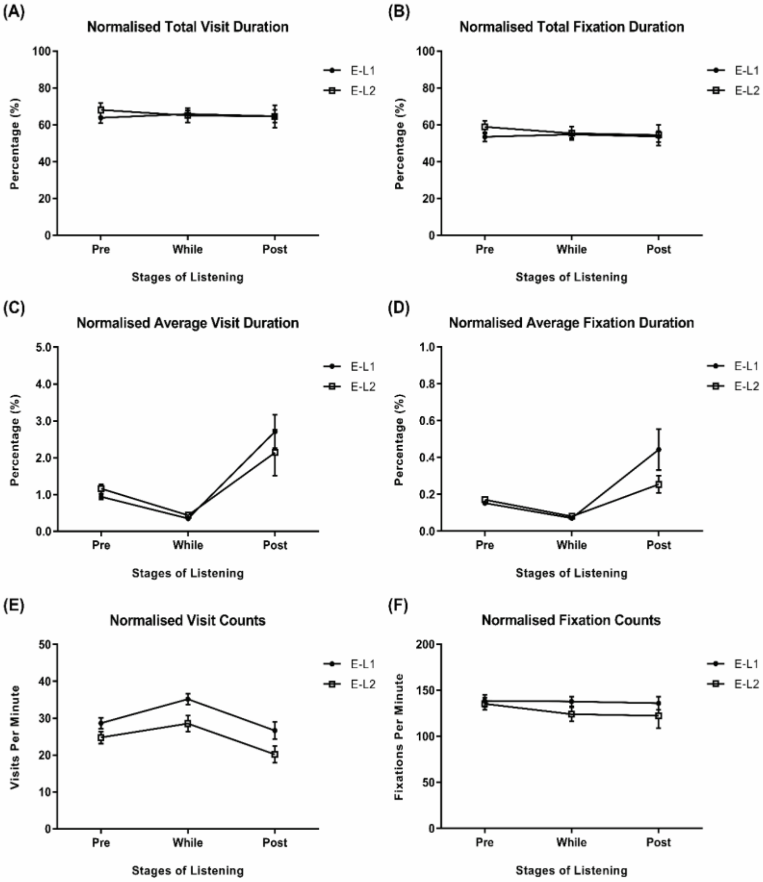


Fig. 3 Normalized gaze behavioral data across the pre-, while-, and post-listening stages. Notes: The data on each plot reflect the means, while the error bars represent the 95% confidence intervals. Abbreviations: E-L1=Native English-speaking candidates, and E-L2=Non-native English-speaking candidates

Gaze behavioral measures as predictor(s) of E-L1 and E-L2 overall measured listening performances

For E-L1 candidates, the automatic linear modelling procedure returned a parsimonious model that comprised normalized average visit duration on the test items during the post-listening stage as its sole predictor term. The results from the forward stepwise method indicated that the normalized average visit duration on the test items during the post-listening stage did not significantly predict E-L1 candidates' overall listening test scores, $R^2=0.04$, $F(1,40)=2.51$, $p=.12$. In contrast, the regression analysis generated a parsimonious model that included average normalized visit duration on the test items during the while-listening stage as its sole predictor term for E-L2 candidates. The results indicated that the normalized average visit duration on the test items during the while-listening stage significantly predicted the E-L2 candidates' overall measured listening performance, $R^2=0.19$, $F(1,18)=5.21$, $p=.036$.

Discussion

Through a WLP test, the aims of this study were to examine (i) whether E-L1 and E-L2 candidates differ in their overall measured listening performances; (ii) whether the candidates' L1 can affect their gaze behaviors on the test items across the three SLC; and (iii) whether E-L1 and E-L2 candidates' gaze behaviors on the test items across the three SLC can predict their overall measured listening performances. The results of this study showed that E-L1 candidates outperformed E-L2 candidates with higher overall test scores across the two sections of the CAEL CE WLP test. While the overall dynamicity of gaze behaviors on the test items across the three SLC was similar between E-L1 and E-L2 candidates, there were distinctive differences in gaze behaviors between the two groups. Lastly, the overall measured listening performance for E-L2 candidates was significantly predicted by the average proportion of time that they spent visiting/dwelling on the test items while simultaneously listening to the text and answering the test items. In contrast, none of the gaze behavioral measures across the three SLC significantly predicted the overall test scores for E-L1 candidates. The three research questions of this study are discussed below.

Research Question 1: Can E-L1 candidates outperform E-L2 candidates in a WLP test?

