THE IMPACT OF VISUAL CUES ON ITEM RESPONSE IN VIDEO-MEDIATED TESTS OF FOREIGN LANGUAGE LISTENING COMPREHENSION

Aaron Olaf Batty

B.A. (*cum laude*), Colorado State University, USA
M.A. (with distinction), Colorado State University, USA

Thesis submitted to
The Department of Linguistics and English Language Lancaster University, UK,
for the degree of Doctor of Philosophy

September 2016

Abstract

The present thesis employed a mixed-methods research design spanning over two studies and an intermediate instrument development step to investigate the interactions between the presence of nonverbal and other visual cues with examinees, individual items, and item task types in video listening tests. The first study employed eye-tracking methodology to determine the specific visual cues to which examinees attend when interacting with a video-mediated listening test. The findings of Study I then informed the development of a new video-mediated listening test for Study II, which investigated the effect of the presence of visual cues on item and item task difficulty via many-facet Rasch modeling and qualitative item analysis. Individual examinee differences were also explored in both studies with respect to gender, proficiency, and perceptions of the two formats.

Study I found that examinees view facial cues an average of 81% of the time regardless of task type, but spend significantly more time oriented toward the listener (i.e., the character who is not speaking) when presented with an implicit item. Direct viewing of gestures, despite their prominent place in the nonverbal communication literature, only accounted for approximately 1.35% of the total video time. Study II found that the presence of visual information exerted a facilitative effect on all items, but that it was significantly more pronounced with implicit items, despite the fact that they were more difficult than the explicit items in the audio format. Finally, no substantive interactions between proficiency, gender, or perception of the formats were observed between either viewing behavior or the facilitative effect of video.

The thesis ultimately raises questions about the usefulness of video listening tests for listening comprehension assessment, as the effect appears to do little more

that raise scores, and, in the case of implicit items, may obviate the need to comprehend the linguistic input in order to make correct inferences.

Acknowledgments

I would like to express my deep appreciation for my supervisor, Luke Harding, who has been a joy to work with from beginning to end. His ideas propelled me in interesting new directions, and his comments were always insightful and encouraging. It has been a great pleasure working with him. I would also like to thank Tineke Brunfaut, whose trenchant and detailed comments on my confirmation document guided my revision and expansion of that material into a full thesis. I would like to thank my colleagues Yoko Hasebe, David O'Donnell, Juanita Heigham, Keith Olfers, Stephen Hofstee, Tim Hemington, Mimi Ohta, and Thomas Hardy for generously lending me their time and energy during data collection for my studies, and to Yuki Kishi, my research assistant and practice subject for Study I. I must also thank Jeffrey Durand, Kathrin Eberharter, Luke Harding, Franz Holzknecht, Yo In'nami, Glyn Jones, Rie Koizumi, Benjamin Kremmel, David O'Donnell, Gary Ockey, Anchana Rukthong, and my frequent research partner Jeffrey Stewart for taking part in item moderation for the Study II instrument. Special thanks also Nerida Rand for casting and performing in the Study II videos, and to David Mashiko, Kim Tierney, and Bob Werley for their performances. Finally, and most importantly, I extend my deepest and most heartfelt gratitude to my wife, Sonoko Kasagi, who not only devoted many hours to checking my translations, assisting with data collection, scouting locations, operating sound equipment, and driving during the filming of scenes for Study II; but—more importantly—endured with grace and patience the countless lonely days and nights during which her husband disappeared into his home office, emerging only to eat and sleep. Without her support at every step, this project simply would not have happened.

Declaration

This thesis has been written by me, and the work reported in it is my own. It has not been previously submitted for the award of degree in this or any other institution.

Aaron Olaf Batty

30 September 2016

Table of Contents

List of Tables

List of Figures

CHAPTER 1. INTRODUCTION

1.1. Background

Nonverbal communication is well-established as a major component of communicative listening in humans, comprising somewhere between 66% and 70% of the information in any face-to-face interaction (Burgoon, 1994; Burgoon, Guerrero, & Floyd, 2016). It appears earlier in our physical/cognitive development (Burgoon et al., 2016), and it has been argued to represent communication in its most primitive state (McNeill, 1992). Nonverbal communication scholars Bavelas and Chovil argue that:

> … listener gaze occurs in large part because what the speaker is "saying" is not just auditory but also visual. Listeners need to see speakers' hands and faces as well as hear their words. (2000, p. 170)

Facial expressions in particular have been convincingly argued to constitute a kind of "paralanguage" (Chovil, 1991a), while gestures' integral link to language has been demonstrated by their synchronous timing with relevant words and phrases, with the use of unfamiliar words, which take longer to recall and process, being accompanied by similarly slowed gestural support (Morrel-Samuels & Krauss, 1992).

Nonverbal communication in the form of gestures can be used to effectively communicate messages even without comprehensible attendant verbal input or knowledge of sign language (Fay, Arbib, & Garrod, 2013), and, in cases where comprehension of verbal input is low, access to such information can bring comprehension levels nearly to those of native speakers (Dahl & Ludvigsen, 2014). Finally, of interest to anyone working in the field of language assessment, nonverbal communication is an important component of cross-cultural communication, with cultures sometimes employing subtly different gestural conventions, which can lead to confusion or embarrassment (Burgoon et al., 2016).

Beyond the wealth of information to be found in nonverbal communication, there is the comparatively more prosaic, but nonetheless important, information granted by context. When listening, the visual channel can aid in setting expectations for the kind of information or interaction to follow, by tapping into the listener's pre-existing knowledge of scripts (patterns of behavior and roles which are well-known and well-understood) and schemata (archetypes of people, places, etc.), allowing the listener to spend less time and energy "catching up" to the content of the stimulus, and devote more to the comprehension of the verbal channel (Buck, 2001).

1.1.1. Video-Mediated Listening Tests

Citing these facts about the importance of nonverbal communication and other visual cues, many have argued that they should feature in tests of foreign language listening comprehension by way of video, rather than audio-only, listening passages (e.g., Gruba, 1997; Ockey, 2007; Wagner, 2002). Such tests would more closely resemble the target language use (TLU) domain (Bachman & Palmer, 1996), resulting in a more authentic test task, which has long been held to be an important part of test construct validity (Messick, 1989). In fact, some widely-used language tests, most notably the TOEFL ITP, have incorporated "slideshows" of still images in the hopes of capturing some of this information. Unfortunately, however, even that practice may not take adequate advantage of the visual channel, as it has been demonstrated that we do not "read" nonverbal communicative cues in short flashes, but in "thin slices"—chunks of time from thirty seconds to up to five minutes—before we reach a decision on what the nonverbal message truly is (Ambady, LaPlante, & Johnson, 2001). Given this finding, it becomes all the clearer why the use of video in such tests would more closely represent actual language use cases, and with that, confer greater test validity.

There has been fairly consistent interest in the topic of video-mediated foreign language listening tests for the past several decades; however, serious gaps remain. To date, research on this topic has largely relied upon total-score comparisons between audio-only and equivalent audio-video instruments (e.g., Baltova, 1994; Batty, 2015; Brett, 1997; Chung, 1994; Coniam, 2001; Cubilo & Winke, 2013; Gruba, 1993; Hernandez, 2004; Londe, 2009; Parry & Meredith, 1984; Shin, 1998; Sueyoshi & Hardison, 2005; Suvorov, 2009, 2013, Wagner, 2006, 2010b, 2013), but many of the studies suffer serious methodological deficiencies that undermine their findings. Even, however, in the best-designed studies, many have small *N* sizes, or feature item formats which are not well-suited to large-scale use (e.g., limited production, essay). Despite the long history of work on the topic, seemingly simple questions such as "which is harder, audio or video" have not been answered conclusively.

Beyond such comparative studies, others have investigated examinee interaction with and use of the visual channel in such tests, whether in terms of time spent oriented toward the video (Ockey, 2007; Wagner, 2007, 2010a), or examinees' perceptions of their own attention to visual cues (i.e., specific, individual behaviors or features in the visual channel) while taking the test via retrospective thinkaloud or recall protocols (Ockey, 2007; Suvorov, 2013; Wagner, 2008). Studies investigating the amount of time examinees orient toward the video can typically only report general statistics on percentage of time with their heads looking up and forward, and report a wide range of such percentages. Studies attempting to determine the specific visual cues to which examinees refer have required the participants to recall and report their own unconscious behavior, which may be fallible.

Recent work by Suvorov (2013, 2015), however, has attempted to further increase the resolution at which the question of how much examinees use the visual

channel in video-mediated listening tests is examined by employing eye-tracking methodology. However, his study, although capable of presenting higher-detailed and more theoretically-defensible measures of time spent oriented toward the video, does little to address the question of what, exactly, examinees are looking at, as the only data collected to address this question was interview data, similar to previous studies. Furthermore, the main interest of the Suvorov studies was in the differences in the amount of time spent looking at contextual video (i.e., a lecturer speaking with no visual aids) and content video (e.g., a presentation slide), rather than the more fundamental questions addressed by the present thesis.

1.1.2. Remaining Questions

As suggested above, despite the long history of research on the topic of video listening tests, some critical gaps in knowledge remain:

It is unclear what information is added to listening tests when video is included. Although there exists a great deal of theory on the informational content of the visual channel in human communication, there is very little research that attempts to determine how much and what kinds are actively used by examinees in video listening tests.

It is unclear how information from the visual channel interacts with test items and item task types. Listening items can be broadly categorized into two task types: "explicit" items, requiring the examinee only to comprehend the unambiguous, propositional content of the stimulus, and "implicit" items, which require the examinee to make correct inferences of information not expressly stated in the content. It seems likely that these two task types would interact with the inclusion of nonverbal or other visual information differently. Curiously, however, no published

studies have sought to investigate how the visual information in video listening tests impacts responses to items of differing task types.

It is unclear whether the visual channel is accessed and utilized similarly by all examinees. Without a closer inspection of individual examinees' interaction with video-mediated listening tests, it will be impossible to know whether such tests assess the same traits for everyone in the tested population.

It is unclear whether any extra information gleaned from video-mediated listening tests represents a useful addition to listening comprehension measures. The most critical question that has dogged video-mediated listening comprehension research since its outset has been one of construct validity. The question of whether audio-only tests are equivalent to video-only tests, or, if they are not, whether the scores from the latter are more informative to test users than those from audio-only measures is one that has been particularly difficult to resolve.

If the interactions between examinees, visual cues, task types, and test items were better understood, video could be more confidently incorporated into listening comprehension assessment in a theoretically-defensible way, with an understanding of the construct implications of the incorporation of video on such tests. An expansion of the knowledge on this topic could guide language testers in their use of video for nonverbal communication in test design, item writing, and test scoring.

1.2. Purpose

This thesis seeks to fill the above gaps in knowledge and practice by employing a fixed mixed-methods exploratory sequential design (Creswell & Plano Clark, 2010) spanning two full studies, with an intermediate instrument development step. The first study aims to determine the nonverbal and visual cues most viewed by examinees in a foreign language listening comprehension test, addressing this question both at the test

and task levels, using eye-tracking technology for quantitative measures and structured interviews based on gaze displays for qualitative data collection and analysis. The findings of Study I are then used to inform the development of a video-mediated foreign language listening comprehension test. Finally, the resulting test is administered in two formats (audio-only and audio-video; Study II), and the results are analyzed with many-facet Rasch modeling to investigate the impact of the presence of visual and nonverbal cues on item difficulty, and the difficulty of two broad item task types: explicit and implicit.

As individual examinee differences may also play a part in any observed effect of the presence of nonverbal or other visual cues on listening tests, whether it be due to natural differences in sensitivity to such extralinguistic information (Costanzo & Archer, 1989; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979), proficiency (Batty, 2015; Sueyoshi & Hardison, 2005), or gender (Burgoon, 1994; Costanzo & Archer, 1989; Noller, 1985; Rosenthal et al., 1979), these individual differences will also be considered. Finally, examinee perceptions of the formats and their interactions with the effect of video will be investigated using survey and test item response data.

1.3. Organization of the Thesis

The thesis is presented in seven chapters. The first chapter (the present chapter) provides a background to the thesis, states the purpose, and presents an outline of the overall organization.

The second chapter, Review of Literature, presents the theoretical background to the work. It opens with a brief introduction to listening comprehension, then moves to a detailed section on the role of visual information in listening comprehension. This section first presents an overview of nonverbal behavior and its importance to human communication, then moves to its importance to listening comprehension in

particular. The chapter then applies these topics to the realm of second language acquisition, moving from studies in the field which addressed nonverbal communication to foreign language listening comprehension tests in particular.

A detailed review of the many video-mediated listening test studies carried out since the early 1980s follows, beginning with those of most importance to the present work: comparative studies between audio-only and video-mediated listening tests. A detailed critique of the methodologies of these studies is then presented, demonstrating the research gap for the present work. The comparative section concludes with a review of the present author's previous published work in this vein.

Additional themes appearing in the video listening test literature are then considered, and further gaps are identified. The section following reviews the literature on task types in listening comprehension tests. The chapter concludes by outlining the concept of test validity that will be used throughout the thesis.

Chapter 3 sets forth a concise rationale for the studies carried out in this thesis and presents the likely implications of the findings for test developers. The research questions are then presented. A thorough overview of the overall research design, as well as that of the studies therein, concludes the chapter.

Chapter 4 reports on Study I, an eye-tracking study of examinee visual interaction with a video-mediated listening test. It begins with a review of the eye-tracking literature, from a discussion of the movements and metrics, through its use in the field of second language assessment, to serve as a foundation for the study report to follow. The study and its results are described and discussed, with particular attention paid to the implications for the design of Study II.

Chapter 5 describes the development of the instrument for Study II, beginning with a brief description of the test specifications, then the process of item

development and video production. The process of item moderation, by which the classifications of items into either “explicit” or “implicit” categories was carried out, is then reported. Finally, the chapter details the process of constructing the final instrument to be administered to the Study II sample.

Chapter 6 reports on Study II, which sought to quantify any effect attributable to the presence of visual information on item and task type difficulty. The chapter opens with a description of the method, followed by a detailed explanation of the data analysis methods used in Study II. Of particular importance is this section’s elucidation of the Rasch models employed. The results are then reported and discussed.

