Check for updates

Article

LANGUAGE TESTING

Language Testing
2021, Vol. 38(4) 511–535
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0265532220951504
journals.sagepub.com/home/ltj
SAGE

An eye-tracking study of attention to visual cues in L2 listening tests

Aaron Olaf Batty
Keio University, Japan

Abstract

Nonverbal and other visual cues are well established as a critical component of human communication. Under most circumstances, visual information is available to aid in the comprehension and interpretation of spoken language. Citing these facts, many L2 assessment researchers have studied video-mediated listening tests through score comparisons with audio tests, by measuring the amount of time spent watching, and by attempting to determine examinee viewing behavior through self-reports. However, the specific visual cues to which examinees attend have heretofore not been measured objectively. The present research employs eye-tracking methodology to determine the amounts of time 12 participants viewed specific visual cues on a six-item, video-mediated L2 listening test. Seventy-two scanpath-overlaid videos of viewing behavior were manually coded for visual cues at 0.10-second intervals. Cued retrospective interviews based on eye-tracking data provided reasons for the observed behaviors. Faces were found to occupy the majority (81.74%) of visual dwell time, with participants largely splitting their time between the speaker's eyes and mouth. Detected gesture viewing was negligible. The reason given for most viewing behavior was determining characters' emotional states. These findings suggest that the primary difference between audio- and video-mediated L2 listening tests of conversational content is the absence or presence of facial expressions.

Keywords

Eye-tracking, facial expressions, gesture, language assessment, listening assessment, nonverbal communication, video listening test, visual cues

Nonverbal communication is accepted as a major component of communicative listening in humans (Burgoon et al., 2016). Many language testing researchers have argued that a valid test of foreign or second language (L2) listening comprehension would include such information, and can easily do so through video (e.g., Gruba, 1997; Ockey, 2007; Wagner, 2002). Nevertheless, what visual cues draw examinee attention in such tests

Corresponding author:
Aaron Olaf Batty, Keio University, 4411 Endo, Fujisawa, Kanagawa 252-0883, Japan.
Email: abatty@sfc.keio.ac.jp

remains largely undetermined. Although researchers have investigated examinee attention, most work has only addressed time spent oriented toward the video (Ockey, 2007; Wagner, 2007, 2010a) and examinees' perceptions while taking the test (Wagner, 2008), or has attempted to describe attention via retrospective verbal reports (Ockey, 2007; Suvorov, 2018). Recent work (Suvorov, 2015, 2018) has employed eye-tracking methodology in an attempt to understand how much examinees use the visual channel in video-mediated listening tests, but none have applied eye-tracking methods to investigate the attentional draw of specific visual cues in the video listening content. The present research seeks to fill these gaps in knowledge and method by employing eye-tracking methodology to determine the nonverbal communicative and other visual cues most attended to by examinees in a video-mediated L2 listening comprehension test, as well as the test takers' motivations for attending in the ways that they do.

Background

Nonverbal behavior

Nonverbal communication is typically understood to encompass all communicative events which are neither spoken or written (Knapp et al., 2014). It is well established as a major component of communicative listening in humans, comprising up to 66% of the information in any face-to-face interaction (Burgoon et al., 2016). For sighted people, in most cases (excepting, e.g., telephone conversations, radio/podcast listening, and public address announcements), listeners are able to use nonverbal cues such as facial expressions, gestures, and the setting as a "co-text" to the verbal signals to enhance comprehension and aid interpretation (Rost, 2016, p. 42). Nonverbal behavior is particularly useful for revealing the speaker's attitudes, emotions, personality, and interpersonal relations (Burgoon et al., 2016; Knapp et al., 2014). It is also instrumental in discerning social cues, especially when those cues are incongruous with the verbal message, operating as a kind of "fact check" for ambiguous statements. Nonverbal communication is an important carrier of social meaning in informal human conversational interactions, whereas the verbal channel is discussed as typically information laden when the content is factual, abstract, or persuasive in nature (Burgoon et al., 2016).

There are a number of taxonomies of nonverbal cues in the literature, but most are based on the work of Ekman and Friesen (1969), which remains the most well-known classification of nonverbal behaviors (Kendon, 2004). Ekman and Friesen (1969) break nonverbal behavior into two broad categories, namely, facial expressions and gestures, each of which are described in turn below.

Facial expressions. Facial expressions (also called *affect displays*) are probably the most important carrier of nonverbal information, and have been repeatedly found to be essentially universal to the human species (Ekman, 2017; Jack et al., 2016), although cultures differ in how and when they display them (Matsumoto, 2006) or where on the face they look to discern them (Jack et al., 2012; Yuki et al., 2007). Many facial expressions even appear among children who were born blind (Eibl-Eibesfeldt, 1972; Magnusson, 2006), and some are shared with other primates (Eibl-Eibesfeldt, 1972). Within the first few

days of life, sighted infants already discern them, showing a preference for viewing happy facial expressions (Farroni et al., 2007).

Although they are often unconscious, facial expressions are paralinguistic, and are sometimes used deliberately to communicate information beyond what is included in the verbal stream (Bavelas & Chovil, 2000). They can be used to signal emphasis, signal a question, represent an emotion or an opinion, add commentary to the verbal channel, or be used in the listener role as a form of backchanneling (Chovil, 1991). Although much of the history of research into facial expressions dating back to Darwin posited a direct link between specific emotions and specific facial expressions, this link appears to be weaker than was long assumed, as has been found both in the lab (Durán et al., 2017) and in natural settings (Fernández-Dols & Crivelli, 2013), owing to cultural and even individual differences in display preferences (Ekman, 2017). Although the relationship between emotion and facial expressions can be complex, they are nonetheless an important component of the communication of emotional or psychological states, upon which humans rely in face-to-face communication.

Gesture. Gesture may be the foundation of all human communication systems (Fay et al., 2013), and appears in infants prior to the ability to use verbal language (Burgoon et al., 2016). In fact, gesture alone has been found to be more communicative than gesture accompanied by non-linguistic (i.e., gibberish) vocalizations in communicating novel information to a naive interlocutor (Fay et al., 2013), and may even communicate as much alone as spoken language does without gesture (Holler et al., 2009). Gestures can be placed on a continuum from less- to more-language-like, ranging from unconscious, non-conventional movements while speaking, through increasingly representative movements all the way to full sign language (McNeill, 1992, 2000), but although several classification schemes exist, they are always somewhat arbitrary (Kendon, 2004). McNeill's (1992, 2000) classification system—a refinement of the system first suggested by Ekman and Friesen (1969) which was based on the work of Efron (1941)—is often used in the gesture literature. It is, however, more granular than necessary for the present research. Furthermore, the McNeill classifications can be mapped back to the original Ekman and Friesen nonverbal classification system, which remains the most widely known (Burgoon et al., 2016; Stam & McCafferty, 2008). As such, the Ekman and Friesen system is used in the present research. It is described below.