Our findings resonate with other studies reporting that E-L1 candidates performed better with higher test scores than E-L2 candidates across various educational levels and PLP tests (Babayiğit, 2012; Burgoyne et al., 2009; Conrad, 1985; Dunkel et al., 1989; Hutchinson et al., 2003; Major et al., 2002; Marx et al., 2017). These results were expected as Rost (2014) had previously highlighted several challenges in L2 listening that were beyond mere disparities in language acquisition between E-L1 and E-L2 candidates. Examples of the differences discussed included attention derailment and the inability of (some) E-L2 candidates to suppress their natural tendency to listen from an L2 perspective (Rost, 2014).

This finding has two implications for assessing English listening comprehension in the context of high-stakes tests. The first implication relates to the standardizing of test scores. As the goal of test score standardization is to rank all candidates on the same scale, E-L2 candidates will be at a disadvantage compared with E-L1 candidates during the ranking process as test scores are used for high-stakes decision-making without further information. Notably, many high-stakes tests of English that are used for career and migration purposes (e.g., the CAEL CE and IELTS) are taken by diverse groups of E-L1 and E-L2 candidates. While any observed differences in measured listening performances between the two groups are not necessarily indicative of test bias (Aryadoust et al., 2011), E-L2 candidates may still be particularly disadvantaged in job and/or migration application processes in English-speaking countries especially when they have similar qualifications as E-L1 candidates. Perhaps future research can incorporate differential item functioning analysis and measurement invariance to investigate whether differences in test scores between E-L1 and E-L2 candidates during such high-stakes listening tests are associated with bias against the latter group.

The second implication of our finding concerns the design of listening tests. While the listening construct is defined as a process of literal and inferential comprehension (Buck, 2001), operationalizing this definition as a listening test needs to define the population of interests and their characteristics (Luoma, 2004). If the test designers were motivated by the notion of ‘native-speakerism’ whereby E-L2 candidates should ideally listen and perform similarly as E-L1 candidates (Holliday, 2005, 2006), then the assessment designed may be tailored for a particular group who not only speak with a specific variety of English, but are also from a certain culture. Since language and culture are inseparable, it is likely that the test developer may devise the test items to fit the culture of E-L1 candidates and over-represent one variety of English while under-representing others. This can in turn affect the level of cognitive difficulty that test developers build into the test. Perhaps test developers need to consider a ‘common identity’ that is proposed by Holliday (2005) when testing English listening comprehension. Including various English varieties in listening assessments and considering an international population for the assessment may help mitigate the problem of under- and over-representing the varieties of English.

Research Question 2: Do E-L2 candidates gaze at test-items differently from E-L1 candidates across the three SLC of a WLP test?

Irrespective of the SLC, both normalized fixation and visit counts were significantly higher for E-L1 than E-L2 candidates. The proportion of time that E-L2 candidates spent fixating on the test items in total regardless of the SLC was significantly greater than E-L1 candidates (i.e., approximately 2%) yet the proportion of time spent visiting the test items in total was not significantly different between the two participant groups (i.e., approximately 1%). Although saccadic frequency was not specifically measured in this study, the absence of normalized total visit duration on the test items suggests the data may have been offset by E-L1 candidates’ higher saccadic frequency on the test items across the three SLC compared with E-L2 candidates as a visit to the test items encompassed both saccades and fixations in this study (see definitions in Methods). To the best of our knowledge, no study has specifically compared E-L1 and E-L2 candidates’ saccadic eye movements behaviors during listening comprehension tests. Nevertheless, only one research study has provided in depth comparisons of gaze behaviors between E-L1 and E-L2 participants during reading comprehension (Conklin et al., 2020). Importantly, our findings and interpretations are congruent with Conklin et al. (2020) who reported that E-L1 participants were more fluent and read story passages faster than E-L2 candidates (i.e., shorter total reading time), with shorter fixation durations and higher word-skipping.

While the normalized average fixation and visit durations were higher for E-L2 than E-L1 candidates across both pre-listening and while-listening stages, E-L1 candidates spent larger proportions of their time on average fixating and visiting/dwelling on the test items during the post-listening stage than E-L2 candidates. In accordance with the 'eye-mind hypothesis', there is negligible lag between the word(s) being fixated upon and what a person is attending to or cognitively processing in real-time (Just & Carpenter, 1980). Thus, the duration of any fixation(s) should reflect the attentional or cognitive effort that was needed to process the word(s) that was(were) being gazed upon (Conklin et al., 2018). This in turn suggests that E-L2 candidates require more cognitive effort than E-L1 candidates when reading of test items during question preview and answering while listening to the text. In contrast, E-L1 candidates read the test items more thoroughly than E-L2 participants when completing and checking their responses. While the rationale behind this difference in gaze behaviors during the post-listening stage is largely unknown, we suggest that this may arise due to differences in cultural traits between E-L1 and E-L2 candidates. In this study, most of the E-L1 candidates are from Singapore. It is well-established that the students in Singapore are examination-oriented and often carry the mindset of not wanting to lag behind their peers academically (Bedford & Chua, 2017). Such dispositions may have resulted in E-L1 candidates adopting a more cautious approach during the listening tests by spending larger proportions of their time fixating and reading the test items more thoroughly per visit compared with E-L2 candidates.