The seventh and final chapter is devoted to a discussion of the findings of both Studies I and II. It begins with a brief summary of the findings with regards to the research questions, then discusses implications of the findings. Both theoretical implications, comprising those of the construct validity of video-mediated listening tests as well as test use cases for the format, and practical implications are addressed. The chapter then provides recommendations for language assessors who may be interested in developing video-mediated listening tests. Finally, limitations of the studies in the thesis are briefly discussed in the context of recommendations for future work on this topic.

1.4. Chapter Summary

This chapter has briefly introduced the background for the present thesis and presented an overview of the organization of the remainder. The next chapter reviews the literature relevant to the topic of video-mediated listening tests.

CHAPTER 2. REVIEW OF LITERATURE

This section offers a review of the literature concerning the assessment of foreign language listening comprehension with the inclusion of nonverbal cues in the visual stream. It focuses on the overall themes and issues which form the motivation of and backdrop to the studies carried out and reported in Chapters 4 through 6, with topics relevant only to Study I or II reviewed within their respective chapters.

This chapter opens with a brief introduction to the construct of listening comprehension, then moves to a focused review of the literature pertaining to nonverbal behavior from the fields of psychology and communication, with particular attention paid to the classification and descriptions of the nonverbal cues that will feature in Study I (Chapter 4). The next section turns its focus back to second language listening comprehension, demonstrating the importance of nonverbal and other visual information to second language comprehension.

Section 2.4 presents a detailed review of the research on video in foreign language listening tests, beginning with a close reading of the studies comparing audio-only to video-mediated listening tests to serve as a historical backdrop to the issues investigated in this thesis, especially those addressed by Study II (Chapters 5 and 6). Following this overview is a section of criticisms of this work, highlighting methodological concerns that raise questions about the veracity or generalizability of their findings, and missed opportunities for greater understanding of the issues. The author's own published work on this topic (Batty, 2015), which attempted to address some of these issues, and which serves as a pilot study to those contained in this thesis, is then reviewed with an eye toward further instrument and research design issues to be addressed by the present studies. Section 2.4 concludes with a review of

additional themes explored by the video listening test literature, which motivate other aspects of the present thesis.

Section 2.5 reviews the literature on item task, providing the framework for the implicit/explicit delineation to be used throughout the studies herein. The final section presents the various conceptualizations of test validity through the history of psychometrics as a field, as questions of validity have dogged the discussion of video-mediated listening tests from the outset, and feature prominently in the rationale for the present work. The chapter concludes with a brief summary of the chapter.

2.1. Listening Comprehension

Second- or foreign-language listening comprehension received relatively little research attention until the late 1970s with the advent of communicative language teaching and its attendant shift in focus from learning and teaching *about* a language to one more relevant to *using* the language (Flowerdew & Miller, 2010; Taylor, 2013). It is typically understood to involve the processing of auditory input into a mental representation of its meaning (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Taylor, 2013), and has also frequently been noted as being the most difficult to describe or target for assessment because it is invisible, occurring entirely within the listener's mind (Brindley, 1998; Lynch, 1998). Buck succinctly described the construct of listening comprehension for the purpose of assessment as the following:

> It is the ability:
>
> - To process extended samples of realistic spoken language automatically and in real time,
> - To understand the linguistic information that is unequivocally included in the text, and

- To make whatever inferences are unambiguously implicated by the content of the passage.

(2001, p. 114)

Most definitions of listening skill focus solely on what listeners do with the auditory input (e.g., Wolvin & Coakley, 1996), but beyond the obvious informational input of sound waves, listeners derive and enrich meaning through other channels such as physical setting, cultural expectations, and non-verbal cues such as gestures or facial expressions (Brindley, 1998; Buck, 2001; Lynch, 1998). The process of extracting meaning from aural input is both unconscious and highly complex, and despite a generation of research into second- or foreign-language listening comprehension, no single agreed-upon definition of it has been adopted (Bejar et al., 2000; Gruba, 2006). However, the descriptions appearing in the recent literature are not without their similarities.

Rost (2011) broke his definition of the process of listening down into four sets of sub-processes: neurological, linguistic, semantic, and pragmatic. In the neurological stage, the listener scans the environment for meaningful sounds and turns his or her attention to them if they appear relevant. Imhof's cognitive model of listening features a "selection" stage that is essentially equivalent, and describes listening as "the process of selecting, organizing, and integrating information" (2010, p. 98). It is the selection phase wherein we (typically unconsciously) determine what is worthy of our attention, such as our name being used in a conversation across even a crowded, noisy room. Attention is cognitively resource-intensive, so it is only tuned to stimuli deemed worth the effort (Rost, 2011). When our name is used across the crowded room in the example above, it is our current conversation partner who suffers, as we drop the

thread of whatever we were talking about with that person to attend, instead, to the people elsewhere talking about us.

Rost's next stage of processing is the linguistic stage, and is the first stage wherein the language resources of the listener come into play (2011). Both this Rost stage and the next correspond to the "organization" stage in the Imhof model (2010). This stage is the "bottom-up" comprehension process, wherein raw aural data from the auditory channel is organized into language in order to begin constructing meaning, and was the main model for the entire listening comprehension process through at least the 1950s (Flowerdew & Miller, 2010), with researchers believing that listeners began with the smallest pieces of verbal information—the individual phonemes in the spoken stimuli—and combined them into ever-larger and more-meaningful structures until the message was comprehended. Later researchers, however, came to understand that bottom-up processing does not represent the entire comprehension process, but is only one tool used by the listener to construct a complete understanding of the aural input (Rost, 2011).

The next stage in the Rost (2011) model of listening comprehension is the semantic stage, which also corresponds to Imhof's "organization" stage (2010). This step can be described as one of "top-down" processing, and was developed after the bottom-up model described above (Flowerdew & Miller, 2010). Top-down processing relies heavily on the listener's memory (Imhof, 2010; Wolvin, 2010), linking the information available in the verbal (and nonverbal) data streams to known patterns (e.g., vocabulary, syntax, idiom) in memory in order to "find coherence and relevance" (Rost, 2011, p. 53). In addition to linguistic resources, this stage includes increased reliance on knowledge and assumptions shared between the listener and the speaker (Rost, 2011), as the listener forms inferences about the speaker's message,

blending this stage also with that described as “integration” in the Imhof cognitive model (2010).

Finally, Rost describes a pragmatic processing stage, wherein the listener draws from the situation and other cues in order to “enrich” the speaker’s meaning (2011, p. 85). This kind of extralinguistic facilitative information is that which Imhof places within the “integration” stage of her cognitive model of listening comprehension (2010), and includes information about the speaker’s emotional state, the situation, and the social expectations attendant to them (Rost, 2011).

Imhof includes one additional stage, “monitoring,” which describes the process by which listeners continue to cycle back through their expectations to continually build and refine their construction of the meaning of the message (2010). This integration of bottom-up and top-down processing to construct meaning represents a kind of parallel processing, drawing from all of the listener’s linguistic and situational resources to accurately comprehend the spoken message (Flowerdew & Miller, 2010), “based on whatever information seems relevant at the time” (Buck, 2001, p. 29).

Wolvin and Coakley have concisely summed up all models of listening as “the process of receiving, attending to, and assigning meaning to aural and visual stimuli” (1996, p. 69), easily accommodating the more detailed models of listening discussed above, and therefore serving as a simple and convenient description by which to understand the complex process of listening comprehension. Furthermore, its explicit inclusion of visual information as a key component of listening, where Imhof makes no mention of it and Rost (2011) treats it as a separate “co-text” (p. 52) in addition to the aural input, is of clear relevance to the present research.

Given the growing recognition of the importance of visual information to the process of listening comprehension, an increasing number of language assessment researchers have turned their attention to the impact of nonverbal or other visual information on comprehension in foreign language listening tests. The psychological and communication literature describing the importance of this information to language comprehension will be reviewed in the next section.

2.2. Nonverbal Behavior

This section provides a focused introduction to the literature on nonverbal communication. It begins with an overview of the communicative functions served by nonverbal behavior, then addresses the two main categories of nonverbal behavior examined in the present research: facial expressions and gestures.

2.2.1. Communicative Functions of Nonverbal Behavior

Listening rarely occurs "in the dark." For sighted people, in virtually all situations and circumstances, listeners are able to glean extra-linguistic information from such sources as facial expressions, gestures, and the setting, with these visual signals operating as a "co-text" to the verbal signals (Rost, 2011, p. 37), as some second-language assessment researchers have noted (e.g., Bejar et al., 2000; Rost, 1990; Rubin, 1994, 1995). The range and depth of literature on nonverbal communication are vast, and largely well outside the scope of the present research, but a brief introduction to the broad themes and relevant findings is in order. Nonverbal behavior has been of interest for centuries, but much of the current work on it draws from seminal work completed in the middle of the last century (Knapp, 2006), with later work largely applying the frameworks established then to newer research questions (Fridlund & Russell, 2006).

Nonverbal behavior can be either involuntarily or intentionally communicative (Russell, Fernández-Dols, & Chovil, 1997), and can be redundant or not with the verbal output (Chovil, 1991a). It can be used to deliberately signify our thoughts or feelings, and frequently changes with the audience (Carroll & Russell, 1996; Chovil, 1991b). When used deliberately, it can be used for "ceremony and ritual, persuasion and propaganda" (Graham & Argyle, 1975, p. 57). It can also be used to modify or augment the verbal channel, amplifying the message (e.g., by leaning in while speaking), augmenting it (e.g., making a pained expression when refusing a request to imply that one would rather comply) or contradicting it (e.g., using "air quotes" to indicate that one does not agree with the use of a particular word) (Scherer, 1980). In general, the face is thought to relay the affective state and the body, its intensity (Ekman, 1984), although gestures of the hands and arms can also be argued to operate very similarly to verbal language (McNeill, 1992).

As a listener, nonverbal behavior allows us to support and check our understanding of the verbal channel. Burgoon argues that nonverbal communication is the primary form of communication for our species—developing earlier in our evolution, as well as in our bodily development (1994; Burgoon et al., 2016). In her review of the relevant literature, she finds that between 66% and 70% of the meaning in a social situation is encoded in the visual, as opposed to the verbal, channel. Listeners watch those who are speaking to them because they understand that what people are "saying" is not just verbal, but visual as well (Bavelas & Chovil, 2000).

Nonverbal behavior is particularly useful for revealing the speaker's attitudes, emotions, personality, and interpersonal relations (Graham & Argyle, 1975). Listeners tend to watch speakers' movements over time, taking in a range of behaviors rather than relying on single observations, to gain these insights (Ambady et al., 2001).

Nonverbal behavior 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 (Burgoon, 1994; Burgoon et al., 2016; Folger & Woodall, 1982). Although cultures, genders, and individuals may be sensitive to different cues to differing degrees (Costanzo & Archer, 1989; Noller, 1985; Rosenthal et al., 1979), adults tend to rely on them more (Burgoon, 1994; Burgoon et al., 2016). Nonverbal communication is an important carrier of social meaning in human interactions such as conversations, but may be ignored if it is found to be irrelevant to the central task or message (Weidenmann, 1989; Burgoon, 1994; Burgoon et al., 2016). By contrast, the verbal channel is typically favored when the content is factual, abstract, or persuasive in nature (Burgoon, 1994).

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 itself was based on the earlier work of Efron (1941), and updated by Ekman in 1999. The framework laid out by Ekman and Friesen remains highly influential (Burgoon et al., 2016), and later taxonomies typically add further levels of detail (e.g., Chovil, 1991a; McNeill, 1992), retaining many of the same terms or essential classifications from Ekman and Friesen. It remains the most well-known classification of nonverbal behaviors (Kendon, 2004), and it seems likely that the reason for this is its relative simplicity, providing a useful framework for further general exploration of the topic.

For the purposes of the present research, which seeks to investigate to what nonverbal behaviors examinees attend, and how this impacts performance, the Ekman and Friesen categories are both clear enough and broad enough to facilitate that investigation. In the present research, nonverbal behavior will be broken into two broad categories: facial expressions and gestures, each described in turn below.

2.2.2. Facial Expressions

Facial expressions (also classified as "affect displays," or "emotional expressions") are probably the most important carrier of nonverbal information, and have been repeatedly found to be essentially universal to the human species (Eibl-Eibesfeldt, 1972; Ekman, 1999; Ekman & Friesen, 1969; Ekman, Friesen, & Ellsworth, 1972; Floyd, 2006; Matsumoto, 2006), even appearing among children who were born blind (Fulcher, 1942; Thompson, 1941). Humans even share some facial expressions with other primates (Eibl-Eibesfeldt, 1972). Most of the work on this topic has focused on what individual expressions mean, which is of little direct relevance to the research at hand; however, a brief overview of the general findings of this enormous vein of the psychological literature will serve to conceptualize the types of information contained in the nonverbal cues to be studied in later chapters.

2.2.2.1. Facial expressions and emotion

Most work on facial expressions adheres to the so-called "Facial Expression Program," which is a set of assumptions about what facial expressions mean. Although it is not explicitly stated or codified anywhere, it can be summed up as an understanding that there are a handful of basic emotions (ranging from five to nine), for which each has but one single facial expression, and that those expressions were laid down and "hard-wired" during the process of natural selection, due to the effective communication of these emotional states being critical to survival. In fact, much of this work harkens back to Darwin's writing on the subject (Russell & Fernández-Dols, 1997). The most well-known list of basic emotions is surprise, fear, disgust, anger, happiness, and sadness, laid down originally by Ekman and Friesen (1975/2003), although others have argued for the addition of contempt and shame, and for surprise and fear being combined. The Facial Expression Program recognizes any

other emotions or psychological states to be a mixture of these basic emotions (Russell & Fernández-Dols, 1997). At least part of the reason for the broad application of the assumptions of the Facial Expression Program is that it includes methods for carrying out studies of encoding (i.e., production) and decoding (i.e., interpretation) of facial expressions via Ekman and Friesen's (1978) Facial Action Coding System, which has been used in a multitude of facial expression studies (Rosenberg, 1997)

Contrary to the assumptions of the Facial Expression Program, however, continued work on people's decoding of facial expressions has revealed only moderate correlations between the expressions and people's interpretation of them (Parkinson, 2005), leading researchers to soften their claims of a strict 1:1 relationship between expressions and emotions, considering them, rather, to be "loosely coupled" (Knapp, Hall, & Horgan, 2014, p. 267). Even Ekman has adopted the stance that "for each emotion there is not just one facial expression but a family of them" (Ekman, 1999, p. 52). These expressions are influenced by "intensity, efforts to control the expression, and perhaps also the particular form of the emotion" (p. 52).

