BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

# INTERACTING WITH VISUALS IN L2 LISTENING TESTS: AN EYE-TRACKING STUDY

AR-A/2015/1

**Ruslan Suvorov, University of Hawai'i at Mānoa**

**ARAGs RESEARCH REPORTS ONLINE**
**SERIES EDITOR: VIVIEN BERRY**
**ISSN 2057-5203**

# ABSTRACT

**This research project investigated how visuals affect second language learners' listening comprehension and listening test performance. The use of a remote eye-tracking system enabled the researcher to conduct an in-depth examination of the language learners' use of visual information during a video-enhanced academic listening test.**

Visual information plays an important role in second language (L2) listening comprehension (Field 2008; Rost 2011), yet visuals have seen limited use in L2 listening assessment. The limited use of visuals in listening tests can be attributed to the lack of solid empirical evidence about how visuals are viewed during such tests and what impact they have on test performance. To address this gap, this study employed eye-tracking technology to investigate the extent to which L2 learners view two types of visuals—context and content videos—during a Video-based L2 Academic Listening Test (VALT), how the learners perceive and use the two video types, and what effect these visuals have on their test performance.

This mixed-methods study investigated:

- differences between scores on the subtests associated with different video types and between scores on the video and audio-only versions of the test;
- learners' viewing patterns with regard to context and content videos; and
- learners' use of visual information when watching the two types of videos and when answering individual test questions.

Test performance data, eye-tracking data, and retrospective verbal data were collected and analysed in the study.

Results demonstrated that, although visuals had no effect on L2 learners' test scores, the use of eye-tracking technology was instrumental in detecting the different effects of context and content videos. Moreover, the results revealed differences between context and content videos in terms of their perceived use during the test-taking process and their perceived helpfulness for answering questions on the VALT.

# Author

**Ruslan Suvorov**
Ruslan Suvorov holds a Ph.D. degree in Applied Linguistics and Technology with a minor in Curriculum and Instructional Technology from Iowa State University, USA. He currently works as Language Technology Specialist at the Center for Language and Technology, University of Hawai'i at Mānoa, USA. Ruslan's research interests include computer-assisted language learning and language testing, instructional technology and design, project-based language learning, language program evaluation, second language listening, multimodality and eye-tracking. He has presented at various regional, national and international conferences, and published in *CALICO Journal, Canadian Journal of Applied Linguistics, Language Testing, University of Cambridge ESOL Examinations Research Notes* and *TESL-EJ,* as well as conference proceedings. He has co-authored an entry to *The Encyclopedia of Applied Linguistics* and a book chapter for *The Companion to Language Assessment.*

# CONTENTS



## LIST OF TABLES



## LIST OF FIGURES

# 1. INTRODUCTION

## 1.1 Theoretical background

With the rapid pace of globalization, international standardized language tests—such as IELTS, TOEFL iBT, and PTE Academic—have high stakes for millions of prospective students and professionals from all over the world. To meet the growing demands of test-takers, many leading language-testing companies have moved from paper-and-pencil tests to computer-assisted language testing (CALT) and adopted technology for more efficient test administration (Suvorov & Hegelheimer 2013).

One of the main advantages of CALT, as argued by many scholars (e.g. Douglas & Hegelheimer 2007; Jamieson 2005), is its potential for multimedia input, which is believed to result in a greater level of authenticity of test tasks and to create testing conditions that closely resemble situations from the target language use domain. Considering that visual information is an indispensable component of multimedia (Mayer 2009), the use of visuals in CALT has generated significant interest among language assessment specialists. Of particular interest for CALT is the use of visuals for assessing second language (L2) listening, a widely used skill that is indispensible for effective communication and overall language proficiency (Ockey 2009).

Although visuals are believed to play an important role in L2 listening comprehension (Anderson & Lynch 1988; Field 2008; Rost 2011), they have seen limited use in L2 listening tests for two main reasons. First, there is a lack of agreement among researchers about what construct—or ability—visually enhanced L2 listening tests should assess (Alderson & Banerjee 2002; Buck 2001; Ockey 2007; Taylor & Geranpayeh 2011). On one hand, some language testing experts (e.g. Ockey 2007; Wagner 2007, 2008) contend that a construct measured by media-enhanced L2 listening tests should include the ability to understand both the verbal and the visual information because in most real-life situations oral information is accompanied by visual information. On the other hand, the opponents of including visuals in L2 listening tests argue that the ability to utilize information from visuals should not be part of the listening construct because "we are usually interested in the test-takers' language ability, rather than the ability to understand subtle visual information" (Buck 2001, p. 172).

Second, visuals are not widely used in L2 listening assessment due to inconclusive research on the effect of images and videos on L2 learners' performance on media-enhanced L2 listening tests. Specifically, these studies showed that while in some cases the use of visuals helped L2 learners perform better on L2 listening tests (Ginther 2002; Wagner 2010b), in other cases, visuals had a detrimental effect (Suvorov 2009) or no effect on the participants' performance (Coniam 2001; Gruba 1993).

These inconclusive findings can be partially attributed to the types of visuals used in L2 listening tests. Traditionally, researchers have differentiated between *context* visuals and *content* visuals (Bejar et al. 2000; Ginther 2002). *Context* visuals are those that provide visual information about the environment in which communication takes place, whereas *content* visuals contain visual information that is related to the verbally delivered information. Interestingly, researchers rarely specify whether their listening tests include context or content visuals. The review of literature, however, allows for the conclusion that most of the existing studies seem to have used context visuals, whereas content visuals have not been implemented in research much.