Despite the distinctive between-group differences in gaze behaviors, the dynamicity across the three SLC were similar between E-L1 and E-L2 candidates. Importantly, this similarity in the dynamicity of gaze behaviors suggests that the test items in this study not only specify the type of the information that both E-L1 and E-L2 candidates attended to and perceived as possibilities for actions, but it also determined the cognitive processes that are engaged for processing the information from the test items (Gibson, 1979; Kirsh, 2005; Moon et al., 2019; Zhang, 1997). Together, these findings support the verbal reports from Badger and Yan (2009) who reported that both E-L1 and E-L2 candidates used similar strategies, sub-strategies, and tactics during an IELTS WLP test.

Research Question 3: Can the gaze behaviors of E-L1 and E-L2 candidates predict their measured listening performances?

Lastly, the contributions of the measured gaze behaviors towards listening test performances were different for E-L1 and E-L2 candidates. In general, these results were in line with an early study by Conrad (1985) who reported that E-L1 candidates relied more on contextual semantic cues whereas E-L2 candidates attended to the syntactic cues in the text (i.e., cloze passage) during listening comprehension. Specifically, the results from our regression analysis suggest that thorough reading of test items in the post-listening stage did not predict the overall listening test scores for E-L1 candidates. As discussed above, the proportion of time that E-L1 candidates spent visiting the test items during the post-listening stage were likely associated with their cultural traits of not wanting to lag behind their peers academically (Bedford & Chua, 2017). Notably, our finding resonates with a survey study by Ho et al. (1998) who reported that there was no significant linear relationship between the undergraduates' tendency in exhibiting such cultural traits and their academic performances.

In contrast, the regression analysis indicated that 19% of the observed variance in the E-L2 candidates' test scores was explained by their normalized average visit duration on the test items during the while-listening stage. This result suggests that the larger the proportion of time E-L2 candidates spend on the test items per visit, the better their measured listening performances. Interestingly, this relationship appeared to contradict the works of Just and Carpenter (1980) and Conklin et al. (2018) as these authors considered fixation duration (instead of visit duration) as an indirect measure of attention/cognitive processing. However, Orquin and Holmqvist (2018) suggested that fixations and attention may not be perfectly coupled as attention is often shifted even prior to the end of a fixation. Perhaps, this may be a reason why normalized average visit duration on the test items instead of normalized average fixation duration during the while-listening stage predicted the overall listening test scores for E-L2 participants even though both sets of data were similar in this study (see Figure 3C and 3D). However, it is important to note that this finding does not invalidate the works of other researchers who considered fixation duration as an indirect measure of attention/cognitive processing as the choice of dependent variables are task/stimuli dependent.

More importantly, our finding emphasizes that the focal construct of the listening tests used in this study is likely confounded by E-L2 candidates' reading behaviors on the test items (Messick, 1996; Millman et al., 1965; Rogers & Bateson, 1991). This is in line with the works of Friedman and Ansley (1990) who previously found that the amount of reading, as investigated through the manipulation of printed information on the multiple-choice questions of the Iowa Tests of Basic Skills Listening Supplement, introduced sources of CIV that inflated the candidates' listening test scores. Additionally, our finding further extends the study of Winke and Lim (2014) and suggests that the ability to spend more time processing the text information in the test items can provide E-L2 candidates an advantage on the listening test. This finding has an important implication for the assessment of listening comprehension as it suggests that the WLP test method can significantly advantage E-L2 candidates with higher test scores by reading the test items more thoroughly while listening to the text. As such behavior is systematic, this may suggest that the candidates used a test-specific strategy known as "keyword matching" where they listened out for keywords or phrases in the lecture and matched them against those in the test items (Field, 2009). Presently, none of the previous studies presented evidence that this strategy can significantly influence test performance (Aryadoust, 2020; Field, 2009; Holzknecht et al., 2017; Winke & Lim, 2014). Our study is the first to show that keyword matching by means of normalized average visit duration on the test items during the while-listening stage significantly contributes to WLP test performance for E-L2 candidates. Additionally, our finding resonates with Field (2009) and questions the cognitive validity or authenticity of WLP test methods as keyword matching strategy is not used in real-life situations such as academic lectures. Under such circumstances, students often refer to the lecture handouts as a structural scaffold rather than as a source that is overwriting their listening processes and strategies.