Gestures typically referred to as *emblems* (orig. Efron, 1941) are very near the "sign language" end of the language-likeness continuum. Much like words, these gestures have specific, semantic meanings, which are not immediately retrievable from the gesture form itself. A demonstrative example is the "okay" hand symbol used in North American and other Anglophone countries (i.e., the thumb and index finger joined in a circle, with the remaining digits extended). It is understood by anyone in that culture, but must be learned at some point, as its meaning is not immediately obvious. Roughly the same gesture, however, can mean "money" in some East Asian cultures. As such, emblems are necessarily culturally bound (Ekman, 1999; Ekman & Friesen, 1969), resulting in language-like differences in meaning even for the same gesture.

The category of gestures called *illustrators* comprises virtually any gesture that is not an emblem. It also originates in the work of Efron (1941), and was later popularized by

Ekman and Friesen (1969; Ekman, 1999). These are gestures used to signify or illustrate ideas, objects, or processes that are difficult to communicate otherwise, as a kind of "gestural onomatopoeia" (Efron, 1972, pp. 121–122), used to help organize utterances (Ekman, 1999; Ekman & Friesen, 1969; Kellerman, 1992) and aid difficult-to-understand utterances (Holler & Beattie, 2003). Illustrators encompass virtually all gestures occurring during or instead of speech, including those which are used to emphasize a word or phrase, those that "sketch a path or direction of thought" (Ekman, 1999, p. 47); deictic gestures, depictions of bodily action or spatial relationships, demonstrations of shapes with the hands, and gestures that "depict the rhythm or pacing of an event" (p. 47). Virtually any meaningful movement carried out while engaged in face-to-face communication is some form of illustrator.

There are two further forms of gesture in the Ekman and Friesen (1969) classification system: *adaptors*, which are forms of self-touching (e.g., scratching one's nose); and *regulators*, which are movements to facilitate turn-taking in conversation, although these functions can also be carried out with illustrators or emblems (Ekman, 1999). As the former is not typically communicative, and the latter is superseded by either illustrators or emblems, the present research classifies all gestures as either emblems or illustrators as described above.

Nonverbal communication in SLA. Much of the research into nonverbal communication in SLA has been concerned with gestures. Representative gestures have been found to greatly improve the comprehension of L1 speech by disambiguating it (Holle & Gunter, 2007; Wu & Coulson, 2007), freeing up cognitive resources for the listener to focus on meaning instead (Skipper et al., 2007). Dahl and Ludvigsen (2014) investigated the impact of gestures on L1 and L2 comprehension, finding that the effect of their presence in an information-gap task was especially substantial in the L2 context, with performances that were not only significantly better than when completing the task without gestures, but that were not significantly different from those completing the task in their L1. Kida (2008) investigated the comprehension of nonnative speakers of French engaged in in-person conversations with the interlocutors either visible to one another or occluded by a screen to remove gesture from the interaction. He found that gesture was important to comprehension overall, but especially to lower-level users of the L2, who utilized it as a coping strategy when unknown words appeared. In the language classroom, teachers' use of representative gestures has been also found to aid L2 comprehension (Lazaraton, 2004; Sime, 2006) and to improve the effectiveness of corrective feedback (Nakatsukasa, 2016; Sato, 2019; Wang & Loewen, 2016).

Several SLA researchers have investigated the impact of a wider array of nonverbal cues on L2 comprehension. One of the earliest investigations was conducted by Riseborough (1981), wherein the researcher presented a mix of audio, face-only video, video with vague gestures, and video with more explicitly meaningful gestures, finding that the presence of more nonverbal cues led to better recall and comprehension. A similar design was employed by Sueyoshi and Hardison (2005), who found that the presence of nonverbal information (face or face and body) significantly facilitated comprehension, especially for low-proficiency listeners. Several other researchers (Baltova, 1994; Brett, 1997) have also investigated the impact of video containing nonverbal communicative

cues as well as other visual information (e.g., physical context, visual aids), and have found it broadly facilitative of comprehension. Overall, it appears that the more visual information included, the more L2 listeners comprehend.

Video-mediated listening tests

Given the importance of nonverbal and other visual cues to comprehension, many L2 assessment researchers have investigated the use of video in listening comprehension tests. Most studies compare video listening tests to audio-only counterparts (e.g., Baltova, 1994; Batty, 2015, 2018; Coniam, 2001; Cubilo & Winke, 2013; Lesnov, 2018; Pusey & Lenz, 2014; Suvorov, 2009; Wagner, 2010b, 2013) and typically find that video tests are easier. In addition to the question of comparative difficulty, however, a number of other themes more pertinent to the present research have repeatedly appeared in the literature.

Time spent watching. Several studies have investigated the amount of time examinees spend watching the video content in video-mediated listening tests. Wagner (2007) found that there was a significant text-type effect, with examinees orienting themselves toward the video 72% of the time during dialogues and 67% of the time during lecture videos. Ockey (2007) observed a similar effect, as did Wagner (2010a) several years later, again finding a significant difference in the amount of time spent watching dialogue (58.5%) and lecture (41.6%) videos. These results suggest that the information examinees find most useful in the videos is either more present or more beneficial in informal dialogues than in academic lectures.

Attention to visual cues. Examinee attention to specific visual cues has appeared a number of times in the literature. In the study discussed above, Ockey (2007) used retrospective verbal reports as a means of exploring the specific visual cues to which participants referred in the videos. Participants reported a wide range of nonverbal cues, from no cues at all through the entire range of lip movements, facial expressions, hand gestures, and body gestures. The two nonverbal cues reported by the most participants were "facial gestures to indicate opinion" and "body gestures to indicate emphasis" (p. 530). Wagner (2008) used while-listening verbal reports both while the videotext was playing as well as while answering comprehension questions. The former was achieved by stopping the video playback at predetermined intervals so that the examinee could verbalize in English (L2) what he or she was looking at and why; the latter was via think-aloud protocols. The "pausing" method is problematic, however, as it required the examinees to repeatedly "pause" their thinking as the video was paused, which likely fundamentally altered the examinees' interactions with the stimulus (Gorin, 2006). The total number of references to visual cues ranged from zero through sixteen, but the reliability of these results is unclear, as transcripts of the interviews seem to suggest that the examinees' English level may not have been adequate to explain their thoughts. Both studies also relied entirely on self-reports, which may themselves be unreliable.

Suvorov (2018) partly addressed this methodological problem with the use of an eye-tracker and cued retrospective reporting. Participants' retrospective verbal reports were conducted while viewing. These interviews revealed that when participants viewed

lectures without visual aids, the visual cues most commented on were nonverbal cues related to the "speaker's mouth, face, head, hands, eyes" (91% of participants), followed by "speaker's gestures, body movements" (61%) (2018, p. 150). Despite the use of sophisticated eye-tracking equipment, however, no attempt was made to quantitatively examine the visual attention to any specific visual cue within the videotext, relying, as in previous studies, solely upon participant reports of visual attention made in the L2, which is one issue the present research seeks to address.