Another factor that might have led to mixed results is related to research designs used in the studies. In particular, most of the existing studies that investigated how visuals affect L2 listeners' test performance entailed the comparison of test-takers' scores on media-based L2 listening tests with their scores on the audio-only versions of the same tests (e.g. Coniam 2001; Gruba 1993; Suvorov 2009; Wagner 2010b). Such research was based on the assumption that a statistically significant difference between L2 test-takers' scores on a visually enhanced L2 listening test and their scores on an audio-only listening test could be attributed to the effect of visuals. The main problem with this assumption, however, is that it does not take into account L2 test-takers' viewing behaviour: Since the test-takers are not forced to watch a screen during visually enhanced listening tests, they vary in the extent to which they use visual information, with some of them not looking at the visuals at all (Wagner 2007). If those L2 test-takers who do not watch the visuals—or watch them to a small extent—obtain different scores on the two tests, the difference in their scores cannot be attributed to the effect of visuals.

Surprisingly, researchers have mostly ignored the viewing behaviour of L2 learners during listening tests accompanied by visuals. The only exceptions are the studies done by Ockey (2007) and Wagner (2007, 2010a), in which the researchers used a video camera to record their participants during a visually enhanced L2 listening test, and then measured the amount of time the participants made eye contact with the screen while taking the test. While the use of video recordings can be useful for learning about L2 test-takers' interaction with visuals, this type of data can generally yield information about *how long* the test-takers look at the screen, but not *what exactly* they look at, *how long* they focus on certain elements of the visual, or *why* they look at them. Specialized technology such as an eye-tracking system, however, can provide much more precise data (Duchowski 2007; Holmqvist et al. 2011) that include detailed information about test-takers' eye movements during visually enhanced L2 listening assessment.

## 1.2 Purpose of the study

Taking into account (a) the inconclusive results of existing studies that have analysed the effect of visuals on L2 learners' test performance, (b) the lack of research comparing the effects of context and content videos on L2 listening test performance, and (c) a surprising dearth of research examining the actual viewing behaviour of L2 learners during visually enhanced L2 listening tests, the overall purpose of this study was to address these gaps.

Specifically, the study had three main objectives:

1. to determine whether context videos and content videos had a differential effect on L2 learners' performance on a video-based L2 academic listening test
2. to investigate L2 learners' viewing behaviour during a visually enhanced L2 listening assessment
3. to explore how L2 listeners use visual information from context and content videos during the test.

# 2. METHODOLOGY

## 2.1 Study design

The design of this study was based on Creswell and Plano Clark's (2007) data transformation model of the triangulation design that involved the concurrent collection of quantitative data sets (i.e. test performance data and eye-tracking data), followed immediately by the collection of qualitative data (i.e. retrospective verbal data) that were subsequently quantified. This model enabled the researcher to use inferential statistics for analysing the data and generalize the results to a larger population.

## 2.2 Materials

To collect the data, the researcher developed a Video-based Academic Listening Test (VALT) and its audio-only version, Audio-based Academic Listening Test (AALT) using the Quiz module in the Moodle course management system. The 45-minute VALT consisted of six short academic video lectures (i.e. three context and three content videos) and 30 multiple-choice questions, with the AALT being the same except for the lectures being presented in an audio-only format. Table 1 outlines the structure of the test.

*Table 1: Structure of the Video-based Academic Listening Test (VALT)*

| Stimulus | Items | Visual type | Discipline | Topic | Stimulus length, sec. |
|---|---|---|---|---|---|
| **Video 1** | 1-5 | Context video | Psychology | Neurons | 188 |
| **Video 2** | 6-10 | Content video | Astrophysics | Exoplanets | 196 |
| **Video 3** | 11-15 | Context video | Political Science | Enlightenment | 212 |
| **Video 4** | 16-20 | Content video | Economics | Rent control | 226 |
| **Video 5** | 21-25 | Context video | Philosophy | P-functions | 223 |
| **Video 6** | 26-30 | Content video | Biology | Mushrooms | 234 |

Decisions as to whether a specific video clip was context or content were made based on the definitions of context and content visuals provided by Bejar et al. (2000) and Ginther (2002). Each context video displayed a professor giving a lecture in a classroom, thus providing visual information about the context and the speaker (see Figure 1). Content videos selected for the VALT utilized different forms of content visuals, such as an image of a star with an exoplanet (Astrophysics, see Figure 2); a graph representing the interaction among demand, supply, and price (Economics); and a drawing of the mushroom structure on a blackboard (Biology).

*Figure 1: A screenshot of a context video from the VALT*


*Figure 2: A screenshot of a content video from the VALT*


Both the VALT and the AALT were piloted and revised several times before being used to collect the data for the main study. Detailed information about test specifications, test development and validation can be found in Suvorov (2013).

## 2.3 Eye-tracking equipment and software

A remote eye-tracking system EyeTech Vision Tracker 2 (0.5° accuracy, 80 fps data sampling rate, 65-100 cm operating range, 1680 × 1050 display) was employed to collect eye-tracking data. The eye-tracker was physically connected to a computer display and run on an iMac station (27 inches, 3.7 GHz) using Windows 7 64-bit OS. The display was also equipped with a web camera Logitech Webcam Pro 9000. In addition, the second display was used by the researcher to monitor the data collection process. The eye-tracking data were recorded and processed using Attention Tool Usability Module (version 4.8), which is an eye-tracking software application for market research, scientific research, and website usability developed by iMotions. Dynamic Media Module, which is an add-on module in Attention Tool for analyzing dynamic media such as videos, was used for the subsequent analysis of eye-tracking data.