Limitations and future research

This study is not without limitations. First, the listening passages were produced based on standard Canadian English by test developers. It may appear that developers assumed that accent, syntax, and phraseological features of such listening passages represent the dominant variant of English in the North American context. However, the spread and diffusion of English or "the emergence of local varieties of English in, for example, India or Singapore" warrant a comprehensive engagement with and an inclusive perspective on English assessment (Davies, 2009, p. 80). In countries like Canada and the United States, there is significant cultural and linguistic diversity. Thus, establishing one variant of English as the gold standard of testing appears to limit the scope of the listening construct. There is an outstanding need to investigate the assessment of the diverse variants of English such as World Englishes, English as a lingua franca, and what Bolton (2006, p. 240) termed the "New English" in multicultural and pluralinguistic societies. The onus of choosing a variant of English as the language of assessment not only falls upon the speech community (Davies, 2009), but also educational policymakers and assessment developers.

The second limitation of the study is the effect of possible extraneous variables which were not controlled for. Notably, the format and length of the test items and the type of listening process or subskill that test items probably engage need to be examined and controlled for in future research. This limitation ensues from the fact that the study was a post-hoc investigation of an operationalized listening test rather than an exploratory attempt to validate the test prior to its official launch. To maintain the ecological validity of the test, we did not alter the test items and preserved its original format. We suggest that future researchers should consider controlling for extraneous variables such as those mentioned above to maximize the validity and replicability of findings.

Conclusion

In summary, this study found that while there were differences between E-L1 and E-L2 candidates' overall measured listening performances and specific gaze behaviors during WLP tests, only the gaze behaviors of E-L2 candidates (i.e., normalized average visit duration during the while-listening stage) on the test items predicted their listening test scores. Although this study may be limited by the smaller and uneven group sample sizes due to our stringent data processing cutoffs (i.e., presence ≥80% gaze data), the statistical methods used were carefully applied to account for this limitation. Nevertheless, our finding suggests that E-L2 candidates' reading behaviors on the test items is a source of group-specific CIV that could contaminate subsequent interpretations and use of listening test scores. Based on our data, the use of the WLP test method that demands multitasking during the test (i.e., concurrent reading, listening, and answering) for the assessment of L2 listening comprehension is not recommended.

To better reflect authentic listening and minimize sources of CIV when developing listening tests, we suggest that test designers incorporate eye-tracking technologies as a stand-alone and include samples of both E-L1 and E-L2 candidates to better understand the cognitive mechanisms that candidates engaged during the tests. Given that both E-L1 and E-L2 candidates make use of visual cues, perhaps more research can be conducted in this field to assess if auditory questions may be a better alternative to visual questions in assessing listening comprehension.

A counterargument is that the deployment of reading strategies in listening tests may not necessarily invalidate listening tests, since reading is integrated with listening in academic contexts such as lecture comprehension wherein the listener alternates their attention between the auditory input (spoken lecture), the slides or written words on blackboards/whiteboards, and possibly their own notes. We concur that academic listening entails multitasking and integrated listening. Nevertheless, the test-taking process of listeners is determined by test format, while in academic lectures, the listener has the freedom of choosing the source of input to attend to at any time. Although using visual materials such as PowerPoint does not seem to have any effect on learning and engagement of college students (Baker et al., 2018), our study showed that the use of written test materials exerts a significant influence on test takers' visual search strategies and test scores. Thus, the use of written test materials does not seem to maximize authenticity in WLP assessments.

Finally, the study also has a pedagogical implication. It is evident that test methods or how test items are presented to test takers can determine a significant share of the variance in test scores. Language teachers should exercise caution in designing and using tests of listening. We suggest that test scores from WLP test formats may not constitute a highly trustworthy indication of students' listening ability. Therefore, teachers may consider combining WLP and PLP formats to partial out test format effects.