This newer conception of the relationship between emotion and facial expression is often referred to as "Minimal Universality" (Fridlund & Russell, 2006; Russell, 1995; Russell & Fernández-Dols, 1997). This model takes into account cultural and personal differences observed in the research, and makes a series of careful assumptions, including "most people everywhere can infer something of another's psychological state from facial movement," and that "people everywhere can form nonrandom association between faces and emotions" (Fridlund & Russell, 2006, p. 307). However, it should be pointed out that much of the research on facial expressions relies on still images of "highly selected, posed, melodramatic faces of unknown ecological validity" (p. 307), despite the fact that others have noted that

people do not rely on static information such as still frames to "read" nonverbal behavior in normal conversation, and instead observe behavior over time in order to reach a conclusion about the psychological state of the speaker (Ambady et al., 2001).

Facial expressions' meanings have also been demonstrated to be highly contextual, meaning different things at different times (Bavelas & Chovil, 2000; Carroll & Russell, 1996; Chovil, 1991a, 1991b; Fernández-Dols & Carroll, 1997; Fridlund & Russell, 2006; Russell et al., 1997). A cropped photo of an apparently grief-stricken woman can be revealed to be one who is overjoyed to the point of tears when the remainder of the photo makes it clear that she is winning a gold medal at the Olympics, for example (Fernández-Dols & Carroll, 1997). This is partly because facial expressions are paralinguistic, and are sometimes used intentionally to communicate extra information beyond what is included in the verbal stream, as a full component of the conversation (Bavelas & Chovil, 2000; Chovil, 1991a; Russell et al., 1997). They can be used intentionally 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, 1991a). Furthermore, facial expressions are frequently used to mirror those of an interlocutor in order to signal empathy and build rapport (Chovil, 1991b). Clearly, the relationship between emotion and facial expressions is somewhat complex, but suffice it to say that facial expressions are an important component of the communication of emotional or psychological states, upon which we all rely in face-to-face communication.

2.2.2.2. Differences in facial expression encoding and decoding

Not all people or peoples encode or decode nonverbal behavior, including facial expressions, in the same ways. In the context of second language listening

assessment, the issue of cultures' differing use of facial expressions obviously warrants discussion.

Generally speaking, cultures' uses of facial expressions can differ to some extent, but these differences do not appear to lie in the expressions themselves as much as they do in the cultural "display rules" governing how much and when to express emotions in certain social contexts (Ekman & Friesen, 1969; Ekman et al., 1972; Matsumoto, 1990, 2006; Matsumoto et al., 2002; Matsumoto, Kazri, & Kooken, 1999; Matsumoto, Takeuchi, Andayani, Kouznetsova, & Krupp, 1998). For example, American culture places a high degree of value on appearing happy, and signaling as such with a broad smile. As a result, Americans, especially when interacting with other Americans, amplify their emotional display of happiness beyond what they are actually feeling (Matsumoto, 2006). On the contrary, for example, Japanese de-amplify felt emotions across the spectrum by reducing facial expressions, as the display of strong emotions is somewhat looked down upon in Japan (Matsumoto, Yoo, Hirayama, & Petrova, 2005).

These display rules also influence how listeners from various cultures judge the emotion of speakers, with listeners from individualistic cultures like those of the West rating actual, felt emotion lower than those from more collectivist cultures. In the case of the former, the listener assumes amplification, and the latter assumes de-amplification of felt emotions—essentially applying the listener's cultural display rules even to those of other cultures (Matsumoto et al., 2002, 1999).

Probably due to these differences in cultural display rules, cultural differences have been observed in the parts of the face that are given the most attention when making judgments of felt emotion. Those from East Asian cultures have been found to pay closer attention to the eyes than Westerners, who pay more attention to other parts

of the face, likely due, in large part, to the display rules of the cultures—the eyes being much harder to control, and therefore more communicative of actual felt emotional states (Jack, Garrod, Yu, Caldara, & Schyns, 2012; Yuki, Maddux, & Masuda, 2007).

Finally, differences exist between individuals in their ability to decode facial expressions, with women typically being more adept at decoding than men (Burgoon, 1994; Burgoon et al., 2016; Costanzo & Archer, 1989; Hall, 2006; Noller, 1985; Rosenthal et al., 1979). Age is another factor. Both the very young and the very old are less capable of decoding nonverbal behavior in general (Burgoon et al., 2016; Feldman & Tyler, 2006). Furthermore, personality differences can play a part in one's ability to decode (Gifford, 2006), with those who are "sociable, nonanxious, publicly selfconscious, empathic, independent, psychologically flexible, and intellectually efficient" being better at both encoding and decoding nonverbal behavior than those with other personality traits (Burgoon et al., 2016, p. 26). The construct implications of this with regard to video-mediated listening tests will be addressed in the Conclusion chapter.

2.2.3. Gestures

In addition to facial expressions, the other main form of nonverbal communication is gestures, which have been the focus of most research into nonverbal communication in recent years (Knapp, 2006). According to Kendon, a prominent researcher on the topic, the reason for this interest is largely due to the demonstrated close association between verbal speech and the use of gesture, and its importance to the act of communication (2004). Kendon defines gesture as "actions that have the features of manifest deliberate expressiveness" (p. 7), but in practice, most work focuses on movements of the hands and arms. This section will first give an introduction to

gestures and their importance to human communication, and then describe the classifications of gestures to be used for the remainder of the thesis.

2.2.3.1. The importance of gesture to communication

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, 1994; Burgoon et al., 2016). In fact, gesture alone has been found to be more communicative than gesture accompanied by non-linguistic (i.e., nonsense, gibberish) vocalizations in communicating novel information to a naïve interlocutor (Fay et al., 2013), something to which anyone who has played the party game "charades" can attest.

A further study in this vein, one with clear and serious implications for the present work, is that of Dahl and Ludvigsen (2014). The researchers presented both L1 and L2 English speakers with a video of a person describing a cartoon image that the listeners could not see. Half of each group was able to see the speaker's gestures as they spoke; the other half could not. The listeners were then to attempt to draw the described cartoon image, and the drawings were rated for similarity to the original picture. The L1 participants fared no better or worse with or without gestures. The L2 participants, however, performed significantly and substantially better with the addition of gestures [$t(44) = 3.52$, $p < 0.01$]. In fact, the effect was so pronounced that there was no significant difference between L2 and L1 performances in the gesture condition [$t(38) = -1.83$, $p = .07$]. Such is the impact of gesture on communication, even in cases where the spoken language is not well understood.

Gestures also allow speakers to clarify ambiguous speech by intentionally providing extra information for the benefit of the speaker (Holler & Beattie, 2003), and they even appear to be so closely linked to lexical access that speakers take longer

to form gestures when recalling unfamiliar words (Morrel-Samuels & Krauss, 1992). They appear to be an intrinsic part of human communication, and processed as a component of language, rather than simply an addition to it.

2.2.3.2. Types of gesture

Gestures can be placed on a "Kendon's continuum" from less- to more-language-like, ranging from unconscious, non-conventional movements while speaking, through full sign language (McNeill, 1992, expanded in 2000), but although several classification schemes exist, they are always somewhat arbitrary (Kendon, 2004). However, the categorization first suggested by Ekman and Friesen (1969), based in part on the work of Efron (1941), is perhaps the most widely known (Burgoon et al., 2016; Kendon, 2004), and provides a well-established framework by which to classify and delineate gestural behaviors in the present research. The present work, then, will treat gestures as members of two broad categories: emblems and illustrators, to be described separately below.

2.2.3.2.1. Emblems

Gestures typically referred to as "emblems" (orig. Efron, 1941) are very near the "sign language" end of Kendon's continuum in terms of language likeness. These gestures each have a specific, semantic meanings. These meanings, much like those of words, are not immediately retrievable from the gesture 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. As such, emblems are necessarily culturally bound (Burgoon, 1994; Ekman, 1999; Ekman & Friesen, 1969; Ricci Bitti & Poggi, 1991), resulting in language-like differences in meaning even for the same

gesture. To complete the example, the "okay" gesture does not mean "okay" in all cultures. It is, in fact, an obscene gesture in parts of Europe, and a variant of it (with the palm turned toward the presenter, rather than the recipient) has the meaning of "money" in several East Asian cultures. Emblems are so language-like that what amount to multilingual dictionaries have been written to catalogue and define them for various cultures (e.g., Morris, 1977, 1994). Most gestures, however, have more immediately-understandable meanings, and fall into the next broad category.

2.2.3.2.2. Illustrators

This category 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), used to help organize utterances (Ekman, 1999; Ekman & Friesen, 1969; Kellerman, 1992) and aid difficult-to-understand utterances (Graham & Argyle, 1975; Holler & Beattie, 2003; Riseborough, 1981). Ekman and Friesen described them as "movements which are directly tied to speech, serving to illustrate what is being said verbally" (1969, p. 68). Although he proposed a detailed breakdown of what he called "gesticulations," it was essentially illustrators within the Ekman and Friesen framework that McNeill described as "unwitting accompaniments of speech" (1992, p. 72), and believes to be utterances in their most primitive state.

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 an illustrator.

There in fact exist two further forms of gesture in the Ekman and Friesen (1969) taxonomy: "adaptors" (renamed "manipulators" in Ekman, 1999) which are forms of self-touching (e.g., scratching one's nose), but these are not very meaningful; 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 both of these groups are rather minor, and the research questions pertain much more to the presence or absence of all nonverbal behavior, the present research will not address them separately.

2.3. The Role of Visuals in Second Language Acquisition

Since nonverbal communication is so important to human communication, various language researchers have investigated its impact on the learning of a foreign language, especially since the advent of the videocassette recorder (VCR). Whereas in previous generations, the only option beyond the teacher him/herself for foreign language listening instruction was audiotape, the VCR (and all video technologies succeeding it) allowed teachers to incorporate visual information into their listening lessons. One of the earliest investigations of the efficacy of video in L2 instruction was that 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 more non-verbal cues led to better recall and comprehension. These findings confirmed those of several studies from the 1950s through the very early 1980s, reviewed by MacWilliam (1986), which found that, in general, comprehension is benefitted by seeing the speaker. Interestingly, several of

these studies also found that other types of visual stimuli, e.g. pictures or other video, as in the case of a newscast, had a deleterious effect on overall comprehension, a finding shared by Mayer (1997) within the context of multimedia instruction materials. Citing these studies, both Kellerman and MacWilliam recommend using video for L2 instruction.

Beyond the visual cues listeners receive from the speaker, there are external cues which can aid in comprehension. These usually provide the listener information about the physical and situational context, which aids in his or her ability to understand the information presented (Rost, 1990; Lynch, 1998). An example might be a university lecture hall, signaling to the listener to expect an academic topic and perhaps a somewhat formal register; another may be a coffee shop, setting an expectation of a more informal exchange between multiple people. Such cues draw upon background knowledge, therefore helping listeners interpret what they hear. Two common theories to explain this effect are script and schema theory (Imhof, 2010; Rost, 2011; Wolvin, 2010; Wolvin & Coakley, 1996). Buck (2001) uses an example of the typical activities at a *sento*—a Japanese public bath—to explain the concept of scripts. Anyone who has been to one knows that one enters, undresses, showers, and then sits in a large, very hot tub of water with others. This is the sento script, and it is shared by anyone who has been to a sento. In the case of this particular script, we would expect Japanese and Japanese-resident listeners to understand more of what was said about a sento because they have this basic understanding of the sento script. Schemata are more general than scripts. Whereas scripts contain detailed procedural and role information, schema are perhaps best understood as archetypes of concepts, events, people, places, and so on (Buck, 2001). Listeners who are cued to these sources of background knowledge can spare themselves the cognitive processing

overhead of struggling to understand the scene and frame the verbal information in terms of the relevant activated scripts and/or schemata.

2.4. Video in Listening Comprehension Tests

Given the importance of nonverbal and other visual information to the comprehension of verbal language under normal circumstances, many researchers have investigated the use of video in listening comprehension tests. Video-mediated foreign language listening tests have been in use since the 1980s at least (Feyten, 1991). A tabular overview of the available published studies can be seen in Table 2.1. This section reviews the video listening test literature, beginning with studies comparing video-mediated tests to audio-only tests. Problems appearing in these studies are then discussed. The section following reviews the present researcher's own published work attempting to address some of these issues, the study in question functioning as a pilot for the work presented in this thesis. After the comparative section, additional themes that have appeared in the video listening test literature are presented, as they also inform the research questions and/or design of the studies contained in this thesis.

2.4.1. Comparative Studies

This section reviews the studies comparing video to audio in foreign language listening tests (available in Table 2.1). Since most listening tests are still essentially audio-only, there has been considerable interest in comparing that format to video-mediated listening tests. The results have been mixed.

Table 2.1.

Overview of Studies of Video in Foreign Language Listening Tests

Study		Type	Purpose	Input	Item Format(s)	Text Type(s)	Variable(s)	*N*	Findings
Parry & Meredith	1984	Compar-ative	Determine whether people who took a video dialog Spanish test would outperform those who took an audio test.	Audio, Video	MC	Conversation	Scores	178	The video group performed significantly better for all proficiency groups but the top, which was un-comparable because all participants scored 100%.
Gruba	1993	Compar-ative	Compare a video and an audio-only version of a test based on an air traffic safety lecture.	Audio, Video	MC, T/F	Lecture	Scores	91	No difference observed.
Baltova	1994 (1)	Compar-ative	Study 1: Determine how much video aided global understanding. Used video with audio, video without audio, and audio-only. Control had no input at all.	Audio, Video	MC	Conversation	Scores	53	Video with audio performed better than audio-only; video with audio and silent performed roughly equally.
Baltova	1994 (2)	Compar-ative	Study 2: Determine how much video aids in text comprehension. Used video with audio and audio-only	Audio, Video	MC	Conversation	Scores, Preferences	43	No difference between formats. Participants found shorter scenes easier with video; longer scenes were still difficult. People preferred video.
Chung	1994	Compar-ative	Compare audio-only, audio and a single image, audio and multiple images, video, and visuals-only.	Audio, Still(s), Video	Likert, Essay, MC	Conversation	Scores	75	Increased visual information resulted in many significant differences from the levels below, with more visuals decreasing difficulty for all groups.
Progosh	1996	Single	Investigate what examinees think about video in listening tests.	Video	MC, Short answer	Conversation, Monologue	Preferences	62	Participants preferred video.
Brett	1997	Compar-ative	Compare audio and video to multimedia for learning.	Audio, Video	Short answer, T/F	Conversation	Scores, Preferences	49	Video scores were usually higher than audio, but statistics are lacking. People preferred multimedia.