Eye-tracking methodology

Eye-tracking research employs the use of hardware and software to measure the movement of the eyes (i.e., oculomotor events), store them, and provide data and visual outputs, such as location data in tabular format and/or visualizations. Most systems can provide data on participants' looking at user-defined areas of interest (AOIs) in the visual field. For complex AOIs, however, manual scanpath analysis may be necessary in cases where human perception remains superior to that of software (Holmqvist et al., 2011). Although some eye movement metrics can be used to infer certain cognitive processes (e.g., skips, regressions), the most common eye-tracking variable is the total duration of time spent looking at a certain AOI (Godfroid, 2019). Most eye-tracking research also incorporates verbal data collected as participants view videos of their own eye movements (i.e., scanpath-overlaid video), as the use of both the quantitative data from the eye-tracking apparatus and the qualitative data from the verbal reports allows the researcher to triangulate a reasonable interpretation of the participants' viewing behavior (Godfroid, 2019; Holmqvist et al., 2011).

Eye-tracking in listening comprehension research. Although most eye-tracking language research, both in L1 and L2 contexts, has been focused on reading (Roberts & Siyanova-Chanturia, 2013), there have been a number of influential studies of eye movements while comprehending spoken language. Broadly speaking, it has been demonstrated that listeners focus their visual attention on objects as they appear or are anticipated to appear in the linguistic stream (Altmann, 2011), even if told to ignore the auditory stream and instead look elsewhere (Salverda & Altmann, 2011). Although subjects do tend to glean information about a scene by looking at the setting, this typically only happens at the very beginning of an activity (Boland, 2004). Eye-tracking has also often been used as a method by which to capture psycholinguistic processing, such as in studies of word recognition (e.g., Marian & Spivey, 2003) or grammatical processing/acquisition (e.g., Winke, 2013). A thorough overview of the use of eye-tracking with aural input may be found in Tanenhaus and Trueswell (2006).

Eye-tracking in L2 assessment research. Eye-tracking methodology has seen increasing use within the field of L2 assessment. Although it is not possible to link eye-tracking measures to any specific cognitive process (Conklin et al., 2018), most work has focused on the cognitive behavior of reading test examinees (e.g., Bax, 2013; Bax & Weir, 2012; Brunfaut, 2016; Brunfaut & McCray, 2015; McCray & Brunfaut, 2018), although Winke and Lim (2014) have also employed eye-tracking to examine listening test examinees'

looking at keywords in the written questions accompanying the audio stimulus, and have conducted a study (Winke & Lim, 2015) of raters' use of rubrics in writing assessment. As of the time of this writing, however, the only published application of eye-tracking methodology to video L2 listening assessment remains Suvorov's (2015, 2018) previously discussed work.

Most of these studies, however, relied on relatively simple and static AOIs to capture eye-tracking metrics. AOIs can be, and often are, automatically generated over text by the eye-tracking software, which limits the data collected to that of AOIs that can be relatively easily mapped by the software. Most eye-tracking software defaults to, or may even limit the user to, rectangular AOIs. There are normally extra steps involved in setting non-rectangle or dynamic AOIs: when AOIs are irregularly shaped, change shape, or are set to move dynamically within or over an image on screen (or to follow an image as it moves on screen), the shape, placement, and/or path of the AOIs must be aligned over the image by hand (video frame by video frame) or through extra programming, thus the capturing of attentional metrics with non-automatically-set or non-rectangular AOIs can be difficult or time consuming. Because of these methodological complexities, which can also be resource depleting, the use of eye-tracking methodology in video listening assessment has, as of yet, not explicitly described the amount of time spent attending to specific visual cues. The present research seeks to address this methodological shortcoming.

Research questions

Despite incremental improvements in studies of examinees' visual attention during video-based listening tests, important questions remain unexamined. As such, the following research questions are posed:

RQ1. What are the specific nonverbal communicative or other visual cues to which examinees attend when taking a video-mediated L2 listening test?

Although much research has been conducted into how long examinees view the videos in such tests and/or the visual cues they mention in verbal reports, none has employed eye-tracking methodology to objectively quantify to what nonverbal or other visual cues examinees actually attend in video-mediated listening tests.

RQ2. What are the reasons for examinee viewing behavior when taking a video-mediated L2 listening test?

As discussed previously, qualitative interview data based on eye-tracking scanpath replay can aid in interpretation of observed behavior by providing information on the viewer's motivations for that behavior.

If examinees' use of visual cues in video-mediated listening tests were better understood, test developers could make more theoretically grounded decisions in test specification, task design, and item development.

Table 1. Participant demographics.

ID	Sex	Age	TOEFL ITP	CEFR
1	F	21	407	A2
2	F	20	473	B1
3	F	22	530	B1
4	M	22	533	B1
5	F	20	373	A2
6	F	21	390	A2
7	F	19	470	B1
8	F	19	437	A2
9	F	19	447	A2
10	F	19	483	B1
11	F	19	453	A2
12	M	19	460	B1

Method

Participants

Participants were 12 Japanese students of English studying at a large university in Japan. Given the close hand-coding of the eye-tracking data (explained below), a larger sample was neither practical nor necessary given the exploratory nature of the study, and the concomitantly large amount of data collected from each participant. All had studied English for a minimum of six years. Compulsory English classes in Japan tend to be based in reading and grammar translation, and the likelihood that video materials played more than a minor role is low. All participants had TOEFL ITP (Educational Testing Service, 2016) scores available as an external measure of general English proficiency. Participant demographics are displayed in Table 1. Common European Frame of Reference (CEFR) levels are according to the TOEFL cut scores published by ETS (Tannenbaum & Baron, 2011).

Equipment

The eye-tracker used was the open source, head-mounted Pupil Dev system (Kassner et al., 2014). The use of head-mounted equipment prevents the problem encountered by Suvorov (2015) wherein gaps in the captured data resulted from participants looking down from the screen to take notes. The Pupil Dev headset uses dark pupil detection with a refresh rate of 30 Hz. Gaze points are mapped to video captured by the "world" camera mounted above the eye, also with a refresh rate of 30 Hz. Although this sampling rate may not be appropriate for a psycholinguistic study wherein the main metrics of interest would be those of fixations, saccades, and other very fast movements, the sampling rate necessary depends on the objectives of the study (Conklin et al., 2018; Holmqvist et al., 2011). Holmqvist et al. (2011) advised using an eye tracker with a speed of at least double the speed of the movement one desires to measure. As the present study is only concerned with total dwell time measured at tenths of a second, a sampling rate of 30 Hz is

Table 2. Video descriptions.

Video	Total duration	Hand duration	Illustrative gestures count	Illustrative gestures duration	Emblematic gestures count	Emblematic gestures duration	Objects count	Objects duration
1	95	93.1	12	25.0	0	0.0	2	89.6
2	45	42.9	4	28.8	0	0.0	0	0.0
3	38	21.6	6	14.9	0	0.0	4	15.9
4	99	36.9	13	24.5	0	0.0	4	17.6
5	88	53.5	12	36.7	1	0.9	6	50.4
6	53	37.4	1	1.6	0	0.0	2	44.0

Note: Duration times in seconds. Durations represent the amount of time the cue was present in the video frame. Counts are the number of times the cue was present in the video frame.

sufficient (Godfroid, 2019, p. 326). The program Pupil Capture captures the gaze positions, world video (i.e., the video of what the participant sees), pupil video (i.e., the video of the participant's eye), and audio. Visualization of the data is achieved via the partner application Pupil Player. See the supplemental file (downloadable from this paper's *Language Testing* website location) for further technical details.