Supplementary Material 1

Gaze Behaviors	Effects	Inferential Statistics	Post-Hoc Inferential Statistics (Bonferroni Corrections, α=0.017)		E-L1	E-L2	PreL	WL	PostL	E-L1(PreL)	E-L1(WL)	E-L1(PostL)	E-L2(PreL)	E-L2(WL)	E-L2(PostL)
Normalized Total Visit Duration (%)	**L1**	*F*(1,389)=2.91, *p*=.089	-		65±1	66±1									
	SLC	*F*(2,389)=0.68, *p*=.51	-				65±1	66±1	65±2						
	L1xSLC	*F*(2,389)=1.06, *p*=.35	-							64±1	66±1	65±2	68±2	65±2	65±3
Normalized Total Fixation Duration (%)	**L1**	*F*(1,389)=5.63, *p*=.018	-		54±1	56±1									
	SLC	*F*(2,389)=0.09, *p*=.91	-				55±1	55±1	54±1						
	L1xSLC	*F*(2,389)=1.06, *p*=.35	-							53±1	55±1	54±1	59±1	55±2	54±3
Normalized Average Visit Duration (%)	**L1**	*F*(1,389)=10.02, *p*=.002	-		1.33±0.10	1.24±0.12									
	SLC	*F*(2,389)=426.26, *p*<.0001	-				1.01±0.03	0.38±0.01	2.52±0.19						
	L1xSLC	*F*(2,389)=17.85, *p*<.0001	**E-L1:** χ^2(2,N=86)=150.58, *p*<.0001 **E-L1 vs. E-L2 (PreL):** *Z*=-3.19, *p*=.0014 **E-L1 vs. E-L2 (WL):** *Z*=-3.78, *p*<.001 **E-L1 vs. E-L2 (PostL):** *Z*=-2.12, *p*=.034 **PreL vs. WL:** *Z*=-7.07, *p*<.0001 **WL vs. PostL:** *Z*=-8.05, *p*<.0001 **PreL vs. PostL:** Z=-8.05, *p*<.0001	**E-L2:** χ^2(2,N=45)=60.40, *p*<.0001 **E-L1 vs. E-L2 (PreL):** *Z*=-7.07, *p*<.0001 **E-L1 vs. E-L2 (WL):** *Z*=-8.05, *p*<.0001 **E-L1 vs. E-L2 (PostL):** *Z*=-8.05, *p*<.0001 **PreL vs. WL:** *Z*=-5.91, *p*<.0001 **WL vs. PostL:** *Z*=-2.93, *p*=.003 **PreL vs. PostL:** *Z*=-5.75, *p*<.0001						0.94±0.04	0.35±0.01	2.71±0.23	1.16±0.06	0.43±0.02	2.14±0.31
Normalized Average Fixation Duration (%)	**L1**	*F*(1,389)=95.66, *p*<.0001	-		0.222±0.021	0.168±0.010									
	SLC	*F*(2,389)=363.25, *p*<.0001	-				0.159±0.002	0.074±0.001	0.379±0.039						
	L1xSLC	*F*(2,389)=41.35, *p*<.0001	**E-L1:** χ^2(2,N=86)=150.58, *p*<.0001 **E-L1 vs. E-L2 (PreL)**: Z=-6.00, *p*<.0001 **E-L1 vs. E-L2 (WL):** *Z*=-3.19, *p*<.0001 **E-L1 vs. E-L2 (PostL):** *Z*=-2.91, *p*=.004 **PreL vs. WL:** *Z*=-7.07, *p*<.0001 **WL vs. PostL:** *Z*=-8.05, *p*<.0001 **PreL vs. PostL:** *Z*=-8.05, *p*<.0001	**E-L2:** χ^2(2,N=45)=60.40, *p*<.0001 **E-L1 vs. E-L2 (PreL):** *Z*=-8.06, *p*<.0001 **E-L1 vs. E-L2 (WL):** *Z*=-8.05, p<.0001 **E-L1 vs. E-L2 (PostL):** *Z*=-7.20, *p*<.0001 **PreL vs. WL:** *Z*=-5.91, *p*<.0001 **WL vs. PostL:** *Z*=-2.93, *p*=.003 **PreL vs. PostL:** *Z*=-5.75, *p*<.0001						0.152±0.002	0.071±0.001	0.444±0.056	0.171±0.002	0.080±0.001	0.254±0.023
Normalized Visit Counts (visits/min)	**L1**	*F*(1,390)=47.33, *p*<.0001			30±1	25±1									
	SLC	*F*(2,390)=41.37, *p*<.0001	**PreL vs WL:** *p*<.0001 **WL vs PostL:** *p*=.008 **PreL vs PostL:** *Z*=-8.05, *p*=.003				27±1	33±1	24±1						
	L1xSLC	*F*(2,390)=1.23, *p*=.30	-							29±1	35±1	27±1	25±1	29±1	20±1
Normalized Fixation Counts (fixations/min)	**L1**	*F*(1,390)=9.11, *p*=.003	-		137±2	127±3									
	SLC	*F*(2,390)=0.90, *p*=.41	-				137±2	133±2	131±3						
	L1xSLC	*F*(2,390)=1.07, *p*=.35	-							138±3	138±3	136±4	135±3	124±4	122±7