## 2.4 Data collection

Data collection took place at a large public university in the Midwest of the USA, and involved 121 participants who were non-native English-speaking students with different levels of English language proficiency. Test performance data comprised the scores on the Video-based Academic Listening Test (VALT) and its audio-only version (AALT) that were developed for this study. Test performance data were collected from all study participants (n=121), with 75 participants taking the VALT and 46 participants taking the AALT.

Eye-tracking data were collected using a remote eye-tracking system and consisted of the eye-movement recordings of 33 participants while they were taking the VALT. These recordings were used to calculate three eye-tracking measures—namely, fixation rate, dwell rate, and total dwell time—that represented the viewing behaviour of L2 test-takers when they were watching context and content videos during the VALT.

Finally, retrospective verbal data were gathered using cued retrospective reporting (Van Gog et al. 2005), which is a method for collecting retrospective verbalizations by showing participants the recordings of their eye movements and asking them to verbalize their cognitive processes that occurred during the initial visual examination of the stimulus. Specifically, the 33 participants who participated in the eye-tracking experiment were shown the recordings of their eye movements and asked to share their perceptions regarding their use of different aspects of visual information while they were completing the VALT. Their answers were used for investigating how L2 learners use visual information when watching context and content videos and answering the questions on the test.

## 2.5 Research questions

The three data sets were used to answer the following research questions.

**Research Question 1:** To what extent are the statistical properties of the scores on the VALT and on the AALT appropriate for making norm-referenced decisions?

**Research Question 2:** To what extent do L2 test-takers perform differently on the subtest enhanced by context videos versus the subtest enhanced by content videos in the VALT and in the AALT? To what extent do L2 test-takers perform differently on the VALT versus the AALT?

**Research Question 3:** To what extent do L2 test-takers watch context videos differently from content videos in the VALT, as indicated by eye-tracking measures? To what extent do L2 test-takers' viewing patterns, as indicated by eye-tracking measures, correlate with their scores on the subtest enhanced by context videos and on the subtest enhanced by content videos?

**Research Question 4:** How do L2 test-takers use visual information when watching context and content videos in the VALT, as indicated by cued retrospective reporting? In particular:

- **Research Question 4.1:** What aspects of visual information, and why, do L2 test-takers focus on when watching context and content videos in the VALT?
- **Research Question 4.2:** What aspects of visual information in the VALT, and why, do L2 test-takers find helpful and/or distracting?

**Research Question 5:** How do L2 test-takers use visual information when answering individual questions on the VALT, as indicated by cued retrospective reporting? In particular:

- **Research Question 5.1:** What is the difference between L2 test-takers' perceptions of the helpfulness of visual information for answering questions on the context subtest vs. questions on the content subtest of the VALT?
- **Research Question 5.2:** To what extent are L2 test-takers' perceptions of the helpfulness of visual information for answering each individual question associated with their scores on individual questions?

## 2.6 Data analysis

Table 2 summarizes the types of data analyses that were carried out to answer each question.

*Table 2: Types of data analyses used for answering research questions*

| Research question | Data | Analysis |
|---|---|---|
| **RQ1:** Statistical properties of scores | Item scores, context subtest scores and content subtest scores, and overall test scores on the VALT (n=75) and the AALT (n=46) | Descriptive statistics, reliability analysis, item analysis, and distractor analysis |
| **RQ2:** Difference between context subtest scores and content subtest scores within the VALT and the AALT; difference between overall VALT scores and overall AALT scores | Context subtest scores and content subtest scores on the VALT (n=75) and the AALT (n=46), overall scores on the VALT (n=75) and on the AALT (n=46) | Two paired-samples *t* tests comparing subtest scores within the VALT and within the AALT, an independent-samples *t* test comparing overall VALT scores and overall AALT scores |
| **RQ3:** Difference between patterns of viewing context videos and patterns of viewing content videos; correlation between viewing patterns and context/content subtest scores | Eye-tracking data (n=33), context subtest scores and content subtest scores on the VALT (n=33) | Descriptive statistics for eye-tracking measures, three paired-samples *t* tests comparing eye-tracking measures for content and context videos, Pearson product-moment correlation coefficient |
| **RQ4:** Use of visual information when watching context and content videos | Retrospective verbal data (n=33) | Transcribing cued retrospective reports, coding for emergent themes, and counting instances of the themes and the number of participants who commented on each theme |
| **RQ5:** Difference between perceptions of the helpfulness of visual information for answering questions on the context subtest vs. questions on the content subtest of the VALT; association between perceptions of the helpfulness of visuals and item scores | Retrospective verbal data (n=33), item scores (n=33) | Quantification of perceptions regarding the helpfulness of visual information; paired-samples *t* test comparing perceptions of the helpfulness of visuals for answering questions on the context subtest vs. questions on the content subtest; Cochran-Mantel-Haenszel chi-square statistic |

# 3. RESULTS

## 3.1 Research question 1

The first research question examined the extent to which the quality of test items created a test that was appropriate for making norm-referenced decisions. The results of four types of analyses that were conducted using the performance data indicated that the statistical properties of the VALT scores and the AALT scores were overall appropriate for making norm-referenced decisions regarding the test-takers' L2 listening ability. Specifically, the results of descriptive statistics showed that the distribution of scores was relatively normal (see Table 3).

*Table 3: Descriptive statistics for overall scores and subtest scores on the VALT (n=75) and the AALT (n=46)*

| Type of scores | | Number of test items | Mean | *SD* | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| | Overall VALT scores | 30 | 16.81 | 5.54 | .179 | -.426 |
| | Overall AALT scores | 30 | 16.65 | 5.29 | .335 | -.750 |
| VALT | Context subtest scores | 15 | 8.21 | 3.28 | .205 | -.654 |
| | Content subtest scores | 15 | 8.60 | 2.81 | .079 | -.406 |
| AALT | Context subtest scores | 15 | 8.50 | 2.90 | .421 | -.919 |
| | Content subtest scores | 15 | 8.15 | 2.79 | .132 | -.545 |

Next, the results of reliability analyses provided in Table 4 revealed that internal consistency reliability estimates of the overall scores on both tests were adequate: $\alpha$=.81 for the VALT and $\alpha$=.79 for the AALT.