Instrument

The instrument was comprised of six short videos of conversations with one multiple-choice item per video to ensure item independence. In order to model the types of items typically found on listening tests, three of the items were explicit (testing comprehension of information stated explicitly in the scene); three were implicit (requiring inference). Conversations were chosen as the stimuli due to the consistent finding in the literature that examinees view conversation content more than they do other types of content (see Ockey, 2007; Wagner, 2007, 2010a discussed previously). Video durations ranged from 38 to 99 seconds. Most videos contained gestures and objects in addition to facial expressions. Please see Table 2 for numeric data on the videos, including durations and counts of non-facial visual cues present. Textual descriptions of the video content, as well as English translations of the questions, are available in the supplemental file.

The videos were scenes from the American television program *Curb Your Enthusiasm*, chosen because it is improvised by the actors based on scene outlines rather than performed according to a script. This results in quasi-authentic language use, complete with hesitation, misspeaking, and false starts. Nonverbal behavior such as gesture and facial expression, which has been found to differ in production when not produced spontaneously (Namba et al., 2017), also closely resemble authentic language production. Finally, the series features no "laugh track," and the selected scenes contained no background music. Each featured only one man and one woman.

The instrument was administered in a web browser via the quiz module in Moodle 2.6. Each video was preceded by a preview of the item stem to provide a purpose for listening, following the recommendation of Buck (1991). As a result, examinees did not

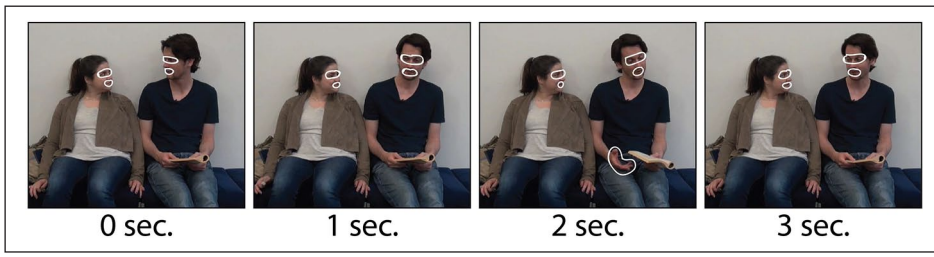

Figure 1. Demonstration of complex, dynamically moving AOIs.

need to remember every detail of the conversation, only that which was pertinent to the task at hand. The participant then viewed the video once and was re-presented with the stem along with four answer options. All instructions and items were presented in the participants' L1 of Japanese in order to eliminate any ambiguity surrounding the purpose of the task, as in several studies already discussed (see Baltova, 1994; Brunfaut, 2016; Brunfaut & McCray, 2015).

Procedure

Eye-tracking data collection. The data collection procedure was developed according to the recommendations of Pernice and Nielsen (2009). The participants first completed a practice task while wearing the headset to familiarize them with the test format, and to allow time to settle into a natural position in front of the monitor. After the practice task, calibration was carried out, adjusting as necessary until the fraction of used datapoints surpassed 0.75, in accordance with the developer's recommendations. Once completed, the participants began the test, with their eye movements recorded. As the participant watched the videos, the researcher monitored the data quality and watched the gaze positions on a separate monitor obscured from the participant's view, making note of behaviors to address in the post-test interview. See the supplemental file for more information on calibration, data monitoring, and drift correction.

Retrospective verbalization data collection. As eye-tracking metrics are best interpreted with the aid of verbal reports (van Gog et al., 2005), retrospective verbalizations were collected via semi-structured interviews based on gaze displays (Holmqvist et al., 2011). After the test, the researcher returned to the experimental area and viewed the eye tracker video with the participant, stopping it periodically to ask what the participant remembered about his/her behavior. A basic interview script was used to ensure similar phrasing of questions between participants. To ensure that participants' English (L2) proficiency would not impact the quality or content of their responses, all interviews were carried out in the participants' L1 by the present researcher, who has studied and worked in Japan for nearly twenty years. Interviews were also attended by a native Japanese speaking assistant.

Data transformation

Eye-tracking data. The scanpath-overlaid videos were imported into NVivo qualitative data analysis software (*NVivo Qualitative Data Analysis Software*, n.d.) for manual

scanpath analysis. Figure 1 provides an example of the challenges inherent in tracking nonverbal behavior. Even in a scene with a stable camera, as shown, the positions and shapes of the AOIs are in constant motion. Although some recent advanced eye-tracking software packages allow for dynamic AOIs, many still require them to be rectangular, and of those which allow more complex shapes, the process of animating the continual changes would not be practical, as it would become necessary to essentially re-draw them at intervals so frequent that doing so would offer little advantage over manual scanpath analysis, especially when using a head-mounted eye-tracker, where the video geometry is not constant. As a result of this, dynamic AOIs are still usually coded manually (Godfroid, 2019). Furthermore, with respect to nonverbal communicative cues, what may be at one moment an AOI associated with the speaker may be that of the listener the next. What may simply be a "hand" one moment may become an "illustrative gesture" the next. As such, it is inadvisable to attempt to define AOIs that move and change over longer periods of time, and sometimes it is simply impossible to define such AOIs beforehand (Conklin et al., 2018). For these reasons, manual scanpath analysis was the only practical option for addressing the research questions.

Each of the 72 scanpath-overlaid videos (12 participants × 6 videos) was coded at NVivo's maximum coding resolution of one-tenth of a second according to an iteratively developed list of viewing behaviors of interest. A detailed list and descriptions can be found in the supplemental file. This method results in not only a count of codes, but their durations, which are then available for further analysis, and can be cross-referenced with any other codes used (e.g., video being viewed, reason given for the behavior in interview). The code durations were converted to percentages of the duration of the videos, following the methodology of Suvorov (2015). These percentages can be understood as measures of total dwell time.

Interview data. Interviews were transcribed and imported into NVivo for qualitative analysis. Specific answers to the researcher's queries were coded with the oculomotor event in question, the reason offered for the behavior, and whether the participant felt that it was related to the item associated with the video. Codes were developed according to the same iterative approach as above and applied to all interviews. The full list of codes and their detailed descriptions can be found in the supplemental file.

Data analysis

Eye-tracking data were analyzed quantitatively with the qualitative (interview) data aiding in interpretation. An overview of the dwell times was achieved through descriptive statistics. Coded qualitative reports from participants were collated into contingency tables for comparative analysis.

Results

RQ1: Attention to visual cues

Overall descriptive statistics for the oculomotor events are displayed in Table 3 and visualized in Figure 2. Table 4 breaks out the percentages for the face events only; values

Table 3. Descriptive statistics for dwell time percentages.