Study		Type	Purpose	Input	Item Format(s)	Text Type(s)	Variable(s)	*N*	Findings
Shin	1998	Compar-ative	Explore concurrent validity of video tests.	Audio, Video	Essay	Lecture	Scores	83	The video group performed significantly better.
Coniam	2001	Compar-ative	Explore the use of video on a high-stakes test for English instructors.	Audio, Video	Short answer, Essay	Conversation	Scores, Preferences	104	No difference between formats. Participants did not like the video because they did not like to look up and down. They found it distracting.
Wagner	2002	Single	Explore the construct validity of test for language ability and task characteristics.	Video	MC, Short answer	Conversation, Lecture	Scores	85	Two factors were revealed: Top-down and bottom-up.
Hernandez	2004	Compar-ative	Not technically a testing paper. Compare audio and video tests with and without closed-captioning.	Audio, Video	MC, Short answer	Conversation	Scores, Preferences	115	Video-based tests were significantly easier, but the effect size was very small. Preference was difficult to interpret.
Sueyoshi & Hardison	2005	Compar-ative	Investigate how much gestures and facial cues contribute to comprehensions.	Audio, Video	MC	Lecture	Scores, Preferences	42	Presence of video (both face and full-body) improved scores. No difference found between face and full-body. Participants preferred video. Low-proficiency participants benefitted the most from video. High-proficiency participants performed best with face only.
Ockey	2007	Single	Investigate the addition of stills and/or video to computer-based tests.	Video		Lecture	Preferences, Amount watched	6	Some participants preferred video; three reported that the video was distracting at some time, and one of those always reported it as distracting. People looked at video more than stills. Participants referred to a wide variety of stimuli.
Wagner	2007	Single	Determine how much examinees watch videos, and if it differs by text type.	Video	MC, Short answer	Conversation, Lecture	Text type, Amount watched	36	On average, participants watched 69% of time, 72% on dialogs, and 67% on lectures. Individuals vary widely.
Wagner	2008	Single	Determine what individual examinees think while taking a video test.	Video	MC, Short answer	Conversation, Lecture	Text type	8	Participants reported the use of many different stimuli. They did not refer to hand gestures very frequently.

Study		Type	Purpose	Input	Item Format(s)	Text Type(s)	Variable(s)	*N*	Findings
Londe	2009	Compar-ative	Compare head-only video to full-body video to audio-only.	Audio, Video	Essay	Lecture	Scores	101	No differences observed.
Suvorov	2009	Compar-ative	Compare scores on audio, video, and context-still listening tasks.	Audio, Video	MC	Conversation, Lecture	Scores, Preferences, Text type	34	The video test was harder. Significant differences between text type by format. Dialogs were easier than lectures. Video-audio lectures were the hardest passages on the test. No difference observed between performance on the types based on preference, although those who preferred audio performed significantly better on audio. Slight preference for audio-only.
Schroe-ders et al.	2010	Single	Investigate dimensionality to determine if visual comprehension is a separate construct from listening comprehension.	Audio, Video	MC	Lecture, Monologue	Scores	485	Finds that a confirmatory factor analysis model with reading comprehension, listening comprehension, and "video comprehension" all separate but correlated displayed the best fit.
Wagner	2010a	Single	Investigate how video viewing interacts with score.	Video	MC, Short answer	Conversation, Lecture	Scores, Preferences, Text type, Amount watched	56	Finds negative correlations between watching the video and score. Again, people watch dialogs more than lectures. Participants weakly in favor of video.
Wagner	2010b	Compar-ative	Compare audio to video to determine what the effect of non-verbal features is on listening testing.	Audio, Video	MC, Short answer	Conversation, Lecture	Scores, Text type	202	Video was found to be easier on both conversation and lecture text types.
Cubilo and Winke	2013	Compar-ative	Compare video with audio-photo.	Audio, Video, Stills	Essay	Conversation, Lecture, Monologue	Scores, Preferences	40	No substantive differences in overall score. The language use category for the essay rating was significantly better with video. Participants who saw the video took fewer notes. Participants liked the video, overall.

Study		Type	Purpose	Input	Item Format(s)	Text Type(s)	Variable(s)	*N*	Findings
Suvorov	2013	Compar-ative	Compare academic video with either context or content video to an audio-only test. Used eye tracking to compare context and content visual use and compare use to score.	Audio, Video	MC	Lecture	Scores, Eye-tracking measures	121	No difference observed between scores on either format of either the context or content tests.
Wagner	2013	Compar-ative	Compare scores on audio and video versions of a listening comprehension test, with questions accessible or not. All groups were allowed to preview the questions.	Audio, Video, Stills	MC	Conversation, Lecture	Scores	192	No effect observed for having access to questions during listening. No interaction between having access and channel of input. Audio-visual is significantly easier than audio-only.
Pusey & Lenz	2014	Compar-ative	Investigate the effect of visual information on listening comprehension and its interaction with working memory.	Audio, Video	MC	Lecture	Scores, Working memory measures	24	Audio-only was found to be significantly easier. No interaction between examinee working memory and delivery format observed.
Suvorov	2015	Single	Compare viewing behavior between tests with content vs. context videos. Used eye tracking.	Video	MC	Lecture	Scores, Eye-tracking measures	33	Participants watched content videos more than context videos. No relation between amount and type of watching to total score. Participants watched about 50% of the time.
Batty	2015	Compar-ative	Compare audio to video multiple-choice questions, including differential distractor analyses.	Audio, Video	MC	Conversation, Lecture, Monologue	Scores, Text type	164	No difference observed between formats; no difference for formats by proficiency level, no difference between performance on text types. Several items exhibited format-based DIF.

The earliest published comparative study of which the present author is aware is that of Parry and Meredith (1984), which administered either an audio- or video-mediated version of a Spanish test to 178 university Spanish students in the United States. The participants were from four separate course levels, and were randomly assigned to the two conditions. The participants only encountered one or the other format. The test content was informal conversation, and the items were sixty (60) multiple-choice questions presented in English "to avoid simultaneous testing of the reading skill in the foreign language" (p. 50). The video test was found to be significantly easier for all levels but the top level. However, this was due to the fact that those examinees nearly all attained perfect scores, rendering them incomparable.

The next study appearing in the literature comparing the audio and video formats is that of Gruba (1993), which he conducted on two intact classes of ESL students at UCLA in 1990, with a total *N* size of 91. The test was fourteen multiple-choice and true/false questions over a lecture on terrorism in airports. He then compared the scores of the two groups with a *t*-test, finding no significant difference. He blamed this inconclusive result upon the poor reliability (0.45) of the instrument. In his closing, he noted that, "second language test developers should be careful in using video as a testing medium because of construct validation concerns" (p. 87), a topic to feature heavily in the present thesis. His reflection on this and later work led him to ultimately advocate an abandonment of comparative studies, as he felt that the construct of foreign language listening comprehension was distinct from that of audio-visual listening comprehension, and that such comparisons were unlikely to prove productive (1997). As can be seen from Table 2.1 and the remainder of this review, his advice went unheeded.

Baltova carried out two of the more intriguing audio/video comparative studies in a middle school French language program in Canada (1994). The content for both was a fifteen-minute video from the school's French curriculum, employing informal conversation. Both forms used multiple-choice questions. Fifty-three (53) eighth-grade students took part in the first study, 43 in the second. In the first study, students were randomly assigned to one of four treatment groups: audio-only, audio-video, video-only, and no input whatsoever. Any items correctly answered by the no-input group, and any that no audio-only participants answered correctly, were removed before analysis, as the former would indicate that the item did not require listening at all, and the latter required visual information to answer correctly.

The results of the first study are remarkable. The audio-video test was significantly easier than the audio-only version, but no significant difference was observed between the audio-video form and the video-only form. Wagner concluded from this finding that there were "obvious problems with the test" (2010b, p. 496), but it seems equally likely that nonverbal behavior is sufficiently information-rich as to allow for very effective coping strategy use, especially given the findings of Fay, et al. (2013) and Dahl and Ludvigsen (2014) discussed previously.

The second Baltova study removed the "control" conditions of video-only and no-input, comparing an audio-only version of a slightly updated version of the test (based on the results of the first administration) to an audio-video version. When scores on these tests were compared, no significant differences between them were observed, likely due to the editing or removal of items deemed clearly biased toward one or the other format.

Chung compared the performances of relatively small groups of university students of French in the United States on four listening testlets, each with increasing

amounts of visual information (1994). Participants were grouped into intermediate French, advanced French, and a group which required no French ability whatsoever, as they would only view visuals. The groups were presented with a dialog first with audio only (except for those in the visual condition), then audio accompanied by one still photo, then audio and a slide show of several still images (similar to the current TOEFL listening section), then full audio with video. Participants completed a Likert item indicating their overall comprehension, wrote a summary, wrote a "résumé"—a list of everything they could remember about the dialog, and answered multiple choice questions written in English. Generally speaking, more visuals resulted in easier testlets for all groups, with many significant differences between the levels observed, once again demonstrating the difference between even multiple still images and full video in terms of informational content.

Brett (1997) compared the results of tests using audio-only, video, and "multimedia" (which appears to be operationally defined as being delivered by computer, with the participant completing tasks as a video plays in a box on the screen). The participants were 49 (or 43—six neglected to attend the second, multimedia day) advanced university students of English in the United Kingdom. The texts were six English business conversations. Unfortunately, the results are fairly difficult to interpret, as statistics are lacking. The two-sample *t*-tests reported compare audio to multimedia, and video to multimedia, but not audio to video. Finally, a broad range of item types were used between the dialogs, damaging comparability. However, in four of the six cases, video was found to be easier than audio, and multimedia was usually found to be the easiest of all. Participants preferred multimedia.

In the comparative study by Shin (1998), 83 university ESL students in the United States took either an audio- or a video-mediated test based on authentic lectures. However, whereas the video version used the original lectures, the audio version was comprised of recreated performances based on the transcripts of the lectures, with pauses, disfluencies, words such as "well" and "alright" as organizers, and casual ungrammaticality removed. The question format was free response, rated for 0 – 3 points by majority vote among three raters. Video was found to be significantly easier than audio, and both groups' scores were found to be significantly correlated with those on a criterion pretest measure. The author cites this finding and concludes that the video test demonstrated concurrent validity.

Coniam (2001) has one of the more-cited studies comparing an audio- and video-mediated version of the same test. He administered a thirty-minute talk-show-based test to 104 education graduate students in Hong Kong, broken into three spans of ten minutes, with five minutes between the sections for catching up, and ten at the end to finish answering. This is one of the longer formats appearing in the literature. The questions were limited to extended free-answer format, "which involved summarising speakers' points and attitudes" (p. 6). The audio and video versions were identical aside from the presence or absence of the video. Although the design was not counterbalanced, Coniam did employ an external proficiency measure (an established cloze test) to ensure that the two groups had equivalent proficiency at the outset. The audio group performed slightly better than the video group, but the difference was not statistically significant.

Hernandez (2004) administered audio and video versions of a test in her investigation of the impact of closed captioning on comprehension. The listening material was four scenes from a film, and the test used a mix of multiple choice and

limited-production items, which could be answered in the participant's native language. The test was administered to 115 university ESL students in the US. Video was found to be easier; however, the study suffers many problems, chief of which was the 0.40 reliability coefficient for the instrument.

Sueyoshi and Hardison (2005) carried out a modest, but nonetheless interesting study into the effect gestures and facial cues had on listening comprehension among 42 university ESL students of two proficiency levels in the United States. They did this by creating three versions of their instrument: audio-only, audio with facial video (i.e., the shoulders and face only), and audio with full video (i.e., the body was also visible). The listening content was a staged lecture delivered from an outline, retaining the features of spontaneous speech as a result. It was split into five sections ranging from two to four minutes in length, with four multiple-choice questions per section. The gestures appearing in the video were classified and tabulated following McNeill's classification system (1992), mentioned briefly in the "Nonverbal Behavior" section of the present chapter, but this information was simply reported, rather than used to predict scores. The two video conditions were found to be significantly easier than the audio-only condition for both proficiency levels, but the difference between the two video conditions was not significant.

Londe carried out a similar study to that of Sueyoshi and Hardison with a much larger sample (2009). Londe presented 101 university ESL students in the United States with either an audio-only, video-with-face, or full video version of a ten-minute lecture. The questions were eleven open-ended, limited-production items. Contrary to that of Sueyoshi and Hardison, however, no significant differences in score were observed.

Suvorov's (2009) first published study on audio- and video-mediated listening tests involved the administration of a 30-item (multiple-choice) listening comprehension test to 34 university ESL students in the United States. The test was comprised of three sections: audio-only, audio-video, and audio-still, ensuring that each participant encountered each of the conditions. This design addresses a common problem in the comparative literature wherein examinees only sit for one format. The content included both conversations and lectures, one listening of each type per format. Overall, the audio-video format was found to be the hardest, with the bulk of that effect appearing to be attributable to the video/lecture items. There was no significant difference by format for conversations. However, it is important to note that the data collection was not counterbalanced, so this could simply be the result of the lecture delivered via video being more difficult than the other passages. This will be revisited in the section immediately following.

Despite Wagner's long list of articles on the topic of video-mediated listening tests, dating back to 2002, his first published comparative study was in 2010 (2010b), and reported the results of the final study of his 2006 doctoral dissertation on the topic. The purpose of the study was to investigate the impact of video on scores, that effect's interaction with text-type (conversation or lecture), and what kinds of visual information affected which items. Wagner, like Coniam before him, used an external proficiency measure (the placement test for the participants' language program) to control for proficiency, since each group would only encounter either audio or video of the experimental instrument. The participants were 202 adult continuing education ESL students in the United States. Video was found to be significantly easier in all cases. A qualitative analysis of the videos led to speculation that the use of pictures in

several of the videos, and illustrative gestures in others may have facilitated certain items. Caveats to these interpretations will be presented in the next section.

Cubilo and Winke (2013) administered two listening-to-write tasks from a practice TOEFL IBT (institution-based test), using the transcript to re-perform the content as a video. This was administered in a counterbalanced design to forty university ESL students in the United States. No significant differences in essay scores were observed, but there was a statistically-significant reduction in the amount of notes taken in the video condition, likely due to the participants viewing the screen instead of their notes.