*Table 4: Reliability analyses of the overall scores and subtest scores on the VALT (n=75) and the AALT (n=46)*

| Type of scores | | Number of test items | Cronbach's alpha ($\alpha$) | Standard error of measurement (SEM) |
|---|---|---|---|---|
| | Overall VALT scores | 30 | .81 | 2.39 |
| | Overall AALT scores | 30 | .79 | 2.39 |
| VALT | Context subtest scores | 15 | .72 | 1.71 |
| | Content subtest scores | 15 | .65 | 1.67 |
| AALT | Context subtest scores | 15 | .64 | 1.72 |
| | Content subtest scores | 15 | .63 | 1.68 |

Finally, item analyses (see Appendix A) and distractor analyses (see Appendix B for the VALT and Appendix C for the AALT) provided empirical evidence that items on both tests were overall of an appropriate level of difficulty for the target population and discriminated among test-takers' with different levels of the targeted L2 abilities.

## 3.2 Research question 2

The second research question addressed the difference between L2 test-takers' performance on the context subtest and their performance on the content subtest of the VALT (n=75), as well as the difference between the performance on the VALT (n=75) and that on the AALT (n=46). The results of the paired-samples *t* test that was carried out to compare the context subtest scores (*M*=8.21, *SD*=3.28) with the content subtest scores (*M*=8.60, *SD*=2.81) within the VALT revealed no statistically significant difference, *t*(74)=1.30, *p*=.20, indicating no variation between the effects of the two video types on L2 learners' test performance (see Table 5).

*Table 5: Results of a paired-samples T test comparing context subtest scores and content subtest scores on the VALT (n=75)*

| VALT scores | *M* | *SD* | *df* | *t* | *p* | Effect size (eta squared) |
|---|---|---|---|---|---|---|
| | | | 74 | 1.30 | .20 | .02 |
| Context subtest | 8.21 | 3.28 | | | | |
| Content subtest | 8.60 | 2.81 | | | | |

In addition, the results of the independent-samples *t* test showed no statistically significant difference between the overall scores on the VALT (*M*=16.81, *SD*=5.54) and the AALT (*M*=16.65, *SD*=5.29), *t*(98.8)=.160, *p*=.87, which implies that both types of videos in the VALT did not have any effect on L2 learners' test performance (see Table 6).

*Table 6: Results of an independent-samples T test comparing overall scores on the VALT (n=75) and overall scores on the AALT (n=46)*

| Overall scores | *M* | *SD* | *df* | *t* | *p* | Effect size (eta squared) |
|---|---|---|---|---|---|---|
| | | | 98.8 | .160 | .87 | .001 |
| VALT | 16.81 | 5.54 | | | | |
| AALT | 16.65 | 5.29 | | | | |

## 3.3 Research question 3

The third research question (a) investigated the viewing patterns of the L2 learners when they were watching context and content videos in the VALT (n=33) and (b) explored the relationship between the three eye-tracking measures and the scores on the two subtests of the VALT. The results of the three paired-samples *t* tests (shown in Table 7) that were carried out to compare each of the three eye-tracking measures (i.e., the fixation rate, the dwell rate, and the total dwell time) for context videos and for content videos demonstrated that L2 learners fixated their eyes on content videos (*M*=.87, *SD*=.42) more frequently than on context videos (*M*=.71, *SD*=.42), *t*(32)=4.73, *p*=.01, and spent statistically significantly more time watching content videos (*M*=57.99, *SD*=19.79) than context videos (*M*=50.70, *SD*=22.49), *t*(32)=5.02, *p*=.01. In contrast, no statistically significant difference was found between the L2 test-takers' dwell rates for context videos (*M*=29.07, *SD*=17.26) and for content videos (*M*=29.40, *SD*=15.49), *t*(32)=.38, *p*=71.

*Table 7: Results of three paired-samples T tests comparing three eye-tracking measures for context videos and content videos (n=33)*

| T test | Eye-tracking measures | *M* | *SD* | *df* | *t* | *p* | Effect size (eta squared) |
|---|---|---|---|---|---|---|---|
| 1 | **Fixation rate** | | | 32 | 4.73 | .01* | .41 |
| | Context videos | .71 | .40 | | | | |
| | Content videos | .87 | .42 | | | | |
| 2 | **Dwell rate** | | | 32 | .38 | .71 | .01 |
| | Context videos | 29.07 | 17.26 | | | | |
| | Content videos | 29.40 | 15.49 | | | | |
| 3 | **Total dwell time** | | | 32 | 5.02 | .01* | .44 |
| | Context videos | 50.70 | 22.49 | | | | |
| | Content videos | 57.99 | 19.79 | | | | |

The results of the correlation analysis illustrated in Table 8 revealed a weak relationship between the context subtest scores and the fixation rate for context videos ($r$=.32), which was not statistically significant at $p$=.07. An even weaker relationship was found between the context subtest scores and the total dwell time for context videos ($r$=.23), and it was also not statistically significant at $p$=.21. All other Pearson product-moment correlation coefficients were close to 0, demonstrating no relationship between the participants' viewing patterns and their scores on the two subtests within the VALT.