Visual cue	Min.	Med.	*IQR*	Max.
Face	53.37	81.74	12.51	95.86
Speaker's face	32.64	60.98	16.67	82.89
Speaker's face regions	4.00	38.13	18.25	69.21
Speaker's eyes	0.00	18.23	31.51	68.11
Speaker's mouth	0.00	16.97	33.02	69.21
Speaker's face scan	0.00	19.85	18.57	67.33
Listener's face	0.00	5.00	7.46	36.79
Listener's face regions	0.00	2.67	5.26	13.58
Listener's eyes	0.00	0.70	4.77	13.58
Listener's mouth	0.00	0.00	0.89	6.42
Listener's face scan	0.00	0.00	3.41	24.34
Alternating between faces	0.00	9.73	18.67	46.60
Hands	0.00	1.00	2.84	9.56
Illustrative gestures	0.00	0.00	1.02	6.89
Emblematic gestures	0.00	0.00	0.00	1.02
Body	0.00	0.00	2.58	13.52
Objects	0.00	4.53	11.39	30.11
Setting	0.00	1.15	4.60	13.58

Note: Dwell times are percentages of their respective videotexts. Aggregate category totals may exceed the sums of those below, as indistinct events were sometimes coded directly as the aggregate category. Sample was 12 examinees interacting with six videotexts.

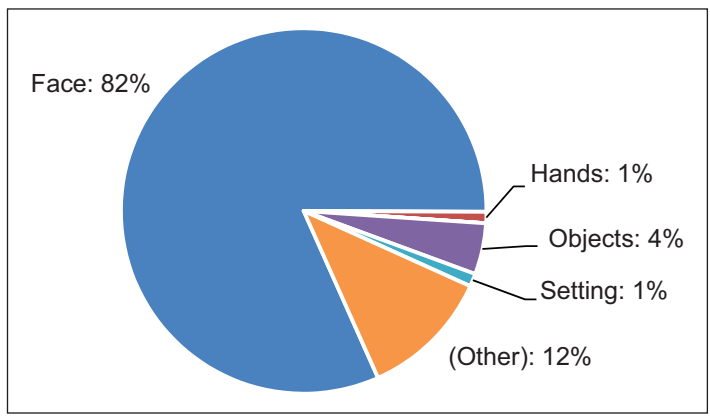

Figure 2. Visualization of median dwell time.

there are not percentages of the total dwell time, but of the facial dwell time. In all cases *speaker* refers to the character speaking at the time, and *listener* denotes the character listening to the speaker. Behaviors tended to be heavily skewed to the right, indicating a

Table 4. Breakdown descriptive statistics for total facial dwell times.

Visual cue	Min.	Median	*IQR*	Max.
Speaker's face	39.05	78.71	16.54	100.00
Speaker's eyes	0.00	20.61	39.19	90.11
Speaker's mouth	0.00	19.89	40.53	83.49
Speaker's face scan	0.00	24.55	22.81	70.79
Listener's face	0.00	6.84	9.30	44.42
Listener's eyes	0.00	0.85	5.56	15.97
Listener's mouth	0.00	0.00	0.76	9.19
Listener's face scan	0.00	0.00	4.76	29.38
Alternating between faces	0.00	12.39	21.21	52.78

Note: Values are percentages of total time spent orienting toward faces.

high degree of variability in individuals' interactions with the videotexts. As such, non-parametric descriptive statistics are reported.

The visual cue accounting for the largest amount of dwell time was the face, with a median of 81.74%. The speaker's face comprised most of the total face dwell times (78.71%) and 60.98% of the dwell time overall. Time spent oriented toward faces was comprised almost equally by the speaker's eyes (20.61%), mouth (19.89%), and scanning the speaker's face (24.55% of facial dwell time); these represented percentages of the total dwell time of 18.23%, 16.97%, and 19.85% respectively. The listener's face drew gaze for considerably less time than the speaker's face. The median dwell time on all cues associated with the listener's face was 5%, representing 6.84% of total facial dwell times.

The second most common facial visual cue was "Alternating between faces" (9.73% of the total, a median of 12.39% of the facial orientation). All hand events only accounted for a median of 1.00% of the total dwell time, which represents 1.40% of the total duration of hands appearing in the videos, and is lower, even, than either body or setting. Objects in the scenes were viewed for a median of 4.53% of the total time, representing 13.28% of the total time objects were visible in the video frame. Scene settings (e.g., a lamp in the background, etc.) were looked at with a median of only 1.15% of the total time.

RQ2: Reasons for attention

Reasons given for the viewing can be seen in Table 5. As the reason "Related to item" necessarily overlaps with other reasons given by the participant, it is summarized separately in Table 6.

The most common reason given for any facial watching was "Determining affect" (39%), that is, determining the person's mood by facial expression. Reasons given for viewing the eyes or scanning the face were largely for determining affect, but this was the reason given 64% of the time for scanning the speaker's face. Notably, nearly half (44%) of the reasons for watching the speaker's mouth was to supplement comprehension. In all such cases, participants reported that it was easier to understand what the

Table 5. Reasons provided for oculomotor events by percentage and absolute count.

Visual cue		Reason given							
		Determining affect	Supplementing comprehension	Cultural explanation	Determining who will speak next	Habit	Unconscious	Other	Total
Face	Pct.	39%	18%	2%	11%	8%	6%	17%	100%
	Cnt.	45	20	2	12	9	7	19	114
Speaker's face	Pct.	42%	23%	3%	–	14%	9%	9%	100%
	Cnt.	28	15	2	–	9	6	6	66
Speaker's eyes	Pct.	53%	17%	7%	–	17%	3%	3%	100%
	Cnt.	16	5	2	–	5	1	1	30
Speaker's mouth	Pct.	6%	44%	–	–	17%	17%	17%	100%
	Cnt.	1	8	–	–	3	3	3	18
Speaker's face scan	Pct.	64%	14%	–	–	–	14%	7%	100%
	Cnt.	9	2	–	–	–	2	1	14
Listener's face	Pct.	50%	6%	–	6%	–	6%	31%	100%
	Cnt.	8	1	–	1	–	1	5	16
Listener's eyes	Pct.	50%	25%	–	–	–	–	25%	100%
	Cnt.	2	1	–	–	–	–	1	4
Listener's mouth	Pct.	–	–	–	–	–	–	100%	100%
	Cnt.	–	–	–	–	–	–	1	1
Listener's face scan	Pct.	100%	–	–	–	–	–	–	100%
	Cnt.	2	–	–	–	–	–	–	2
Alternating between faces		28%	13%	–	34%	–	–	25%	100%
	Cnt.	9	4	–	11	–	–	8	32

(Continued)

Table 5. (Continued)

Visual cue		Reason given							
		Determining affect	Supplementing comprehension	Cultural explanation	Determining who will speak next	Habit	Unconscious	Other	Total
Hands	Pct.	4%	9%	9%	–	9%	52%	17%	100%
	Cnt.	1	2	2	–	2	12	4	23
Illustrative gestures	Pct.	6%	12%	12%	–	12%	41%	18%	100%
	Cnt.	1	2	2	–	2	7	3	17
Emblematic gestures	Pct.	–	–	–	–	–	100%	–	100%
	Cnt.	–	–	–	–	–	1	–	1
Body	Pct.	–	–	–	–	–	50%	50%	100%
	Cnt.	–	–	–	–	–	2	2	4
Objects	Pct.	–	5%	–	–	–	26%	68%	100%
	Cnt.	–	2	–	–	–	10	26	38
Setting	Pct.	–	14%	–	–	–	29%	57%	100%
	Cnt.	–	1	–	–	–	2	4	7

Note: Percentage values represent the row percentages.