Suvorov completed his doctorate in 2013 with an investigation of the impact of content (e.g., presentation slides) video and context (e.g., a lecturer only) video on listening test performance, using eye-tracking methodology. A portion of his dissertation was published in *Language Testing* in 2015, but the published section contains no comparison, and will be discussed in greater detail in the "Eye Tracking" section of Chapter 4. In Suvorov's (2013) comparative study, an academic listening test was administered via the web to 121 university ESL students in the United States. The listening content was six short lectures sourced from YouTube and thirty multiple choice questions. The participants were randomly assigned to an audio-only group (n = 46), an audio-video group, or the eye-tracking group which also sat the video version of the test (n = 75). No significant score differences were observed between the audio and video tests, nor between the content or context video sections of the video test.

Wagner's (2013) first published study on video listening tests to use material developed after his doctoral dissertation examined the interaction between audio, video, and item access. Once again, a pretest was used to ensure comparability of the

groups. The experimental instrument was comprised of two dialogs and two lectures, with 26 multiple choice questions. One-hundred and ninety-two (192) university ESL students in the United States were assigned in intact classes to one of four groups: audio-only with question options accessible during listening, audio-only with questions inaccessible, video with items accessible, or video with items inaccessible. All groups previewed the items before listening. Once again, the video format was significantly easier than the audio format, but access to questions was found to have no effect on scores.

Finally, Pusey and Lenz (2014) carried out a modest study on 24 university ESL students in the United States, administering a twenty-item multiple choice listening test over three short lectures in either an audio-only or video condition. The participants were assigned to groups by coin toss. The participants also took an assessment of their working memory, in order to investigate the interaction between access to visuals, working memory, and scores. The video was found to be more difficult than the audio version, and no interaction between the facilitative effect of video and working memory was observed.

Of the seventeen studies reviewed above, the total number of comparative studies of which the present author is aware, nine found the presence of video to significantly raise scores in comparison to equivalent audio-only tests, two found it to lower scores, and six found it to have no effect. Obviously a wide range of methodologies are represented by the literature on this topic, as well as a number of different lines of inquiry. The question of how reliable and/or generalizable the findings of these studies can be taken to be, however, will be addressed in the following section.

2.4.1.1. Criticism of video listening studies

Despite the large number of studies on this topic, results can be difficult to compare or generalize due to the wide range of test types, narrow range of contexts and populations, and sample sizes. Furthermore, due to a number of issues related to their methodologies and item design, it is unclear to what degree the results are reliable. This section elucidates those issues primarily in the comparative studies described above, with some additional attention paid to single-format studies, in terms of comparability, methodology, item design, and measurement models.

2.4.1.1.1. Comparability of studies

As can be seen in Table 2.1, many different studies employing a wide variety of designs have been carried out on the topic of video-mediated listening tests. Whereas most studies employ multiple-choice (MC) items, those of Progosh (1996), early work by Wagner (2002, 2006, 2007, 2008, 2010a, 2010b), and Hernandez (2004) also include limited-production or short-answer items, and those of Brett (1997) and Coniam (2001) eschew MC items entirely in favor of short-answer or cloze items. Items such as this (with some exception for cloze items) typically result in more variable answers. Moreover, marking such items typically involves a measure of subjectivity, adding to the variability of scores. As such, it is hard to apply findings associated with them to other contexts, as the degree to which these sources of variability affected outcomes is unknown. Even harder to apply to other contexts are the findings of research employing essay questions, such as Shin (1998), Coniam (2001), Londe (2009), and Cubilo & Winke (2013). Although it is fairly straightforward to compare the results of MC-based studies, the generalizability of findings from studies employing these more open-ended item formats are difficult to judge.

Another factor calling the generalizability of results into question is the very narrow range of contexts in which the research has been conducted. Of the 27 studies listed in Table 2.1, nineteen were conducted in the United States at university intensive English programs (or similar). These students are studying in a second-language context, ensuring that they are continually surrounded by native speakers of English, which may make them more attuned to culturally-bound body language.

Finally, sample sizes are fairly small for the bulk of the studies listed in Table 2.1. Seventeen of the studies have sample sizes under 100, which, in the studies relying on statistical evidence of findings, may not be adequate. Only two studies, those of Schroeders et al. (2010) and Wagner (2010b) have sample sizes larger than 200.

Further work on this topic should limit itself to MC items, apply the methodology to language learners in a foreign-language context, and seek a large sample size.

2.4.1.1.2. Methodological concerns

The methodologies employed by these studies are varied, and sometimes not ideal. Parry and Meredith (1984) and Baltova (1994) lack counterbalancing to ensure that all examinees saw both conditions, and that all items were administered under both conditions. Many of the studies (e.g., Baltova, 1994; Gruba, 1993; Parry & Meredith, 1984; Suvorov, 2009) lack an external measure of foreign language proficiency to control for ability. Baltova (1994) changed the items between her two studies to the extent that they are not comparable; Shin (1998) removed pauses in speech from the video (resulting in an inauthentic delivery), and used committee-rated free-answer questions that are difficult to interpret; and the Hernandez (2004) instrument had a KR-20 reliability coefficient of only 0.401. Due to these problems, it

is unclear whether the findings of the body of work on this topic can be considered accurate.

2.4.1.1.3. Item concerns

Although most of these studies do not offer the items in question for examination, Hernandez (2004), Suvorov (2009), and Wagner (2006) are exceptions. The Hernandez items appear to include quite a few double-keys and weak distractors (although no distractor analysis appears to have been performed), which likely explains the poor reliability coefficient noted above. The Suvorov instrument in its entirety was previously available on the author's website, and the items appeared to be very well-crafted, although there did not appear to be much nonverbal behavior to be observed in the videos. The performers in the scenes appeared to be fellow graduate students who were untrained in acting. As a result, the performances lacked much of the natural nonverbal content which underpin the video format's claim to authenticity.

Most of Wagner's work appears to be based on the items that originated in his doctoral dissertation (2006), which is available for examination. Although his work is extremely thorough, some of his items do not seem to be written in a truly unbiased manner. The best example of this is one of his "Wild Bill" items, wherein the actor admitted to having trouble describing the way the famous Wild West lawman drew his pistols from their holsters without demonstrating it with his hands. It is not surprising, therefore, that this item was found to be much easier under the video condition, especially given the findings of Fay, Arbib, and Garrod (2013), who found that gesture both with or without nonsense vocalization was equally communicative, and those of Dahl and Ludvigsen (2014), who found that L2 listeners understand as much as L1 listeners when given access to gesture. Another concern with Wagner's items is his choice to allow his actors to improvise scripts based on an outline. Wagner's

argument for this choice is that it is more authentic than the tightly-scripted passages that normally feature in listening tests, but that tight scripting is usually necessary to ensure that distractor options in multiple choice items duly distract. In Wagner's instruments, distractors were often other, unrelated, single vocabulary items, which could have been answered without the listening content if the words were already known, or phrases describing content that was not included in the listening material at all.

An example of the first issue would be the third item of the "Skunks" listening passage:

3. The thick, yellow fluid that causes a skunk's odor is called _____.
    a. musk
    b. gland
    c. species
    d. omnivorous

(Wagner, 2006, p. 337)

In the example above, there is only one possible answer if one knows the meanings of the words. The lowest-frequency word, "omnivorous," may have been unknown by the participants, but the "-ous" suffix marks it as an adjective, when the stem is clearly asking for a noun, allowing the examinee to quickly eliminate that as an option. As such, it is possible that items such as this did not function as listening comprehension items, but, rather, as vocabulary items. However, all the words offered as options do, indeed, appear in the "Skunks" transcript. This cannot be said of the fifth item for the "Wild Bill" listening passage:

5. Wild Bill was killed in _____.

    a. Abilene, Kansas

    b. Kansas City, Missouri

    c. Deadwood, South Dakota

    d. Rock Creek Station, Nebraska

(Wagner, 2006, p. 339)

In the item above, only options "c" and "d" appear in the transcript, which suggests that examinees could have selected the credited responses based upon phonetic recall alone. Even when all the terms appear, however, they are not typically mentioned in conjunction, which calls into question whether the distractors distracted for reasons related to comprehension, or simply memory. As such, his findings may not be generalizable to more traditional test formats, which typically attempt to link any multiple-choice distractors to the propositional content of the listening stimulus.

2.4.1.1.4. Measurement models

A curious gap in the literature on the topic of audio-video comparability, however, is the use of any form of item response theory (IRT) in the analyses, even in cases in which the sample size was sufficient to make use of it. All of the studies discussed here rely upon raw score comparisons, which can be used to explore differences between groups and conditions by examining item-total correlations or comparing item facilities between groups, but which are sample-dependent. However, such classical test methods suffer from a serious flaw in studies related to construct validity: the construct does not feature in the model at all; such methods are only concerned with the patterns of dichotomous responses to the various items. This is in sharp contrast to IRT methods, which includes the construct as the latent (mathematical) variable underlying the item responses (Borsboom, 2009).

Furthermore, IRT methods provide the researcher item difficulty statistics which are theoretically sample independent (Embretson & Reise, 2000), and individual item-model fit statistics which can be used to detect departures from unidimensionality in the data (Engelhard, 2009; Engelhard, Kobrin, & Wind, 2014)—indication that an item may not be measuring the same construct in two different delivery formats, for example. In contrast, Wagner (2010b) employs classical test theory methods in his examination of individual item functioning under the audio and video conditions of his test. Although these methods are by no means universally inferior to IRT, his method of determining which items to single out for greater scrutiny was to examine any item with a 10%-or-greater difference in scores under the two conditions, which seems somewhat arbitrary. By using bias statistics from many-facet Rasch modeling, differences in examinee performance between the formats (audio and video) can be selected for closer inspection for much more principled reasons, and then examined much more closely (Eckes, 2015; Engelhard, 2009). The specific methods employed will be explained in detail in the Data Analysis section of the Study II chapter.

2.4.1.2. Batty (2015)—A pilot study

Noting the problems with the existing comparative literature, the present researcher undertook a comparative study seeking to address some of the issues discussed above, primarily those related to item format and multiple-choice distractors (Batty, 2015). He sourced items and video texts from two forms of the video-mediated placement test used by his employer at the time, Kanda University of International Studies (KUIS) in Chiba, Japan—the descendant of a video-mediated listening test originally designed by Gruba. Each test form was comprised of four videotexts ranging from 1 min. 47 sec. through 4 min. 30 sec., with one monologue videotext,

two conversations, and one lecture. Each videotext was accompanied by five-to-seven multiple-choice items. The listening sections from the two forms of the placement test were used as sections of the experimental test for the study, designated “A” and “B.” Four forms of the experimental test were produced to counterbalance for format of delivery (video or audio-only) and content (A or B).

The test was administered to 164 students of said university in eight intact classes of four different institutional English ability tiers. Examinees were given one minute to read the items for the next listening passage, and a minute and a half to answer them after listening. Comparative format difficulty as well as interactions between format and text type and format and ability level were sought via many-facet Rasch modeling (MFRM) with the software package FACETS (Linacre, 2014). The format difficulty estimates were essentially identical, and no significant interactions between format and other factors were observed.

To investigate possible format differences at the item level, item difficulty estimates which were significantly different when delivered in the two formats were sought; four were discovered. As it can be very difficult to determine the cause of differential item functioning (DIF), these analyses were paired with differential distractor functioning (DDF) analyses using the software package DDFS (Penfield, 2008, 2010a, 2010b). Only one of the four items exhibiting DIF seemed to possibly stem from a difference in the delivery format.

It is critical to note, however, that the test in question was carefully constructed to exclude any possible effect from the presence of the video, in adherence to the institutional test specification, which was developed with concern for the construct validity of the test should any items make use of the video. Furthermore, the performances on the video were of such an amateurish quality that there were

virtually no useful nonverbal cues to be found. Actors had not memorized their lines, and spent each scene looking diagonally off camera, clearly reading cue cards. Finally, it is unclear as to how independent the items were, given the "testlet" design of the instrument, with several items per videotext, which—though near-ubiquitous in listening tests—violates the assumption of local independence. As such, it remains unclear how much the format of the stimulus interacts with task type.

2.4.2. Additional Themes in Video Listening Test Research

In addition to comparisons of scores on audio-only and video-mediated listening tests, much research has been conducted on a variety of related issues. This section provides a brief overview of the studies, some already discussed, addressing additional issues related to video-mediated listening tests, most of which will also feature in Study I and/or Study II of the present thesis.

2.4.2.1. Examinee preference

Many of the studies in Table 2.1 include data on examinee preferences regarding the formats, with varied results. Whereas Baltova's (1994), Progosh's (1996), Sueyoshi and Hardison's (2005), and Cubilo and Winke's (2013) participants reported a preference for the video-mediated test, the participants in Coniam's (2001) and Suvorov's (2009) studies showed a preference for audio, complaining that it was distracting in the case of Coniam's study. Ockey's (2007) participants' ($N = 6$) reports on their perceived usefulness of the video input ranged from "always helpful" to "always distracting," with a wide range of responses in between. Wagner's study on the topic (2010a) administered a seven-item questionnaire on examinee attitudes toward the use of video in his experimental listening test. The mean score on the five-point Likert scale was 3.47 ($SD = 0.70$), indicating a weak endorsement of the format. Brett (1997) found that participants preferred multimedia presentations to either audio

or video passages. The preferences of Hernandez' (2004) participants are difficult to interpret, as the item wording in her instrument is ambiguous. The lack of consensus between these studies leaves the issue fairly unresolved.

2.4.2.2. Text-type

Another frequently-visited topic is the interaction between test format and text type, drawing on the psychological research demonstrating that the nonverbal channel is of more importance in social settings, and less so in those focused on relaying information (Burgoon, 1994; Burgoon et al., 2016). This general finding led Buck to hypothesize that video would play more of a part in the comprehension of "interactional" (i.e., conversational) than in "transactional" (e.g., lectures) language use (2001, p. 172). Suvorov (2009) found that video-mediated lectures were significantly harder than their audio-only counterparts, as well as conversations under either condition. In contrast, Wagner (2006, 2010b) found that the addition of video improved scores on both conversational and academic passages to approximately the same degree, although effect sizes were small (Cohen's d = 0.29 and 0.34, respectively). These findings alone demonstrate the lack of consensus in the literature on this topic.