*Table 8: Correlations between three eye-tracking measures and VALT subtest scores (n=33)*

| Scores | Fixation rate | | Dwell rate | | Total dwell time | |
|---|---|---|---|---|---|---|
| | *r* | *p* | *r* | *p* | *r* | *p* |
| Context subtest | .32 | .07 | .04 | .81 | .23 | .21 |
| Content subtest | .02 | .93 | .15 | .41 | -.01 | .96 |

## 3.4 Research question 4

The focus of Research Question 4 was on the L2 test-takers' use of visual information when watching context and content videos in the VALT (n=33). This research question comprised two sub-questions. **Research Question 4.1** inquired into the aspects of visual information that the participants focused on when watching context and content videos, and their reasons for focusing on these aspects. The results of the qualitative analysis of retrospective verbal data revealed that the participants focused on two main types of aspects in videos: speaker-related aspects and lecture-related aspects. Specifically, when watching context videos, the participants focused primarily on speaker-related aspects that included the speaker's appearance (i.e., mouth, face, head, eyes, and hands, focused on by 88% of the total number of participants) and body movements and gestures (58%). Additionally, they focused on lecture-related aspects such as a contextual visual aid (e.g., a PowerPoint slide with a picture of John Locke, 40%) and some textual information presented as several key words (6%).

When watching content videos, however, the participants concentrated a lot of attention on both speaker-related aspects and lecture-related aspects. The lecture-related visual aspects in content videos comprised content-based visual aids (i.e., a picture of a star projected on the screen, a graph on a PowerPoint slide, and a drawing of a mushroom on the board, focused on by 97% of the total number of participants) and textual information (i.e., notes on the board and titles of the PowerPoint slides, 43%). The speaker-related aspects of content visuals included the speaker's appearance (i.e., mouth, face, and hands, 55%), movements and actions (e.g., body movements, gestures, and pointing to visual aids, 30%), and presentation of visual content (e.g., writing notes, showing a mushroom, and drawing the structure of a mushroom on the board, 52%).

With regard to the reasons for focusing on the visual aspects, the findings demonstrated that L2 test-takers focused on context videos mostly due to speaker-related reasons. The participants reported focusing on context videos because they had no visual information to look at other than the speaker (33%), they believed that seeing the speaker's mouth facilitated their comprehension of the lecture (18%), that seeing the speaker helped them focus (18%), and that the speaker's personality attracted their attention (21%). With respect to content videos, the results evinced one speaker-related reason (namely that the speaker was pointing to a visual aid) expressed by 18% of the total number of participants, and four lecture-related reasons explaining why the L2 learners focused on this video type. In terms of lecture-related reasons, the participants claimed that visual aids in content videos helped them comprehend the lecture (45%), facilitated their note-taking (9%), and were related to the speaker's talk (55%). Likewise, 9% of the participants focused on these videos because they found the topic of the lectures interesting.

**Research Question 4.2** investigated the aspects of visual information that the L2 learners found helpful and the aspects that they found distracting, as well as the reasons why they found these aspects to be helpful and/or distracting.

In context videos, the following three speaker-related aspects were considered helpful: the speaker's gestures (15%), the speaker's mouth (12%), and the speaker in general (18%). Regarding lecture-related aspects of context videos, 15% of the participants claimed that it was helpful to see a contextual visual aid (namely, a PowerPoint slide with a picture of John Locke) and 36% of the participants expressed similar remarks about seeing textual information (i.e., words on a PowerPoint slide). As far as content videos are concerned, all 33 participants unanimously reported that the most helpful aspect was content-based visual aids (e.g., an image of a star and a graph on a PowerPoint slide), although some participants also claimed to have benefited from seeing notes on the board (39%) and the speaker's gestures (18%).

Several reasons explain why L2 test-takers found these aspects of visual information helpful. For context videos, most reasons were speaker-related: The test-takers believed that seeing the speaker's mouth facilitated their comprehension of the lecture (9%), that seeing the speaker helped them focus (21%), and that the speaker's movements attracted their attention and facilitated their comprehension (21%). In addition, 15% of the test-takers reported that seeing textual information facilitated their comprehension. Regarding content videos, the results revealed one speaker-related and seven lecture-related reasons that the participants provided to explain why the specific aspects of visual information from this video type were helpful. The three most common reasons were that content-based visual aids facilitated L2 learners' comprehension of the lecture (97%), helped the participants answer the questions on the VALT (30%), and were related to the content of the lecture (52%).

In addition to helpful visual aspects, the results yielded from investigation of Research Question 4.2 also showed that some aspects of visual information in both types of video were distracting. In context videos, the speaker's movements were found by 73% of the participants to be the most distracting aspect, followed by contextual visuals aids (21%) and lights going out during one of the lectures (21%). In content videos, the only aspect that distracted 39% of the test-takers was content-based visual aids from Video 4 (namely, the floor plan of an apartment and the graph showing the relationship among the demand, the supply, and the price).

The results of the retrospective verbal data analysis evinced two reasons why context videos were distracting. The first reason was that the speaker's body movements distracted from listening and/or note-taking (reported by 58% of the total number of participants), whereas the second reason was related to the problems with interpreting contextual visual aids (9%). With regard to content videos, 30% of the participants deemed content-based visuals aids distracting due to the problems with their interpretation. Interestingly, it was also found that some aspects of visuals were regarded as both helpful and distracting.

## 3.5 Research question 5

The last research question aimed at investigating how the 33 participants used visual information from context and content videos when answering individual questions on the VALT. Specifically, **Research Question 5.1** focused on studying the difference between L2 test-takers' perceptions of the helpfulness of visual information for answering questions on the context subtest vs. questions on the content subtest of the VALT.