Table 6. Percentages and counts of visual cue viewing behaviors reported to be related to items.

Visual cue		Related to Item
Speaker's face	Pct.	12%
	Cnt.	6*
Speaker's eyes	Pct.	4%
	Cnt.	2
Speaker's face scan	Pct.	6%
	Cnt.	3
Listener's face	Pct.	4%
	Cnt.	2
Alternating between faces	Pct.	16%
	Cnt.	8
Hands	Pct.	2%
	Cnt.	1
Objects	Pct.	66%
	Cnt.	33

*Includes one direct coding of "Speaker's face."

character was saying by watching the mouth. In interview, half of the reasons given for watching the listener's face were related to affect. For listener face scanning behavior, it was the reason given both times it was discussed in interview. Aside from the uninformative "Other," the reasons for "Alternating between faces" were fairly evenly split among "Determining affect" and "Determining who will speak next." The latter is of little theoretical importance, however, as it typically occurred in moments of silence, especially at the beginning of scenes, and was likely unconscious.

The most frequent reason for looking at a character's hands was that it was unconscious (52%). Although this is an aggregate category, it also contained events of looking at hands which were not being used in any communicative way (e.g., holding a pen or reaching for an object). The most common reason given for looking at the hands in these cases can be summed up as "it was moving." The two most informative reasons given for viewing illustrative gestures was to supplement comprehension and cultural explanations. As an example of the latter, Participant 4, who has spent some time studying in the United States, said, "Like—this is just my image, but—foreigners have this specific way of moving, right? That's probably related [to my looking]. It's because I kind of know that they also communicate with that [i.e., gestures]" (Translation mine).

Emblematic gestures had only one viewing event in the collected data. This is likely, however, to be partly a function of the fact that there was only one emblematic gesture in the entire test. The single participant who looked at the emblematic gesture (an "okay" sign), when her scanpath was shown to her, said that she did not remember that it had been an "okay" gesture, only that the character had raised his hand, and that she had looked unconsciously. Of the reasons given for the body events, all were either

"Unconscious" or "Other." For the setting, the large majority of the reasons given were simply coded as "Other."

Although Table 5 shows that 68% of the mentions of object viewing were classified as "Other," this obscures the relatively large number of times that the reason given was directly related to the question asked. Table 6 displays a breakdown of these codings. The largest percentage of behaviors reported to be related to the items was that of objects (66%), the likely reason for which will be discussed in the following section. The next most frequently cited cue was "Alternating between faces" (16%) followed by an aggregation of all "Speaker's face" cues.

Discussion

The results of this study demonstrate that examinees interacting with a video-mediated listening test tend to focus mostly on the face of whomever is speaking, with only small departures from this to directly look at gestures, objects, the setting, and so on. Participants appeared to largely split their time between watching the speaker's eyes or mouth. The reasons given for the former almost all referred to the expression of the speaker; in fact, the word *hyōjō* (表情; *expression* or *countenance*) appeared 20 times among nine of the participants when asked why they were looking at the speaker's eyes. This is as the nonverbal communication theories presented earlier would suggest, but it also seems to agree with the findings of Coniam (2001), whose listening test examinees reported that facial expressions were useful for determining the speakers' attitudes and predicting what they were likely to do or say next. Furthermore, it aligns with Suvorov's (2018) observation that the overwhelming majority of participants reported focusing on the "speaker's mouth, face, head, hands, eyes" (p. 150) while watching the videos in his test.

The reasons given for watching the mouth mostly centered on an increased facility for comprehension. This aligns with the findings of an eye-tracking study by Lansing and McConkie (2003), who found that most viewers of a video watched the speaker's eyes, only shifting gaze to the lips when comprehension became difficult (e.g., poor audio, lower L2 proficiency). The importance of lip reading was noted by Suvorov (2018) as well, and has been repeatedly demonstrated in the SLA literature (e.g., Hardison, 2018; Inceoglu, 2016).

A surprising finding was how little gestures seemed to draw examinees' eyes directly during viewing of the videotexts, given their prominence in the nonverbal communication and SLA literatures, and their importance to previous studies investigating examinee interaction with video-mediated listening tests. However, Gullberg and Holmqvist (2006) observed similar behavior in their eye-tracking study of gesture viewing, with their participants watching faces for 90–95% of the time in face-to-face interaction, and somewhat less when watching videos.

Only two of the participants in the present study, when asked about their looking at a gesture, made comments to the effect that it supplemented their understanding of what was said. This is in stark contrast to the findings of Wagner (2008), whose participants mentioned the gestures appearing in his videos repeatedly. However, it is important to note that the gestures that were most impactful there were representative gestures to facilitate the explanation of famous American Western lawman Wild Bill Hickock's

distinctive and unintuitive gun draw. Furthermore, Wagner interpreted his participants' use of the same gestures to explain the content back to him as an indication of their importance to comprehension. Another, perhaps more-likely interpretation, however, is that the content was simply too difficult to describe in words, and/or that his participants lacked sufficient vocabulary to do so, which would confirm their suggested role as a coping strategy for lower-proficiency L2 users (Kida, 2008; Skipper et al., 2007). The gestures that seemed to affect Cubilo and Winke's (2013) and Suvorov's (2018) participants most were topic organizing gestures in academic lectures, which do not appear in the present study.

Only one participant viewed the single emblematic gesture in the six videotexts, and she did not remember doing so. This perhaps illustrates the difficulty one might have in incorporating emblematic gestures into a test specification. Not only are they fairly uncommon (it was the only example found in six seasons of *Curb Your Enthusiasm*), examinees do not seem to choose to view them directly when attempting to answer comprehension questions. It is perhaps possible that the participants unconsciously glanced at that and other gestures in saccades (i.e., quick movements of the eyes) too fast for the eye-tracking hardware to register, or that the gestures were detected via parafoveal vision. Parafoveal vision is particularly well suited to registering movement, which obviates the necessity of watching gestures directly. Viewers have been found to do so most when the speaker looks at the gesture him- or herself or holds it for longer than usual (Gullberg & Kita, 2009). However, even when participants were specifically questioned as to whether they had noticed a gesture, they did not remember seeing it. As it has been shown that participants in eye-tracking research actually do remember what they looked at (Hansen, 1991), this finding should probably be accepted as genuine.

The large percentage of object viewing that was reported as being related to the item is likely owing to the fact that two of the items either directly or indirectly referred to objects. The first was Item 1, which depicted a man returning from shopping, accompanied by the question, "What did the man buy?" The second was Item 5, which depicted a man walking around a store, and the question was, "What is the man looking for?" In the former, participants looked at the garment bag the character was holding; in the latter, at various objects displayed for sale in the store (e.g., plates, vases). These items illustrate the large effect that task characteristics may have on viewing behavior, which is an effect observed elsewhere in the eye-tracking literature (Gullberg & Kita, 2009).