2.4.2.3. Time spent watching

Wagner (2007) investigated how much time examinees spend oriented toward the screen in video-mediated listening tests, in an attempt to understand whether they found the videos to be a help or a hindrance to comprehension. He 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. However, in the sample of 36 examinees, the ranges were quite high, with some examinees watching the videos up to 90% of the time and others less than 50%.

Ockey (2007) found similar with his six participants, who watched the video in his listening test anywhere from 2.5% to 73.6% of the time. Wagner (2010a) revisited this question with a larger sample size of 56, again finding a significant difference in the amount of time spent watching dialogue (58.5%) and lecture (41.6%) videos, and a small, but statistically significant, negative correlation between time spent watching the videos and total score.

Most recently, Suvorov (2013, 2015) has attempted to improve upon the studies above by employing eye-tracking methodology. Examinees were presented with videos of academic lectures with either content (e.g., slide presentations, blackboard illustrations) or context (i.e., simply a lecturer speaking) visuals, similar to the still-image study by Ginther (2002). The eye tracking software was utilized to determine how much time was spent in oculomotor engagement with (i.e., "looking at") the video window in the web-mediated listening test, granting him similar data to those above, but with greater precision. Content videos were observed to have significantly higher dwell times (time spent looking; 57.99%) associated with them than context videos (50.70%), although no significant correlation with test score was found. Although Suvorov argues that by setting the eye-tracking software's area of interest (AOI) to the video window within the web-browser is superior to merely counting heads oriented toward a single screen in a classroom, as Wagner and Ockey separately did in the studies discussed above, it is unclear as to what else participants in the Wagner and Ockey studies could have been turning their attention to. Given the functions available in high-end eye-tracking software, it is puzzling as to why Suvorov did not track attention to specific visual cues within the videos in question.

2.4.2.4. Visual cues

Both Sueyoshi and Hardison (2005) and Londe (2009) have attempted to isolate the contribution of specific human nonverbal cues' contribution to scores by comparing "talking head" videos to those featuring full body video, but more frequently the investigation of visual cues in video-mediated listening tests employs qualitative methodologies.

In the same study discussed above, Ockey (2007) used retrospective verbal reports as a means of exploring the specific visual cues to which respondents referred in the videos. Respondents reported a wide range of visual cues, from no cues at all through the entire range of lip movements, facial expressions, hand gestures and body gestures. The two visual cues reported by the most respondents were "Facial gestures to indicate opinion" and "Body gestures to indicate emphasis" (p. 530). The list of visual cues, however, was derived from the verbal reports, rather than from nonverbal communication theory, and they are therefore undefined beyond examples from respondent verbal reports.

Another Wagner study (2008) also used verbal report protocols in an attempt to determine what visual cues examinees referred to while the videotext was played and while answering the comprehension questions. The former was achieved by stopping the playback every few seconds so that the examinee could verbalize in English what he or she was looking at and why; the latter was via a thinkaloud protocol. This methodology is problematic, however, as it required the examinees to continually "pause" their thinking as the video was paused, and likely fundamentally alters the examinee's interaction with the stimulus (Gorin, 2006a). The total number of references to visual cues ranged from zero through sixteen, but it is unclear whether these results can be trusted, as transcripts of the interviews seem to suggest that the

examinees' English level may not have been adequate to explain their thoughts, missing such basic vocabulary as "wrist."

Suvorov (2013) also investigated attention to visual cues employing a similar methodology to that of Ockey above, but the retrospective verbal reports were conducted while watching gaze-overlaid videos from the eye-tracker, a method known as cued restrospective reporting. The results of these interviews indicated that when respondents were presented with a lecture without rich visual aids (i.e.,with a "context" video), the visual cues commented on by the largest number of respondents were related to the "speaker's mouth, face, head, hands, eyes" (88% of respondents), followed by "speaker's body movements, gestures" (58%; p. 157). By contrast, when presented with a lecture using rich visual aids (i.e., a "content" video), the cues mentioned by the most participants were the visual aid (97%), followed by "speaker's mouth, face, head, hands, eyes" (55%). It is important to note, however, that despite the use of sophisticated eye-tracking hardware and software, no attempt was made to quantitatively examine visual attention to any specific visual cue within the videotext, relying, as in previous studies, solely upon participant reports of visual attention.

2.5. Item Task Type

As discussed above, the interaction between listening test stimulus format and task type is still largely unexplored, which may not be surprising, as most research into listening task types generally is concerned with the response format (e.g., multiple-choice, short-answer, cloze), rather than the comprehension strategy necessary to engage it (e.g., Berne, 1993; Brindley & Slatyer, 2002). There are, however, some notable exceptions. Shohamy and Inbar (1991) compared the difficulty of "global" items, those which required some degree of inference, and "local" items, those which required the comprehension of the language and the facts of the verbal stimulus,

finding that the global items were significantly more difficult than local items, even across the three text types (news broadcast, lecturette, and dialog) examined. In an investigation of the sources of item difficulty in the TOEFL dialogue items, Nissan, DeVincenzi, and Tang (1996) also included the degree to which examinees must infer to answer questions, also finding that items requiring more inference, i.e., "global" items in the Shohamy and Inbar classification, were significantly more difficult than those wherein the answer was available explicitly within the verbal stimulus. Freedle and Kostin (1999) investigated a similar question in the TOEFL minitalk (i.e., lecturette) items by classifying them as "detail explicit," "detail implicit," and "gist" (pp. 10 – 11), but found that task type differences were not predictive of item difficulty. Finally, recent work by Field (2013) has attempted to categorize listening items by the types of information, and therefore the cognitive processing demands necessary to answer them, but due to the contributions of text and distractors to difficulty, he admits that his five categories "cannot be more than loosely indicative" of actual item difficulty (p. 137).

However, the impact of task type has been largely neglected in the video-mediated listening test literature, despite calls for such work to be undertaken, with specific hypotheses of its relation to the video medium. Wagner (2002, 2006) employed exploratory factor analysis to investigate the factor structure of a video listening test, identifying two factors in the item responses, which he classified as "top-down" and "bottom-up" processing. Most items which required the use of inference loaded more strongly on the latter, while those which required the comprehension of more prosaic features of the listening stimuli, such as details or vocabulary, tended to load on the former. These classifications would be analogous to "implicit" and "explicit" tasks, respectively, in the present research.

Ockey (2007) concludes his video listening paper with a suggestion that the impact of visual cues may be more apparent in items related to the speaker's attitude, and recommends conducting research comparing such items under the formats of video and audio-only. Suvorov (2013) also calls for more research investigating the relationship between visual cues and the type of information elicited by listening test items (p. 227), an interest shared by Batty (2015). As test data are inescapably the product of examinees interacting with items, the question of how these interact with the addition of visual and nonverbal cues in listening tests must be addressed before the format is more widely accepted and used.

2.6. Validity

As the question of whether and how video should be utilized in foreign language listening comprehension assessment is one intrinsically linked to concepts of test validity, a brief overview of the history and competing models of validity and test validation is in order. This section provides a history of the operational definitions of validity to serve as a foundation for the study rationale to be outlined in the following chapter (Chapter 3).

Validity is an elusive concept, with many subtly different interpretations among assessment researchers and through history. A (gross) simplification of the concept of validity may be: "Do the scores mean what we think they mean" (Alderson, Clapham, & Wall, 1995; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Bachman, 1990), or "whether a test measures what it should measure" (Borsboom, 2009, p. 149). However, among validity theorists, it has grown far beyond this fairly simple (but hard to demonstrate) concept, to one encompassing

all manner of ancillary issues related to the use of the test, including any consequences thereof.

Most discussions of the history of the concept of validity (e.g., Borsboom, 2009; Kane, 2001, 2012, 2013a) begin with a discussion of the criterion model of validity, wherein validity was demonstrated by the test's predictive ability with regards to a criterion measure that is thought to be a more accurate demonstration of the attribute of interest (Kane, 2001). An example may be comparing the scores on a university entrance exam to the accepted students' subsequent grades at the university. Performance at the university is the criterion measure that the test is intended to predict; any test whose scores, therefore, correlated strongly with grade-point average would be considered "valid." The problem with this model of validity, however, arises when the criterion measure is perhaps no better a representation of the attribute of interest than the test itself (Kane, 2012, 2013a). For example, what is the criterion measure for a psychological attribute such as "agreeableness?" Many such measures' criteria will be no more valid than the test to which they are being compared, giving rise to a circular argument.

For such problems, criterion-based validity gave rise to content validity, which "is established by showing that the test items are a sample of a universe in which the investigator is interested" (Cronbach & Meehl, 1955, p. 282). This was/is frequently operationalized by consulting experts to verify that test questions adequately reflected important concepts or skills in the field of interest, or attempting to distill skills into questions or test tasks that attempted to represent those of importance to the tested domain (Guion, 1977). This concept of validity, too, suffered from problems, notably that of confirmation bias and subjectivity, as the determination that the test matches

the domain is/was usually determined by the test developer him or herself (Kane, 2001, 2012).

Partially due to these issues, the concept of "construct validity" was developed (Cronbach & Meehl, 1955). Construct validity seeks validity in the theoretical constructs which are intended to be addressed by tests. The validity of scores, therefore, is determined by the validity of the construct being measured by it, as they are seen to be an "interlocking system of laws which constitute a theory as a *nomological network*" (Cronbach & Meehl, 1955, p. 290; italics in original), where observable scores support the description of constructs and vice-versa. Test validity is determined by investigation of the test from various angles (e.g., group differences, factor analysis, internal structure), with departures from the expected results being taken as indication that either the test was not addressing the theoretical construct, or that the construct itself is not correctly defined or understood from a theoretical perspective (Cronbach & Meehl, 1955; Kane, 2013a).

It is important to note, however, that construct validity was originally proposed as an addition to criterion/content validity. Although it was considered most important to tests wherein no criterion existed, but for which a theory could be constructed, demonstrations of construct validity were not intended to be limited only to tests of this nature (Cronbach & Meehl, 1955; Kane, 2001, 2012, 2013a). A problem with the approach, however, is that the kind of overlapping, interlocking nomological networks envisioned by it rarely materialize outside of the physical sciences (Borsboom, 2009; Cronbach & Meehl, 1955), and for that reason, test validation still required as much support as possible from criterion and content studies, in addition to any study that could support the theoretical framework of the test (Borsboom, 2009; Kane, 2001, 2013a). Over time, the construct model of validity subsumed the other forms, leading

to a unified concept of validity which encompassed all of these within the construct conceptual framework. All evidence gathered for the validity of a measure was considered support for the construct validity, taken as an umbrella term (Kane, 2001, 2012, 2013a).

It was at this point in the development of validity theory that Sam Messick of the Educational Testing Service (ETS) wrote a ninety-page treatise on the concept of validity for *Educational Measurement* (1989), which extended the concept of validity beyond the nomological network to the various uses of the results, opening the piece with the oft-cited summary of his viewpoint:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (p. 13; italics in original)

Messick's view, which essentially became that a test's validity includes not only characteristics of the test itself, but a wide range of other concerns related to the use of scores, including consequences (Messick, 1989, 1996), has been extended further by language testing researchers to encompass social consequences and their ethical implications (McNamara, 2006). A criticism of this concept of validity, however, is that it is so over-arching and all-inclusive that it becomes nearly impossible to test through validation, since there exist no "stopping rules" (to borrow a term from computer adaptive testing) at which point a test can be considered "valid enough" for a particular interpretation or use (Kane, 2001).

Kane has, in recent years, famously attempted to harness the unified, Messick-associated concept of validity and the process of validation to a more concrete method of validation that seeks to address the question of when a test can be claimed to be

valid. His model of validity and validation seeks to position validity as an attribute which can be established through the construction of a validity argument, which supports a separate interpretive argument (Kane, 2001, 2012, 2013a). The process of validation, therefore, becomes an exercise in evaluation of the logical structure and data in support of the claims to validity of a test's scores for a particular interpretation or use, including the exploration and rebuttal of counter-arguments based on further data (Kane, 2012, 2013a), based on the Toulmin model of inference (2003). As such, it is largely intended to provide a concrete procedure to validation, collecting data to support assertions by the test makers and that by which one can rebut counter-arguments. If the argument can be deemed sound, the test is assumed to be valid. This form of validation has not only found support among language testing researchers, but has been established in language testing education by its adoption in Bachman and Palmer's latest edition of their influential *Language Assessment in Practice* text (2010).

This approach, and the unified construct model upon which it is based, however, has been sharply criticized by some psychometric scholars, who argue that the definition of "validity" at use here is too broad to serve any practical use. Popham, for example, wishes to see consequences removed from the discussion of validity, keeping it, rather, focused on the "accuracy of test-based inferences," noting that "cluttering the concept of validity with social consequences will lead to confusion, not clarity", and arguing that "[t]est-use consequences should be systematically addressed by those who develop and utilize tests, but not as an aspect of validity" (1997, p. 9). Dutch psychometrician Denny Borsboom and his colleagues take this criticism one step further and argue that there has been a growing confusion over the demarcation between validity, as a property of a test, and validation, which is the process of

establishing whether a test is valid (Borsboom, 2009; Borsboom, Mellenbergh, & van Heerden, 2004). "Validity is not a judgment at all," he claims, "[i]t is the property being judged" (2009, p. 154). For a test to be valid, in Borsboom's view, 1) the attribute being tested must exist, and 2) variations in that attribute must cause variations in the results of the test. If the attribute has not yet been proven to exist, then it is unknown what causes variations in scores on tests designed to measure it (Borsboom, 2009; Borsboom et al., 2004).

Borsboom and Markus have recently engaged in a point-counterpoint with Kane in the pages of the *Journal of Educational Measurement,* wherein the former researchers describe their approach to validity as an example of "true belief"—that they wish to believe only what is true, as opposed to Kane's "justified belief"—whereby any belief that can be justified is accepted. They point out that a compelling validity argument can likely be constructed to support any interpretation one likes without being sure of the actual validity (i.e., that the attribute exists and that variations in it result in variations in measurement outcomes), and that one runs the risk of building very good arguments based upon false premises if the attribute is not first demonstrated to exist (Borsboom & Markus, 2013). Kane (2013b) concedes their point to a certain extent, but counters that in the realm of psychology, what they desire is unlikely to be conclusively demonstrated. "I focus on plausibility and justification," he writes, "and leave Truth in the background" (p. 119).