A paired-samples *t* test was utilized to compare the scores representing the helpfulness of visual information for answering questions on the context video subtest and the scores representing the helpfulness of visual information for answering questions on the content video subtest. The results of the paired-samples *t* test, which was used to compare these scores, indicated that the L2 learners perceived the visual information from content videos ($M$=5.58, $SD$=2.09) to be significantly more helpful than the visual information from context video ($M$=.76, $SD$=1.06) for answering questions on the two subtests of the VALT, $t(32)=12.66$, $p=.01$ (see Table 9).

Coupled with the findings for Research Question 4, these results suggest that unlike context videos, content videos that contain semantically rich visual information are perceived by L2 learners as helpful for answering questions on the listening test.

*Table 9: Difference between participants' perceptions of the helpfulness of visual information for answering questions on the context subtest vs. content subtest (n=33)*

| Subtest | *M* | *SD* | *df* | *t* | *p* | Effect size (eta squared) |
|---|---|---|---|---|---|---|
| | | | 32 | 12.66 | .01 | .83 |
| Context subtest | .76 | 1.06 | | | | |
| Content subtest | 5.58 | 2.09 | | | | |

Finally, **Research Question 5.2** explored the association between L2 test-takers' item scores on the VALT and their perceptions of the helpfulness of visual information from each individual video (i.e. both context and content videos) for answering each individual test item. The results of the Cochran-Mantel-Haenszel chi-square statistic showed a statistically significant positive association between the L2 test-takers' item scores and their perceptions regarding the helpfulness of visuals ($\chi^2$=13.72, $p$<.01), demonstrating that those test-takers who considered visual information from the videos to be helpful for answering individual questions on the test had a tendency to answer those questions correctly.

# 4. CONCLUSION

This research project introduces an innovative approach that employs eye-tracking technology for exploring L2 learners' interaction with visuals during video-mediated L2 academic listening assessment. By triangulating eye-tracking data with retrospective verbal data and test performance data, the study presents evidence about how, why and to what extent L2 learners use visual information from the videos in the test. It is also the first study that compares the effects of two types of videos—namely, context videos and content videos—on L2 learners' listening test performance.

The results of this study make an important contribution to the field of language testing and, in particular, to the body of research on the use of visuals in L2 listening assessment. Specifically, the results revealed differences in the way L2 learners viewed context and content videos while taking the VALT, even though in this study the differences in viewing did not result in a detectable difference in scores on the two subtests. In other words, the use of eye-tracking technology was essential for detecting the different effects of the content and context visuals.

The study also provides novel insights into L2 learners' emic perspectives regarding the aspects of visual information that they find helpful and the aspects that they find distracting, as well as the reasons why they find them helpful and/or distracting.

# 5. ACKNOWLEDGMENTS

I express my immense gratitude to my major professor, Dr. Carol Chapelle, for her expert guidance, invaluable support, and thought-provoking discussions that helped me conceptualize the ideas in this research study. I am also very grateful to all members of my dissertation committee who guided me throughout this project: Dr. Volker Hegelheimer, Dr. John Levis, Dr. Don Payne, Dr. Ana-Paula Correia, and Dr. Ann Smiley-Oyen. Special thanks go to Andrea Peer and Dr. Stephen Gilbert for providing me access to the eye-tracking equipment and User Experience lab at Iowa State University.

This dissertation research was supported by the Small Grant for Doctoral Research in Second Language Assessment from Educational Testing Service and the Assessment Research Award from the British Council.

# REFERENCES

Alderson, C. J. & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching,* vol. 35, pp. 79–113.

Anderson, A. & Lynch, T. (1988). *Listening.* Oxford, UK: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests.* Oxford, UK: Oxford University Press.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S. & Turner, J. (2000). *TOEFL 2000 Listening Framework: A working paper.* Princeton, NJ: Educational Testing Service.

Buck, G. (2001). *Assessing listening.* Cambridge, UK: Cambridge University Press.

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: a case study. *System,* vol. 29, pp. 1–14.

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

Douglas, D. & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics,* vol. 27, pp. 115–132.

Duchowski, A. (2007). *Eye tracking methodology: theory and practice,* 2nd edn, London, UK: Springer-Verlag.

Field, J. (2008). *Listening in the language classroom.* Cambridge, UK: Cambridge University Press.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing,* vol. 19, pp. 133–167.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal,* vol. 15, pp. 85–88.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. & Van de Weijer, J. (2011). *Eye tracking: a comprehensive guide to methods and measures.* Oxford, UK: Oxford University Press.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics,* vol. 25, pp. 228–242.

Kane, M. T. (2006). Validation. In R. Brennen (Ed.), *Educational measurement,* 4th edn. Westport, CT: Praeger.

Mayer, R. E. (2009). *Multimedia learning,* 2nd edn. Cambridge, UK: Cambridge University Press.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing,* vol. 24, pp. 517–537.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal,* vol. 93, pp. 836–847.

Rost, M. (2011). *Teaching and researching listening,* 2nd edn. Harlow, UK: Pearson.

Suvorov, R. (2009). Context visuals in L2 listening tests: the effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun & I. Katz (Eds.), *Developing and evaluating language learning materials.* Ames, IA: Iowa State University.

Suvorov, R. (2013). *Interacting with visuals in L2 listening tests: an eye-tracking study.* (Unpublished doctoral dissertation.) Ames, IA: Iowa State University.

Suvorov, R. & Hegelheimer, V. (2013). Computer-assisted language testing. In A. J. Kunnan (Ed.), *Companion to Language Assessment.* Malden, MA: Wiley-Blackwell.

Taylor, L. & Geranpayeh, A. (2011*).* Assessing listening for academic purposes: defining and operationalising the test construct. *Journal of English for Academic Purposes,* vol. 10, pp. 89–101.