Note: The data above are presented in mean ± standard error. Abbreviations: L1=First language, PostL=Post-listening, PreL=Pre-listening, SLC=Stages of listening comprehension, and WL=While-listening,

Abbreviations

AOI: Area of interest; ART: Aligned Rank Transform; CAEL: Canadian Academic English Language; CE: Computer Edition; CI: Confidence intervals; CIV: Construct-irrelevant variance; E-L1: Native English-speaking; E-L2: Non-native English-speaking; fNIRS: Functional near-infrared spectroscopy; IELTS: International English Language Testing System; L2: Second language; PLP: Post-listening performance; SLC: Stages of listening comprehension; WLP: While-listening performance.

Authors' contributions

The study was conceived and conducted by Vahid Aryadoust and his research team. Vahid Aryadoust and Stacy Foo conducted the data analysis and wrote the report.

Authors' information

Vahid Aryadoust is Associate Professor of language assessment at the National Institute of Education of Nanyang Technological University, Singapore. His areas of interest include language assessment, quantitative methods, eye tracking, and brain imaging in language assessment. Vahid has published his research in *Computer Assisted Language Learning*, *Language Testing*, *System*, *Current Psychology*, *Language Assessment Quarterly*, *Assessing Writing*, *Educational Assessment*, *Educational Psychology*, etc. He has also (co)authored a number of book chapters and books that are published by Routledge, Cambridge University Press, Springer, Cambridge Scholar Publishing, Wiley Blackwell, etc. Vahid has also led a number of assessment research projects supported by educational fund-providers in Singapore, USA, UK, and Canada. He is a member of the Advisory Board of multiple international journals and has been awarded the Intercontinental Academia Fellowship (2018–2019). His YouTube channel, *Statistics and Theory*, has been awarded the John Cheung Social Media Award, 2020, which recognizes exemplary and innovative use of social media. The channel is available from: https://www.youtube.com/user/vahidaryadoust .

Stacy Foo is a research administrator at National Cancer Centre Singapore. She was previously a research assistant at the National Institute of Education of Nanyang Technological University. Her research background includes eye tracking, visual perception, and language assessment. She is interested in the connection between eye movements and brain activation.

Funding

This work was supported by Paragon Testing Enterprises and the National Institute of Education, Nanyang Technological University, Singapore (RI). The abovementioned funding bodies did not play any roles in influencing the study design, report writing, and the decision to submit this article for publication.

Availability of data and materials

The data will be made available upon request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 24 February 2022 Accepted: 5 August 2022
Published: 28 February 2023 (Online First: 6 October 2022)

References

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. https://doi.org/10.1080/10904018.2012.639649

Aryadoust, V. (2020). 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*, *33*(5-6), 510–537. https://doi.org/10.1080/09588221.2019.1574267

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

Aryadoust, V., Ng, L. Y., Foo, S., & Esposito, G. (2020). A neurocognitive investigation of test methods and gender effects in listening assessment. *Computer Assisted Language Learning*, *35*(4), 743–763. https://doi.org/10.1080/09588221.2020.1744667

Babayiğit, S. (2012). The role of oral language skills in reading and listening comprehension of text: A comparison of monolingual (L1) and bilingual (L2) speakers of English language. *Journal of Research in Reading*, *37*(S1), 1–26. https://doi.org/10.1111/j.1467-9817.2012.01538.x

Badger, R., & Yan, X. (2009). The use of tactics and strategies by Chinese students in the Listening component of IELTS. In P. Thompson (Ed.), *International English Language Testing System (IELTS) Research Reports 2009: Volume 9* (pp. 67–98). British Council and IELTS Australia. https://search.informit.org/doi/10.3316/informit.070356543696560

Baker, J. P., Goodboy, A. K., Bowman, N. D., & Wright, A. A. (2018). Does teaching with PowerPoint increase students' learning? A meta-analysis. *Computers & Education*, *126*, 376–387. https://doi.org/10.1016/j.compedu.2018.08.003

Bedford, O., & Chua, S. H. (2017). Everything also I want: An exploratory study of Singaporean *Kiasuism* (fear of losing out). *Culture & Psychology*, *24*(4), 491–511. https://doi.org/10.1177/1354067x17693831

Bolton, K. (2006). World Englishes today. In B. B. Kachru, Y. Kachru & C. L. Nelson (Eds.), *The Handbook of World Englishes* (pp. 240–269). Wiley-Blackwell. https://doi.org/10.1002/9780470757598.ch15

Buck, G. (2001). *Assessing Listening (Cambridge Language Assessment)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Burgoyne, K., Kelly née Hutchinson, J. M., Whiteley, H. E., & Spooner, A. (2009). The comprehension skills of children learning English as an additional language. *British Journal of Educational Psychology*, *79*(4), 735–747. https://doi.org/10.1348/000709909x422530

Charles, M., & Pecorari, D. (2016). *Introducing English for academic purposes*. Routledge.