Critically, however, it is important to note that Kane (2012) does not advocate the use of his argument model of validation at the test development stage:

> Validation tends to have two distinct senses at different stages of assessment development. In the initial, *development stage,* 'validation' involves the development of assessment procedures that support the proposed

> interpretations and uses of the scores, and of evidence to support these proposed interpretations and uses; in this usage, 'to validate an interpretation or use' is to show that it is justified. This first usage implies an advocacy role, in the building of a case for the validity of a proposed interpretation, and tends to have a confirmationist bias.
>
> In the second, *appraisal stage,* 'validation' is associated with a critical evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate. ...
>
> I try to talk about validation mainly in the second of these usages, as an even- handed evaluation of the proposed interpretations and uses, but I accept the advocacy usage in some contexts. (p. 4; italics in original)

Until test development is completed, it is difficult or impossible to collect data to bolster the claims made by a validity argument. As such, the validity argument model is of little use in assessment research, such as the present research, which aims not to develop tests for production, but to investigate the *construct* validity of test formats.

It seems unlikely that a single concept of validity or operationalization of validation will be agreed upon in the soft sciences any time in the immediate—or, likely, distant—future, but the specific aspects of validity theory which form the rationale for the present research will be discussed in the next chapter.

2.7. Chapter Summary

This chapter has reviewed the relevant literature pertaining to the present research. It began with a brief overview of listening comprehension and then moved to a focused review of the psychological and linguistic literature on nonverbal communicative behavior, with particular attention paid to nonverbal communicative cues, which were broken into facial expressions, for communicating the speaker's affect, and gestures,

which can be used either to illustrate the verbal information, or to directly communicate concepts via sematic gestures known as “emblems.” Research on the role played by visuals in second language acquisition was reviewed before the literature on comparative studies between audio-only and video listening tests was presented. It was observed that in most cases, video has been shown to exert a facilitative influence on examinee scores in such studies, and various criticisms of the work completed on this topic up to present were outlined, including those pertaining to the comparability of studies, methodology, the quality of items used in the research, and the measurement models used, with a recommendation for work to adopt item response theory (IRT) as a way to investigate the impact of delivery format at the item level.

The author’s own contribution to the literature on this topic, which sought to address some of the problems noted in the existing studies, was then summarized. The present researcher then reviewed his own published work on this topic, highlighting questions raised by the work to be addressed by the present research. The central questions remaining, it was argued, are related to different task types.

Frequent additional topics addressed by the video listening test literature were then considered, demonstrating an especially important gap in the literature with regards to the visual attention to nonverbal and other visual cues paid by examinees in such tests. Reviews of the eye-tracking literature, of central importance to Study I, and the item response theory literature, central to Study II, will be presented with their respective studies.

Finally, as the present study is largely concerned with the construct(s) addressed by video-mediated listening tests, the literature on the shifting concept of validity over the past century was reviewed. Particular attention was paid to the

concept of construct validity, and it was demonstrated that the current, favored argument model of validation was not relevant to exploratory research such as the present project. Chapter 3 draws from this review of the literature to present a concise rationale for the work presented in the Study I and Study II chapters (Chapters 4 through 6), demonstrates implications of any findings for future test development, elucidates and explains the research questions motivating the studies in the present thesis, and lays out the design of the project as a whole, as well as that of each phase contained therein.

CHAPTER 3. DESIGN

This chapter presents the rationale and designs of the studies contained in the present thesis. It begins with an explication of the rationale for the research, followed by a section highlighting the implications of the work for test developers. Next, the research questions are posed and explained. Finally, the design of the project as a whole, as well as those of the phases, is presented. The chapter concludes with a brief summary.

3.1. Study Rationale

As stated above, the issues surrounding the use of video in foreign language listening tests are not related to any particular use of scores on a particular test, or the consequences thereof, but of something more basic: construct validity. The argument for the construct validity (not to be confused with a formal validity argument *a la* Kane, as discussed above) of video-mediated tests seems to be rooted in the observation that (for sighted people speaking face-to-face or via video chat, at any rate) we typically benefit from multimodal input, using the visual channel as a co-text for the verbal, and that, therefore, a test lacking this information is not valid, due to construct underrepresentation (Wagner, 2006, 2007, 2008, 2010a, 2010b). This argument hinges upon Messick's treatment of authenticity in his previously-discussed 1989 paper, and his later article in *Language Testing* applying his unified model of validity to the context of language assessment (1996). In it, he writes:

> ... in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviours of listening, speaking, reading and writing of the language being learnt. (p. 241)

The purpose of including nonverbal communication and other visual cues in a listening test, then, is to model reality and provide examinees with a more authentic

language experience so that scores are arguably more indicative of real-world ability, and therefore easier to interpret.

While this may seem reasonable at first, the issue of authenticity in testing is not so clear. Messick also cautions about construct *over*representation (although he does not use this term), or construct-irrelevant variance, wherein "the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct" (1996, p. 244). The question that remains, even after decades of research, is whether any effect, if present, of the presence of nonverbal information on a listening test is truly relevant to a construct of listening comprehension in which any test user is interested.

Furthermore, a blanket call for authenticity in testing seems to pre-suppose that an "authentic" test possesses some kind of "inherent validity" (Stevenson, 1985a, p. 42). Without careful and rigorous validation, however, this kind of "validity" is little more than *face* validity—the mere appearance of construct validity (Bachman, 1990)—or, even less flatteringly, "pop validity" (Stevenson, 1985b). According to the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education's (NCME's) *Standards for Educational and Psychological Testing,* validity must be established by accumulating evidence "to provide a sound scientific basis for the proposed score interpretations," and those interpretations must include "a rationale for the relevance of the interpretation to the proposed use" of the scores (2014, p. 11). Few tests truly meet these criteria, but the research thus far into the construct of listening addressed by video-mediated listening tests falls far short. For all the talk of the psychological importance of nonverbal cues in the literature reviewed above, little

work has been done to develop a theory of how such information impacts scores, which contradicts the AERA, APA, and NCME Standard 1.12:

> If the rationale for a test use or score interpretation depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. (2014, p. 26)

As Suvorov notes, most literature on this topic compares two forms of a test, and if a score difference is observed, that difference is attributed to visual cues (2015), but it is still unclear why or how that difference arises (or why it sometimes does not).

Without a closer examination of examinee behavior whilst interacting with a video-mediated listening test, and the interaction between task type and format at a greater degree of detail, it will remain unclear how and where the inclusion of visual cues impacts item scores. By employing eye-tracking methodology at a more granular level than in the Suvorov study (2013, 2015), a more-detailed theory of examinee attention to visual cues can be developed and applied to test design. By investigating item-level score differences with probabilistic models of item response on a test that was created based on a deeper understanding of examinee attention to visual cues, even small effects attributable to the video format can be isolated for scrutiny. Such work would help put to rest concerns over the construct validity of such tests (e.g., Bachman, 1990; Buck, 2001; Taylor & Geranpayeh, 2011) and aid in the development of the more robust and theoretically-sound definition of a construct of visual listening comprehension called for in some of the literature reviewed above (e.g., Batty, 2015; Ockey, 2007).

3.2. Implications for Test Development

A clearer understanding of the comparability of constructs addressed by traditional foreign language listening tests which are audio-only and by those which include video would have a clear, positive impact on test development, in that it would allow test designers to write test specifications from a theoretically-sound position, with a clear purpose and domain for the instrument (AERA et al., 2014). This is not possible without greater understanding of what visual cues attract the most conscious attention on the part of examinees, and further insight into if, when, and how the presence of this information interacts with items to result in examinee item responses. Without this evidence, it will remain unclear as to why an examinee answers an item correctly on such instruments, and whether the items truly test what the test developers intend (Alderson et al., 1995; AERA et al., 2014). A more detailed understanding of the construct of foreign language listening comprehension that is addressed by video-mediated tests of foreign language listening comprehension would enable test designers to make informed decisions on whether and how to incorporate video in their instruments in the future.

3.3. Research Questions

Based on the preceding rationale for the present project, the following research questions are posed:

1a. *What are the specific nonverbal or visual cues to which L2 examinees attend when taking a video-mediated test of foreign language listening comprehension?* Although much research has been conducted into such questions as how long examinees view the videos in such tests and/or the visual cues they mention in verbal reports, no one has employed more sophisticated methods to determine to

what nonverbal or other visual cues examinees actually attend in video-mediated listening tests.

1b. *How does viewing behavior of visual cues differ with respect to videotexts?*

It is likely that viewing behavior is dependent upon the content of the video and the question asked. Differences could indicate different approaches to comprehending the material based on the question asked. It would be beneficial to know what differences from one item to the next can be attributed to the content of the listening videotext, or to the content of the item.

1c. *How does viewing behavior of visual cues change with respect to task type (explicit and implicit)?*

It has been suggested by some in the literature that visuals may play a larger role in responding to implicit questions than to explicit. The present research will attempt to demonstrate that difference, if it can be shown to exist.

2a. *How does the presence of visual cues interact with items on video-mediated listening comprehension tests?*

Although others in the audio-video comparative literature have calculated item facility statistics, etc., for individual items on equivalent audio- and video-mediated listening tests, none have taken advantage of IRT's benefits for this kind of analysis.

2b. *How does the presence of visual cues interact with task types (explicit and implicit) to influence item responses on video-mediated listening comprehension tests?*

Although comparisons between item-equivalent video- and audio-mediated listening tests abound in the literature, none attempt to investigate the interaction

between format and task types requiring different levels or types of comprehension at the item level.

3a. *How do individual examinee differences interact with the presence of visual cues on video-mediated listening comprehension tests?*

Although individual differences between examinees in terms of proficiency, gender, or simple preference may cause any impact of the presence of nonverbal communicative or other visual cues on scores to be non-uniform, this has been little-studied. Further research is warranted to determine whether and how individuals' performance is benefitted or hindered by the presence of visual cues in the form of video in listening tests.

3b. *How do examinee perceptions of video interact with performance on video-mediated listening comprehension tests?*

It is possible that examinees' perceptions and preferences regarding the presence or absence of visual information in listening tests interact with their performances on such tests, resulting in differential person functioning with respect to the format of delivery. This relationship can be explored through the use of many facet Rasch measurement.

The next section describes the research design by which these questions are to be addressed.

3.4. Research Design

This section describes the research design of the present thesis. It begins with an explanation of the overall design of the project, then outlines the design of each phase of separately. A graphical representation of the design can be seen in Figure 3.1.

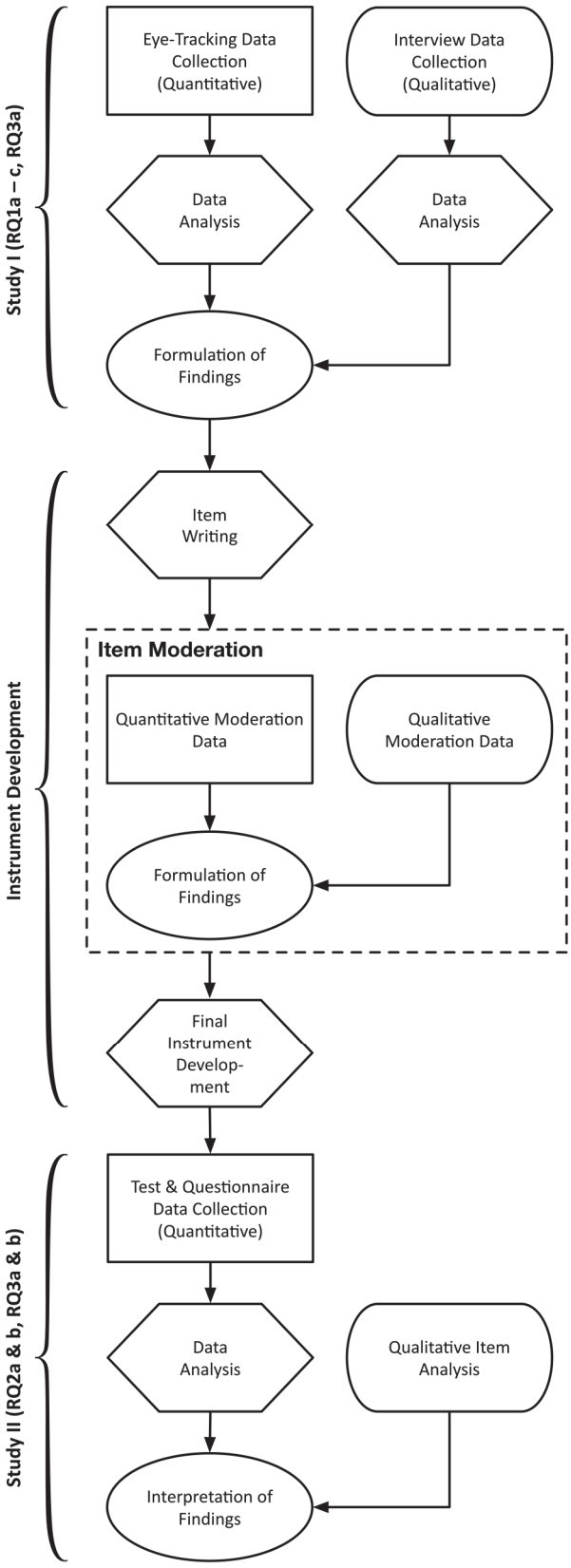

Figure 3.1. Graphical representation of the project design.

3.4.1. Design of the Project

The present research employs a *fixed mixed-methods exploratory sequential design.* Mixed methods designs are those that combine quantitative and qualitative research methods to address a research problem more completely than either method could address it alone. The quantitative and qualitative data can be used in tandem, or used sequentially—the findings of one informing the collection of the other, or one can be embedded within the other. It frequently uses these methods in single studies or in separate stages of a larger study (Creswell & Plano Clark, 2010).

This design is a *fixed* mixed-methods design because the "use of the quantitative and qualitative methods is predetermined and planned at the start of the research process, and the procedures are implemented as planned" (Creswell & Plano Clark, 2010, p. 54). The two studies in the present thesis address mostly different research questions, each drawing from both quantitative and qualitative data. The research is *exploratory* because Study I, although it includes quantitative data, is ultimately most concerned with collecting fairly qualitative data on intentional direction of attention to specific nonverbal and visual cues, rather than setting out to demonstrate or explain a pre-existing hypothesis about this behavior. The present thesis is *sequential* because the results of the eye-tracking study (Study I; Chapter 4) feed into the Study II instrument development (Chapter 5) and inform interpretations of the quantitative results of the Study II (Chapter 6) data analyses (Creswell & Plano Clark, 2010).