Van Gog, T., Paas, F., Van Merriënboer, J. J. G. & Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied,* vol. 11, pp. 237–244.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology,* vol. 11, pp. 67–86.

Wagner, E. (2008). Video listening tests: what are they measuring? *Language Assessment Quarterly,* vol. 5, pp. 218–243.

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System,* vol. 38, pp. 280–291.

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing,* vol. 27, pp. 493–513.

# Appendix A:

## Results of item analyses for the VALT (n=75) and the AALT (n=46)

| Item | VALT | | AALT | |
|---|---|---|---|---|
| | IF | $r_{p\text{-}bis}$ | IF | $r_{p\text{-}bis}$ |
| 1 | .787 | .444 | .717 | .475 |
| 2 | .773 | .462 | .848* | .018** |
| 3 | .600 | .309 | .733 | .190** |
| 4 | .520 | .506 | .522 | .644 |
| 5 | .427 | .465 | .578 | .125** |
| 6 | .440 | .001** | .435 | .008** |
| 7 | .387 | .554 | .435 | .427 |
| 8 | .640 | .328 | .543 | .440 |
| 9 | .413 | .584 | .261 | .371 |
| 10 | .680 | .481 | .717 | .244** |
| 11 | .480 | .469 | .478 | .297** |
| 12 | .467 | .381 | .478 | .446 |
| 13 | .467 | .547 | .413 | .495 |
| 14 | .360 | .404 | .304 | .514 |
| 15 | .440 | .430 | .435 | .394 |
| 16 | .680 | .455 | .804* | .512 |
| 17 | .573 | .593 | .644 | .609 |
| 18 | .267 | .251** | .217 | .357 |
| 19 | .840* | .441 | .822* | .419 |
| 20 | .373 | .332 | .391 | .275** |
| 21 | .573 | .196** | .667 | .286** |
| 22 | .581 | .550 | .522 | .452 |
| 23 | .653 | .347 | .644 | .603 |
| 24 | .427 | .259** | .522 | .219** |
| 25 | .667 | .475 | .696 | .372 |
| 26 | .893* | .177** | .733 | .478 |
| 27 | .608 | .304 | .652 | .309 |
| 28 | .653 | .291** | .674 | .300 |
| 29 | .800 | .183** | .478 | .388 |
| 30 | .360 | .429 | .391 | .411 |
| Mean values | .56 | .39 | .56 | .37 |

*Note.* Test items with IF values above .80 are marked with an asterisk (*); test items with a discrimination index ($r_{p\text{-}bis}$) of less than .30 are marked with a double asterisk (**).

# Appendix B:

## Distractor analysis of item scores on the VALT (n=75)

| Item | Response frequencies | | | | Option point-biserials | | | |
|---|---|---|---|---|---|---|---|---|
| | % A | % B | % C | % D | A $r_{p\text{-}bis}$ | B $r_{p\text{-}bis}$ | C $r_{p\text{-}bis}$ | D $r_{p\text{-}bis}$ |
| 1 | **78.7** | 2.7 | 17.3 | 1.3 | **.444** | -.024 | -.426 | -.144 |
| 2 | 5.3 | **77.3** | 4.0 | 13.3 | -.078 | **.462** | -.327 | -.329 |
| 3 | 5.3 | **60.0** | 33.3 | 1.3 | -.164 | **.309** | -.248 | .025* |
| 4 | 25.3 | 12.0 | **52.0** | 10.7 | -.248 | -.286 | **.506** | -.169 |
| 5 | **42.7** | 6.7 | 22.7 | 28.0 | **.465** | -.020 | -.161 | -.351 |
| 6 | 5.3 | 42.7 | 8.0 | **44.0** | -.057 | .206* | -.329 | **.001** |
| 7 | 13.3 | **38.7** | 22.7 | 25.3 | -.293 | **.544** | -.178 | -.209 |
| 8 | 14.7 | 0.0 | **64.0** | 21.3 | -.383 | -* | **.328** | -.053 |
| 9 | 20.0 | **41.3** | 18.7 | 20.0 | -.389 | **.584** | -.046 | -.286 |
| 10 | **68.0** | 6.7 | 22.7 | 2.7 | **.481** | -.205 | -.381 | -.085 |
| 11 | 26.7 | **48.0** | 2.7 | 22.7 | -.336 | **.469** | -.024 | -.196 |
| 12 | 5.3 | 17.3 | **46.7** | 30.7 | -.089 | -.093 | **.381** | -.293 |
| 13 | 36.0 | 9.3 | 8.0 | **46.7** | -.318 | -.289 | -.133 | **.547** |
| 14 | **36.0** | 26.7 | 30.7 | 6.7 | **.404** | -.204 | -.203 | -.039 |
| 15 | 6.7 | **44.0** | 18.7 | 30.7 | -.020 | **.430** | -.133 | -.340 |
| 16 | **68.0** | 20.0 | 9.3 | 2.7 | **.455** | -.395 | -.197 | .021* |
| 17 | 25.3 | **57.3** | 16.0 | 1.3 | -.281 | **.593** | -.454 | -.038 |
| 18 | 36.0 | 21.3 | 16.0 | **26.7** | -.171 | .006* | -.084 | **.251** |
| 19 | 2.7 | **84.0** | 4.0 | 9.3 | -.130 | **.441** | -.191 | -.356 |
| 20 | 30.7 | 6.7 | 25.3 | **37.3** | -.308 | -.001 | -.042 | **.332** |
| 21 | 4.0 | 18.7 | **57.3** | 20.0 | -.166 | -.301 | **.196** | .132* |
| 22 | **58.1** | 2.7 | 18.9 | 20.3 | **.550** | -.180 | -.231 | -.378 |
| 23 | 6.7 | 5.3 | **65.3** | 22.7 | -.146 | -.046 | **.347** | -.283 |
| 24 | 1.3 | **42.7** | 37.3 | 18.7 | .067* | **.259** | -.074 | -.257 |
| 25 | 5.3 | **66.7** | 10.7 | 17.3 | -.294 | **.475** | -.184 | -.266 |
| 26 | 2.7 | 6.7 | **89.3** | 1.3 | -.115 | -.039 | **.177** | -.228 |
| 27 | **60.8** | 13.5 | 12.2 | 13.5 | **.304** | -.300 | -.142 | .002* |
| 28 | 2.7 | 8.0 | **65.3** | 24.0 | .021* | -.088 | **.291** | -.276 |
| 29 | **80.0** | 5.3 | 12.0 | 2.7 | **.183** | -.111 | -.084 | -.130 |
| 30 | 29.3 | **36.0** | 24.0 | 10.7 | -.063 | **.429** | -.157 | -.357 |