Conklin, K., Alotaibi, S., Pellicer-Sánchez, A., & Vilkaitė-Lozdienė, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research*, *36*(3), 257–276. https://doi.org/10.1177/0267658320921496

Conklin, K., Pellicer-Sanchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Conrad, L. (1985). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition*, *7*(1), 59–69. https://doi.org/10.1017/s0272263100005155

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555

Davies, A. (2009). Assessing World Englishes. *Annual Review of Applied Linguistics*, *29*, 80–89. https://doi.org/10.1017/S0267190509090072

Dunkel, P., Mishra, S., & Berliner, D. (1989). Effects of note taking, memory, and language proficiency on lecture learning for native and nonnative speakers of English. *TESOL Quarterly*, *23*(3), 543–549. https://doi.org/10.2307/3586929

Field, J. (2009). The cognitive validity of the lecture-based question in the IELTS listening paper. In P. Thompson (Ed.), *International English Language Testing System (IELTS) Research Reports 2009: Volume 9* (pp. 17–65). British Council and IELTS Australia. https://search.informit.org/doi/10.3316/informit.070337910725302

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes*, *10*(2), 102–112. https://doi.org/10.1016/j.jeap.2011.04.002

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge University Press.

Field, J. (2015). *The effects of single and double play upon listening test outcomes and cognitive processing*. British Council.

Friedman, S. J., & Ansley, T. N. (1990). The influence of reading on listening test scores. *The Journal of Experimental Education*, *58*(4), 301–310. https://doi.org/10.1080/00220973.1990.10806544

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Graham, S. (2011). Self-efficacy and academic listening. *Journal of English for Academic Purposes*, *10*(2), 113–117. https://doi.org/10.1016/j.jeap.2011.04.001

Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*(2), 163–180. https://doi.org/10.1177/0265532211421161

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*(8), 180502. https://doi.org/10.1098/rsos.180502

Ho, J. T. S., Ang, C. E., Loh, J., & Ng, I. (1998). A preliminary study of *kiasu* behaviour - Is it unique to Singapore? *Journal of Managerial Psychology*, *13*(5/6), 359–370. https://doi.org/10.1108/02683949810220015

Holliday, A. (2005). *The struggle to teach English as an international language*. Oxford University Press.

Holliday, A. (2006). Native-speakerism. *ELT Journal*, *60*(4), 385–387. https://doi.org/10.1093/elt/ccl030

Holzknecht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., & Spöttl, C. (2017). *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test*. British Council. https://doi.org/10.13140/RG.2.2.21966.36166

Hulstijn, J. H. (2003). Connectionist models of Language Processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning*, *16*(5), 413–425. https://doi.org/10.1076/call.16.5.413.29488

Hutchinson, J. M., Whiteley, H. E., Smith, C. D., & Connors, L. (2003). The developmental progression of comprehension-related skills in children learning EAL. *Journal of Research in Reading*, *26*(1), 19–32. https://doi.org/10.1111/1467-9817.261003

Imhof, M. (2010). What is going on in the mind of a listener? The cognitive psychology of listening. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 97–126). Wiley-Blackwell. https://doi.org/10.1002/9781444314908

Juhola, M. (1991). Median filtering is appropriate to signals of saccadic eye movements. *Computers in Biology and Medicine*, *21*(1-2), 43–49. https://doi.org/10.1016/0010-4825(91)90034-7

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295x.87.4.329

Just, M. A., & Carpenter, P. A. (1984). Using eye fixations to study reading comprehension. In D. E. Kieras & M. A. Just (Eds.), *New Methods in Reading Comprehension Research* (pp. 151–182). Routledge. https://doi.org/10.4324/9780429505379-8

Kim, J. (2019). The effects of note-taking strategy training on students' notes during academic English listening tests. *English Teaching*, *74*(1), 25–48. https://doi.org/10.15858/engtea.74.1.201903.25

Kirsh, D. (2005). Metacognition, distributed cognition and visual design. In P. Gärdenfors & P. Johansson (Eds.), *Cognition, education, and communication technology* (pp. 147–180). L. Erlbaum Associates, Publishers.