Implications for video listening test development

The present study has a number of clear implications for video listening test development. The first major implication is related to the finding that, in conversational stimuli, the visual cues with the highest dwell times were the seat of nonverbal cues for displaying affect/emotion: the face. In previous comparative video listening test work, it was unknown what information normally present in real-world listening was removed when an item was administered in an audio-only format, and what, therefore, was the cause of any video effect observed. By establishing that viewers spend most of their time focusing on faces, especially the speaker's face, video listening tests featuring conversational

stimuli can be designed with the knowledge that the major difference between the formats is likely to be the presence or absence of facial expressions.

Moreover, although gestures did not draw much conscious attention from the participants in the present study, there are two related implications for video listening test design. For gestures to play any part in the videotexts, scenes must be framed with angles wide enough that they capture gesture use, as "close-ups" sometimes obscure the hands, which may have partly contributed to the low amount of gesture viewing in the present study. The second implication for gestures in video listening tests is that emblematic gestures may not need to be incorporated unless they occur naturally, as it appears that they are unlikely to be looked at directly (and focused upon intently, except perhaps parafoveally) by examinees.

Limitations and directions for future research

One limitation of the present study is the small sample size. Although the amount of coding compensates for this with regard to the research questions at hand, the sample size complicates between-person analyses. Related to this is the imbalance between male and female participants, owing to difficulty encountered in recruiting males for the study. This renders gender-based analyses impossible as well. A wider range of proficiencies may have also revealed proficiency-based differences in viewing behavior. Finally, although the partial motivation for using a head-mounted eye tracker was to avoid problems encountered by Suvorov (2015), it also complicated data collection and analysis, while most participants ultimately did not take notes. Future research should focus on a larger, more gender-balanced sample with a wider range of proficiencies in order to conduct between-person, between-gender, and proficiency-based analyses, and make use of a remote eye-tracker that produces more easily analyzed outputs. An eye-tracker with a faster sampling rate may also allow closer analysis of looking behavior.

The findings presented here also offer several avenues for future research. The first of these is related to the impact of task characteristics on viewing behavior. Further work may explore this effect in detail, not only at the item level, but at the level of item type, as item type (i.e., explicit vs. implicit) has already been demonstrated to affect item difficulty significantly on video-mediated listening tests (Batty, 2018). Another possible topic for future research would be the specific impact of lip reading on video-mediated listening tests. Additionally, as individual sensitivity to nonverbal behavior has been repeatedly observed in the nonverbal communication literature (Gifford, 2006), the impact of individual differences upon viewing behavior would likely also be a fruitful topic of further inquiry. Finally, although the present study focused on conversational stimuli, further work could investigate patterns when other meaningful visual stimuli are included (e.g., presentation slides) in video L2 listening tests.

Acknowledgements

I would like to thank my Ph.D. supervisor, Luke Harding, for his support and direction in designing and carrying out this study, as well as my Ph.D. examiners, Tineke Brunfaut and Yasuyo Sawaki, and *Language Testing* Editor, Paula Winke, for their invaluable criticisms, suggestions, and insights.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Aaron Olaf Batty https://orcid.org/0000-0002-2760-4918

Supplemental material

Supplemental material for this article is available online.

References

Altmann, G. T. M. (2011). The mediation of eye movements by spoken language. In S. Liversedge, I. Gilchrist & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 979–1004). Oxford University Press.

Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*, *50*(3), 507–531. https://doi.org/10.3138/cmlr.50.3.507

Batty, A. O. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3–20. https://doi.org/10.1177/0265532214531254

Batty, A. O. (2018). Investigating the impact of nonverbal communication cues on listening item types. In E. Wagner & G. J. Ockey (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 161–175). John Benjamins.

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, *19*(2), 163–194. https://doi.org/10.1177/0261927X00019002001

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*(4), 441–465. https://doi.org/10.1177/0265532212473244

Bax, S., & Weir, C. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. *University of Cambridge ESOL Examinations Research Notes*, *47*, 3–14. https://www.cambridgeenglish.org/images/22669-rv-research-notes-47.pdf

Boland, J. E. (2004). Linking eye movements to sentence comprehension in reading and listening. In M. Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs, and beyond* (pp. 51–76). Psychology Press.

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System*, *25*(1), 39–53. https://doi.org/10.1016/S0346-251X(96)00059-0

Brunfaut, T. (2016). *Looking into reading II: A follow-up study on test-takers' cognitive processes while completing Aptis B1 reading tasks* (VS/2016/001; British Council Validation Series). The British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_final_with_hyperlinks_3.pdf

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study* (AR/2015/001; ARAGs Research Reports Online). The British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, *8*(1), 67–91. https://doi.org/10.1177/026553229100800105

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2016). *Nonverbal communication*. Routledge.

Chovil, N. (1991). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, *25*(1–4), 163–194. https://doi.org/10.1080/08351819109389361

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*, *29*(1), 1–14. https://doi.org/10.1016/S0346-251X(00)00057-9

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

Cubilo, J., & Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue Interpretation and note-taking. *Language Assessment Quarterly*, *10*(4), 371–397. https://doi.org/10.1080/15434303.2013.824972

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, *98*(3), 813–833. https://doi.org/10.1111/j.1540-4781.2014.12124.x

Durán, J. I., Reisenzein, R., & Fernández-Dols, J.-M. (2017). Coherence between emotions and facial expressions. In J. A. Russell & J. M. F. Dols (Eds.), *The science of facial expression* (pp. 107–132). Oxford University Press.

Educational Testing Service. (2016). *Test content*. TOEFL ITP Assessment Series. https://www.ets.org/toefl_itp/content/

Efron, D. (1941). *Gesture and environment*. King's Crown.

Efron, D. (1972). *Gesture, race and culture*. King's Crown.

Eibl-Eibesfeldt, I. (1972). Similarities and differences between cultures in expressive moments. In R. A. Hinde (Ed.), *Non-verbal communication* (pp. 297–314). Cambridge University Press.

Ekman, P. (1999). Emotional and conversational nonverbal signals. In L. S. Messing & R. Campbell (Eds.), *Gesture, speech, and sign* (pp. 45–56). Oxford University Press.

Ekman, P. (2017). Facial expressions. In J. A. Russell & J. M. F. Dols (Eds.), *The science of facial expression* (pp. 39–56). Oxford University Press.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1), 49–98.

Farroni, T., Menon, E., Rigato, S., & Johnson, M. H. (2007). The perception of facial expressions in newborns. *European Journal of Developmental Psychology*, *4*(1), 2–13. https://doi.org/10.1080/17405620601046832

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science*, *37*(7), 1356–1367. https://doi.org/10.1111/cogs.12048

Fernández-Dols, J.-M., & Crivelli, C. (2013). Emotion and expression: Naturalistic studies. *Emotion Review*, *5*(1), 24–29. https://doi.org/10.1177/1754073912457229

Gifford, R. (2006). Personality and nonverbal behavior. In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 159–179). SAGE Publications.

Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35. https://doi.org/10.1111/j.1745-3992.2006.00076.x

Gruba, P. (1997). The role of video media in listening assessment. *System*, *25*(3), 335–345. https://doi.org/10.1016/S0346-251X(97)00026-2

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition*, *14*(1), 53–82. https://doi.org/10.1075/pc.14.1.05gul

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior*, *33*(4), 251–277. https://doi.org/10.1007/s10919-009-0073-2

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica*, *76*(1), 31–49. https://doi.org/10.1016/0001-6918(91)90052-2

Hardison, D. M. (2018). Effects of contextual and visual cues on spoken language processing: Enhancing L2 perceptual salience through focused training. In S. M. Gass, P. Spinner & J. Behney (Eds.), *Salience in second language acquisition*. Routledge. http://www.taylorfrancis.com/books/9781315399027

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175–1192. https://doi.org/10.1162/jocn.2007.19.7.1175

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, *3*(2), 127–154. https://doi.org/10.1075/gest.3.2.02hol

Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, *33*(2), 73–88. https://doi.org/10.1007/s10919-008-0063-9

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 University Press.

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics*, *37*(5), 1175–1199. https://doi.org/10.1017/S0142716415000533

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, *145*(6), 708–730. https://doi.org/10.1037/xge0000162

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 1151–1160. https://doi.org/10.1145/2638728.2641695

Kellerman, S. (1992). 'I see what you mean': The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied Linguistics*, *13*(3), 239–258. https://doi.org/10.1093/applin/13.3.239

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kida, T. (2008). Does gesture aid discourse comprehension in the L2? In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 131–156). Taylor & Francis Group.

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2014). *Nonverbal communication in human interaction* (8th ed.). Cengage Learning.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception and Psychophysics*, *65*(4), 536–552. https://doi.org/10.3758/BF03194581

Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A microanalytic inquiry. *Language Learning*, *54*(1), 79–117. https://doi.org/10.1111/j.1467-9922.2004.00249.x

Lesnov, R. O. (2018). Content-rich versus content-deficient video-based visuals in L2 academic listening tests: Pilot study. *International Journal of Computer-Assisted Language Learning and Teaching*, *8*(1), 15–30. https://doi.org/10.4018/IJCALLT.2018010102

Magnusson, A.-K. (2006). Nonverbal conversation-regulating signals of the blind adult. *Communication Studies*, *57*(4), 421–433. https://doi.org/10.1080/10510970600946004

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*(2), 97–115. https://doi.org/10.1017/S1366728903001068

Matsumoto, D. (2006). Culture and nonverbal behavior. In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 219–235). SAGE Publications.

McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, *35*(1), 51–73. https://doi.org/10.1177/0265532216677105

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (Ed.). (2000). *Language and gesture*. Cambridge University Press.

Nakatsukasa, K. (2016). Efficacy of recasts and gestures on the acquisition of locative prepositions. *Studies in Second Language Acquisition*, *38*(4), 771–799. https://doi.org/10.1017/S0272263115000467

Namba, S., Makihara, S., Kabir, R. S., Miyatani, M., & Nakao, T. (2017). Spontaneous facial expressions are different from posed facial expressions: Morphological properties and dynamic sequences. *Current Psychology*, *36*(3), 593–605. https://doi.org/10.1007/s12144-016-9448-9

NVivo qualitative data analysis software (10.0.638.0). (n.d.). [Windows]. QSR International Pty Ltd.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517–537. https://doi.org/10.1177/0265532207080771

Pernice, K., & Nielsen, J. (2009). *How to conduct eyetracking studies*. Nielsen Norman Group. https://media.nngroup.com/media/reports/free/How_to_Conduct_Eyetracking_Studies.pdf

Pusey, K., & Lenz, K. (2014). Investigating the interaction of visual input, working memory, and listening comprehension. *Language Education in Asia*, *5*(1), 66–80. https://doi.org/10.5746/LEiA/14/V5/I1/A06/Pusey_Lenz

Riseborough, M. G. (1981). Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal Behavior*, *5*(3), 172–183. https://doi.org/10.1007/BF00986134

Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition*, *35*(2), 213–235. https://doi.org/10.1017/S0272263112000861

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). Taylor and Francis.

Salverda, A. P., & Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1122–1133. https://doi.org/10.1037/a0023101

Sato, R. (2019). Examining the effects of gestures in providing oral corrective feedback. *Electronic Journal of Foreign Language Teaching*, *16*(1), 22–33. https://e-flt.nus.edu.sg/v16n12019/sato.pdf

Sime, D. (2006). What do learners make of teachers' gestures in the language classroom? *International Review of Applied Linguistics in Language Teaching*, *44*(2), 211–230. https://doi.org/10.1515/IRAL.2006.009

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and Language*, *101*(3), 260–277. https://doi.org/10.1016/j.bandl.2007.02.008

Stam, G., & McCafferty, S. G. (2008). Gesture studies and second language acquisition: A review. In S. G. McCafferty & G. Stam (Eds.), *Gesture: Second language acquisition and classroom research* (pp. 3–24). Taylor & Francis Group.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

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* (pp. 53–68). Iowa State University.

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*(4), 463–483. https://doi.org/10.1177/0265532214562099

Suvorov, R. (2018). Test takers' use of visual information in an L2 video-mediated listening test: Evidence from cued retrospective reporting. In E. Wagner & G. J. Ockey (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 145–160). John Benjamins.

Tanenhaus, M. K., & Trueswell, J. C. (2006). Eye movements and spoken language comprehension. In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 863–900). Elsevier. https://doi.org/10.1016/B978-012369374-7/50023-7

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL® ITP scores onto the Common European Framework of Reference* (Research Memorandum RM–11-33; p. 31). Educational Testing Service. http://www.ets.org/Media/Research/pdf/RM-11-33.pdf

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*, *11*(4), 237–244. https://doi.org/10.1037/1076-898X.11.4.237

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL and Applied Linguistics, Teachers College, Columbia University*, *2*(1). https://tesol-dev.journals.cdrs.columbia.edu/wp-content/uploads/sites/12/2015/05/4.-Wagner-2002.pdf

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67–86. https://www.lltjournal.org/item/2604

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly*, *5*(3), 218–243. https://doi.org/10.1080/15434300802213015

Wagner, E. (2010a). Test-takers' interaction with an L2 video listening test. *System*, *38*(2), 280–291. https://doi.org/10.1016/j.system.2010.01.003

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513. https://doi.org/10.1177/0265532209355668

Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, *10*(2), 178–195. https://doi.org/10.1080/15434303.2013.769552

Wang, W., & Loewen, S. (2016). Nonverbal behavior and corrective feedback in nine ESL university-level classrooms. *Language Teaching Research*, *20*(4), 459–478. https://doi.org/10.1177/1362168815577239

Winke, P. (2013). The effects of input enhancement on grammar learning and comprehension: A modified replication of Lee (2007) with eye-movement data. *Studies in Second Language Acquisition*, *35*(2), 323–352. https://doi.org/10.1017/S0272263112000903

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* (No. 2014/3; IELTS Research Reports Online Series). https://www.ielts.org/-/media/research-reports/ielts_online_rr_2014-3.ashx

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38–54. https://doi.org/10.1016/j.asw.2015.05.002

Wu, Y. C., & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, *101*(3), 234–245. https://doi.org/10.1016/j.bandl.2006.12.003

Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, *43*(2), 303–311. https://doi.org/10.1016/j.jesp.2006.02.004