*Note.* Keys are in bold. Distractors with a positive point-biserial coefficient are marked with an asterisk (*).

# Appendix C:

## Distractor analysis of item scores on the AALT (n=46)

| Item | Response frequencies | | | | Option point-biserials | | | |
|---|---|---|---|---|---|---|---|---|
| | % A | % B | % C | % D | A $r_{p\text{-}bis}$ | B $r_{p\text{-}bis}$ | C $r_{p\text{-}bis}$ | D $r_{p\text{-}bis}$ |
| 1 | **71.7** | 2.2 | 26.1 | 0 | **.475** | -.218 | -.415 | -* |
| 2 | 4.3 | **84.8** | 0 | 10.9 | -.108 | **.018** | -* | .050* |
| 3 | 2.2 | **73.3** | 22.2 | 2.2 | .095* | **.190** | -.229 | -.019 |
| 4 | 17.4 | 26.1 | **52.2** | 4.3 | -.222 | -.481 | **.644** | -.128 |
| 5 | **57.8** | 8.9 | 20.0 | 13.3 | **.125** | -.158 | .126* | -.198 |
| 6 | 15.2 | 30.4 | 10.9 | **43.5** | -.053 | .152* | -.177 | **.008** |
| 7 | 19.6 | **43.5** | 6.5 | 30.4 | -.344 | **.427** | .236* | -.290 |
| 8 | 19.6 | 4.3 | **54.3** | 21.7 | -.240 | -.108 | **.440** | -.247 |
| 9 | 32.6 | **26.1** | 21.7 | 19.6 | -.140 | **.371** | .035* | -.282 |
| 10 | **71.7** | 13.0 | 8.7 | 6.5 | **.244** | -.320 | .227* | -.269 |
| 11 | 21.7 | **47.8** | 2.2 | 28.3 | -.318 | **.297** | -.218 | .033* |
| 12 | 15.2 | 8.7 | **47.8** | 28.3 | -.192 | -.171 | **.446** | -.235 |
| 13 | 39.1 | 10.9 | 8.7 | **41.3** | -.355 | -.003 | -.245 | **.495** |
| 14 | **30.4** | 32.6 | 23.9 | 13.0 | **.514** | -.282 | -.167 | -.098 |
| 15 | 17.4 | **43.5** | 17.4 | 21.7 | -.123 | **.394** | -.156 | -.217 |
| 16 | **80.4** | 10.9 | 8.7 | 0 | **.512** | -.337 | -.348 | -* |
| 17 | 15.6 | **64.4** | 17.8 | 2.2 | -.334 | **.609** | -.459 | .035* |
| 18 | 45.7 | 13.0 | 19.6 | **21.7** | -.231 | -.073 | -.020 | **.357** |
| 19 | 2.2 | **82.2** | 11.1 | 4.4 | -.226 | **.419** | -.300 | -.158 |
| 20 | 34.8 | 6.5 | 19.6 | **39.1** | -.126 | -.067 | -.145 | **.275** |
| 21 | 2.2 | 13.3 | **66.7** | 17.8 | -.225 | -.359 | **.286** | .053* |
| 22 | **52.2** | 2.2 | 17.4 | 28.3 | **.452** | -.161 | .052* | -.494 |
| 23 | 4.4 | 4.4 | **64.4** | 26.7 | -.252 | -.190 | **.603** | -.447 |
| 24 | 6.5 | **52.2** | 28.3 | 13.0 | -.235 | **.219** | -.069 | -.061 |
| 25 | 4.3 | **69.6** | 15.2 | 10.9 | .096* | **.372** | -.261 | -.311 |
| 26 | 2.2 | 17.8 | **73.3** | 6.7 | -.217 | -.295 | **.478** | -.267 |
| 27 | **65.2** | 4.3 | 23.9 | 6.5 | **.309** | -.190 | -.080 | -.302 |
| 28 | 8.7 | 13.0 | **67.4** | 10.9 | -.245 | -.085 | **.300** | -.137 |
| 29 | **47.8** | 10.9 | 26.1 | 15.2 | **.388** | -.364 | .030* | -.261 |
| 30 | 13.0 | **39.1** | 26.1 | 21.7 | .223* | **.411** | -.244 | -.408 |

*Note.* Keys are in bold. Distractors with a positive point-biserial coefficient are marked with an asterisk (*).

# British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

## INTERACTING WITH VISUALS IN L2 LISTENING TESTS: AN EYE-TRACKING STUDY

AR-A/2015/1

**Ruslan Suvorov**

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203