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645. https://doi.org/10.1109/tbme.2010.2057429

Kruger, J.-L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled Academic lecture: L1 vs. L2. *Journal of Eye Movement Research*, *7*(5), 1–15. https://doi.org/10.16910/jemr.7.5.4

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterl*y, *36*(2), 173–190. https://doi.org/10.2307/3588329

Malone, M. E. (2010). Test review: Canadian academic English language (CAEL) assessment. *Language Testing*, *27*(4), 631–636. https://doi.org/10.1177/0265532210384265

Marx, A., Heppt, B., & Henschel, S. (2017). Listening comprehension of academic and everyday language in first language and second language students. *Applied Psycholinguistics*, *38*(3), 571–600. https://doi.org/10.1017/s0142716416000333

Matsumoto, K. (1993). Verbal-report data and introspective methods in Second language research: State of the art. *RELC Journal*, *24*(1), 32–60. https://doi.org/10.1177/003368829302400103

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707–726. https://doi.org/10.1177/001316446502500304

Moon, J. A., Keehner, M., & Katz, I. R. (2019). Affordances of item formats and their effects on test-taker cognition under uncertainty. *Educational Measurement: Issues and Practice*, *38*(1), 54–62. https://doi.org/10.1111/emip.12229

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Olsen, A., & Matos, R. (2012). Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In S. N. Spencer (Ed.), *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12* (pp. 317–320). https://doi.org/10.1145/2168556.2168625

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods*, *50*(4), 1645–1656. https://doi.org/10.3758/s13428-017-0998-z

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on School Leaving Examinations. *Applied Measurement in Education*, *4*(2), 159–183. https://doi.org/10.1207/s15324818ame0402_5

Rost, M. (2014). Listening in a multilingual world: The challenges of Second language (L2) listening. *International Journal of Listening*, *28*(3), 131–148. https://doi.org/10.1080/10904018.2014.937895

Rost, M. (2016). *Teaching and researching listening*. Routledge.

Sasaki, M. (2013). Introspective methods. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1340–1357). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla076

Song, M.-Y. (2011). Note-taking quality and performance on an L2 academic listening test. *Language Testing, 29*(1), 67–89. https://doi.org/10.1177/0265532211415379

Stuart, S., Hickey, A., Vitorio, R., Welman, K., Foo, S., Keen, D., & Godfrey, A. (2019). Eye-tracker algorithms to detect saccades during static and dynamic tasks: A structured review. *Physiological Measurement, 40*(2). https://doi.org/10.1088/1361-6579/ab02ab

Suvurov, R. (2018). *Investigating test-taking strategies during the completion of computer-delivered items from the Michigan English Test (MET): Evidence from eye tracking and cured retrospective reporting (Cambridge Michigan*

Language Assessment (CaMLA) Working Papers, Issue. M. L. Assessment. https://michiganassessment.org/wp-content/uploads/2019/03/CWP-2018-02.pdf

Tobii AB. (2016). *Tobii Studio User's Manual (Version 3.4.5)*.

Tobii AB. (2017). *Tobii Pro Studio (Version 3.4.8)*.

Vandergrift, L. (1999). Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal*, *53*(3), 168–176. https://doi.org/10.1093/elt/53.3.168

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*(1), 3–25. https://doi.org/10.1017/s0267190504000017

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191–210. https://doi.org/10.1017/s0261444807004338

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action.* Routledge.

Wang, H. (2018). *Testing lecture comprehension through listening-to-summarize cloze tasks: The trio of task demands, cognitive process and language competence*. Springer Nature. https://doi.org/10.1007/978-981-10-6202-5

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave McMillan.

Winke, P., & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports Online Series*, *3*(1), 1–30. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2014-3.ashx

Wobbrock, J. O. (2011). *ARTool align-and-rank data for a nonparametric ANOVA*. http://depts.washington.edu/acelab/proj/art/index.html

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems* (pp. 143–146). Association for Computing Machinery. https://doi.org/10.1145/1978942.1978963

Yang, H. (2013). The case for being automatic: Introducing the Automatic Linear Modeling (LINEAR) procedure in SPSS Statistics. *General Linear Model Journal*, *39*(2), 27–37.

Yildiz, N., Parjanadze, N., & Albay, M. (2015). The effect of question position on listening comprehension: A case study. *International Journal of Social Sciences & Educational Studies*, *2*(1), 4–9.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, *21*(2), 179–217. https://doi.org/10.1207/s15516709cog2102_3

Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.