3.4.2. Design of the Phases

Each phase of the thesis, too, follows a mixed-methods approach. Study I employs a *fixed convergent parallel mixed-methods design,* wherein the quantitative and qualitative data are collected during the same phase of the study and are given equal

importance when formulating findings (Creswell & Plano Clark, 2010). The quantitative measures collected by the eye-tracking hardware is combined with the qualitative verbal reports of the participants' structured interviews as they watch playbacks of their viewing behavior. The former is subjected to statistical analysis and the latter, to qualitative data analysis with NVivo (*NVivo qualitative data analysis software*, n.d.).

The instrument development phase contains an item moderation step to ensure that the present author's classifications of items into "explicit" and "implicit" task types are agreed upon by expert judges. This step employs a *fixed embedded mixed-methods design,* wherein qualitative data are added to quantitative data to aid interpretation, but is not weighted as heavily (Creswell & Plano Clark, 2010). In this case, qualitative data in the form of notes from the item moderators is referred to as a secondary source of data, but the main data source is the respondents' answers to the items and their classifications into the types.

Finally, Study II also utilizes a fixed embedded design with the inclusion of qualitative item analysis in order to assist in interpreting the results of the quantitative data analysis. Whereas the bulk of the data analyzed in Study II is quantitative, whether it be item responses or selected responses from a questionnaire, a qualitative item analysis is also carried out in order to better interpret the findings of the various statistical analyses employed. Using these two data sources, RQ2a and 2b, and 3a and 3b, can be fully addressed.

As such, the design of the present research project, which spans over two studies and an intermediary instrument development step, employs mixed-methods approaches both at the project level, and during every phase of the project as a whole.

3.5. Chapter Summary

This chapter outlined the central issues of validity which serve as the rationale for the present research. It then highlighted the implications of any findings of the present research to test development in the future. Next, it posed and explained the research questions. This was followed by a detailed explanation of the overall design of the project, and then an introduction of the designs of each of the project's three phases. The next chapter begins with an introduction to eye-tracking methodology and a review of the relevant eye-tracking literature before it describes Study I, reports its results, and discusses its findings.

CHAPTER 4. STUDY I

This chapter describes Study I, which sought to address RQ1a through 1c (those pertaining to viewing behavior) and RQ3a, "How do individual examinee differences interact with the presence of visual cues on video-mediated listening comprehension tests," via the use of eye-tracking methodology for the purpose of informing test design for Study II. The study employs a fixed convergent parallel mixed-methods design, wherein the uses of both the quantitative and qualitative data are determined from the outset, and both are given equal weight in the formulation of findings (Creswell & Plano Clark, 2010). Study I also conforms to a nomothetic approach, wherein the focus is on the population of interest at large, rather than on a particular individual (Duchowski, 2007). Additionally, eye-tracking was employed in a diagnostic capacity, i.e., to provide "objective and quantitative evidence of the user's visual and (overt) attentional processes" (p. 205). Data were collected in October of 2014. Participants sat a six-item video-mediated listening test while wearing a head-mounted eye-tracking device, followed by a structured interview based on gaze displays.

First, eye-tracking methodology will be introduced and literature applying the methodology to language testing will be reviewed. Next, the method of Study I will be explained, beginning with descriptions of the participants and equipment. The development of the Study I instrument is then described, including rationales for the design choices made. The data collection procedure is then presented. Following the method section is the data analysis section, which begins by relating the process of data transformation necessary to facilitate quantitative and qualitative analyses. These methods are then elucidated. The results section follows, and presents an overview of the observed behaviors, followed by comparisons between participants both

individually and by proficiency and gender. Differences between the videotexts are then explored. Finally, the interaction between task type and viewing behavior is presented. The chapter concludes with a discussion of the findings and their implications for Study II instrument development.

4.1. Eye-Tracking

This section introduces eye-tracking and its application to language and language assessment research. Eye-tracking methodology allows the researcher to capture and plot the movements of participants' eyes by technological means, and has been in use since the late 1800s. It has mostly focused on describing eye movements while reading, and/or attempting to infer the cognitive processes underlying them (Rayner, 1998; Tanenhaus, 2007), on the assumption that eye movement indicates voluntary visual attention (Duchowski, 2007) and can therefore be used to infer cognitive processes, an assumption predicated upon the eye-mind hypothesis (Just & Carpenter, 1976). In recent decades, however, as eye-tracking apparatuses have improved with ever more sophisticated computational resources, its use has been expanded through fields as diverse as neuroscience and marketing, increasingly focused on interactive tasks (Duchowski, 2007).

4.1.1. Eye Movements and Metrics

Eye-tracking research typically concerns itself with a number of eye movements, such as *fixations,* when the eyes stop over an object of interest; *saccades,* when the eyes quickly move to a new position in the field of vision; and *smooth pursuits,* when the eyes follow an object in motion (Duchowski, 2007). How importance is assigned to these movements, however, is typically related to their interaction with *areas of interest* (AOIs), which are simply areas in the visual field in which the researcher is interested (Holmqvist et al., 2011).

The eye movements most commonly tracked in relation to AOIs tend to be the percentage of time the eyes are oriented toward individual AOIs, the number of fixations per AOI, and average duration of time spent looking at an AOI (Jacob & Karn, 2003). Of particular importance to the present research is the last of these, which is typically referred to as *dwell time,* and includes any fixations, saccades, or pursuits associated with it. The meaning of a dwell can change according to the research in question, but dwells are frequently understood to indicate viewer interest in the AOI and/or its informativeness. When dwell time is summed over the course of an observed activity, it is referred to as *total dwell time.* Another event of particular importance to the present research is the *transition,* which is simply the movement of the eyes from one AOI to another (Holmqvist et al., 2011). The path eyes follow around an image or video over time is usually referred to as a *scanpath,* and was the primary subject of analysis for much of the early eye-tracking work, although the term was not coined until the early 1970s (Duchowski, 2007; Holmqvist et al., 2011).

4.1.2. Eye-Tracking Methodology

Eye-tracking research employs the use of hardware and software to measure the movement of the eyes, store them, and usually provide data and visual outputs, such as location data in tabular format and/or visualizations, usually superimposed over the image with which the respondent was interacting.

Eye-tracking hardware comes in many forms, from direct measurement of eye movements by way of electromagnetic fields or the use of specialized invasive contact lenses, to the much-more-common video-camera-based systems currently in wide use (Duchowski, 2007). The camera-based hardware can take the form of specialized cameras affixed to a computer monitor (so-called “remote” units) or units mounted directly to the head, usually in the form of glasses or goggles (Holmqvist et al., 2011).

Both approaches have benefits and drawbacks. Monitor-mounted systems are non-intrusive, but can only track the eyes when the user's head is oriented toward the screen—a limitation which had some deleterious effects on the data collection in the Suvorov studies (Suvorov, 2013, 2015). Head-mounted units, however, can track the eyes regardless of where the user moves his or her head, but, as they require the user to affix some manner of apparatus to his or her body, may be distracting. These hardware systems function at a wide range of sampling frequencies (i.e., how many times per second that a measurement is recorded), ranging from 25 Hz all the way through 2,000 Hz. Generally speaking, faster systems are intended for research concerned with cognitive processes, whereas slower systems are sufficient for more straightforward research on participants' voluntary attention to visual stimuli (Holmqvist et al., 2011).

Software for such systems also varies widely in terms of both data capture and output, from packages capable of little more than capturing *x* and *y* coordinates for gaze positions in a visual space, to those capable of automatically determining the amount of time the user looked at dynamically-changing areas in the visual field. Nearly all software can also superimpose the captured scanpath over a recording of the subject's visual field, using a number of visualizations from simple lines through "heat maps," which reveal or occlude portions of the image based on dwell time in the area. In research with fixed AOIs and remote capture hardware (e.g., reading studies or computer user-interface usability studies), the AOIs can be defined in the software, and dwells, etc., can be automatically tracked and logged numerically.

For more complex AOIs (e.g., semantically-determined regions which move in the visual field), manual scanpath analysis may be necessary, although such analysis necessitates the addition of verbal data to arrive at useful interpretations of the

observed behaviors (Holmqvist et al., 2011). Such verbal data can come in many forms, but cued retrospective reporting, wherein the participant is shown scanpath-overlaid videos of his or her behavior after the captured activity, and is asked to explain his or her thinking and actions, has been found to be particularly effective at providing insight into the meaning of the observed behavior (van Gog, Paas, van Merriënboer, & Witte, 2005), while concurrent thinkaloud methods have been criticized as deleterious to authentic eye movement capture (Bojko, 2005). Both Pernice and Neilsen (2009) and Ehmke and Wilson (2007) argue for a slightly more focused version of cued retrospective reporting: a structured interview of the participant by the researcher as both watch the captured video together. In either case, however, 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.

4.1.3. Eye-Tracking in Listening Comprehension Research

Although most eye-tracking research both in L1 and L2 contexts has been focused on reading, for the obvious reason that it is a visual activity, there have been a number of influential studies of eye movements while comprehending spoken language. Listening studies are simultaneously more complex and yet somewhat simpler than the much-more-common reading studies, however, as the behavior is not based on a regular set of movements intended to directly comprehend linguistic meaning, obviating the use of many of the very precise metrics often found in the reading literature in favor of the more coarse metrics related to dwells and transitions (Boland, 2004).

Broadly speaking, it has been demonstrated that listeners focus their visual attention on objects as they appear in the linguistic stream (Allopenna, Magnuson, &

Tanenhaus, 1998; Cooper, 1974), and their eyes orient to the last known position of that object regardless of whether it is still visible (Altmann, 2004). This process occurs even as the listener anticipates the object to be mentioned based on the syntax and semantics of the sentence (Altmann & Kamide, 1999; Boland, 2005), and 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 in the very beginning of an activity with the first few fixations (Boland, 2004; Rayner, 1998). Overall, this research generally supports the eye-mind hypothesis, in that people's eyes tend to orient toward objects as they begin to think of them (Boland, 2004).

4.1.4. Eye-Tracking in Language Assessment Research

Eye-tracking methodology has seen limited use within the field of language assessment, although there does appear to be increasing interest. As in the general eye-tracking research, attention has mostly been focused on reading tasks.

Although not in the L2 assessment context, Gorin (2006b) employed eye tracking to assist in item difficulty modeling of paragraph comprehension questions on the Scholastic Aptitude Test (SAT), one of the standardized tests used in the United States for college admission. By tracking the examinees' eye movements, Gorin was able not only to catch a glimpse of their cognitive processes whilst answering the items, but, in one case, identify an item which an examinee was able to answer correctly without referencing the text at all. Such information could be invaluable to a test developer, as it has direct consequences on the validity of the test scores with regard to the construct of reading comprehension.

Bax and Weir (2012) have employed eye-tracking and questionnaires to investigate examinees' cognitive processes and the degree to which they matched the

processes targeted by items from the Cambridge English: Advanced (CAE) reading test—the so-called "cognitive validity" of the test (Khalifa & Weir, 2009). Data were analyzed from ten examinees, selected from a population of 35 who used the eye-tracker, among 108 who took the test, on thirteen items from four texts. After answering each item, each participant responded to a brief questionnaire about their cognitive process while answering. Overall, the findings indicated strong support for the cognitive validity of the CAE in that examinees tended to read sections and focus on words as expected. However, they also found that in 31.6% of cases, the reported cognitive processes in the questionnaires did not match the observed eye movements, calling into question the efficacy of such a data collection method.

Bax (2013) has further employed the methodology in his investigation of the differences in cognitive processing of lower- and higher-proficiency examinees responding to two IELTS (International English Language Testing System) reading passages, administering eleven reading items to 38 eye-tracked examinees, twenty of which later sat for cued retrospective reports (called "stimulated recall interviews" in the Bax article). Significant eye movement differences between more- and less-proficient readers were observed in five of the items, arguably demonstrating the cognitive validity of the IELTS reading section.

Most recently, Brunfaut and McCray (2015) have used eye-tracking to investigate the cognitive processes of examinees on the Aptis reading test with regards to examinee proficiency, CEFR level, and task type. Twenty-five participants sat for fifty Aptis reading items while using the eye-tracker, followed by cued retrospective reports (referred to as "stimulated recall sessions" in the study). Of particular importance to the present research, the recall sessions were carried out in the participants' L1s to ensure clear and easy communication of their recollections and

explanations of their viewing behaviors. Following the eye-tracked reading test, participants sat the full Aptis test to provide a separate measure of overall L2 proficiency. Although the study noted construct validity issues with some items, it concluded that the Aptis reading section tests a wide variety of processes, as it is intended to do.

As of this writing, the only application of eye-tracking to foreign language listening assessment available appears to remain Suvorov's PhD thesis (2013), and the *Language Testing* article resulting from it (2015), which has already been discussed in detail in the "Examinee watching behavior" section of Chapter 2. Given the encouraging findings of research employing this methodology to questions of reading assessment, clearly, more applications of it to listening assessment are warranted.

This section has offered an overview of the fundamentals of eye-tracking methodology, and reviewed the research applying that methodology to language comprehension and listening assessment. The next section describes the method of Study I, which will apply this methodology to RQ1a through 1c and RQ3a.

4.2. Method

4.2.1. Participants

Participants were twelve (12) Japanese students of English studying at Keio University's Shonan-Fujisawa Campus (Keio SFC) in Fujisawa, Kanagawa, Japan. They were recruited from English classes and paid ¥1,000 (approximately £7.60 as of this writing; minimum wage in Kanagawa prefecture was ¥880/hr. at the time of data collection) for their participation. All had studied English for a minimum of six years, beginning in middle school.

The majority of participants (n = 10) were female, the result of several unrelated factors. One male participant completed the task, but the test equipment

failed shortly after he began and no data were saved. This did not become apparent until after he had completed the task, and as a result he could not sit the test again, having already completed answering the questions.

Additionally, in order to avoid any possible effect related to cultural differences in attention to facial features (Yuki et al., 2007), respondents were intended to be Japanese only. In the course of interviews of two male participants, it became clear that, although both were in the standard Japanese track (Keio also offers an English-mediated track for foreign students), one had grown up in Korea and another in China. Although their behaviors were not appreciably different from their Japanese peers, their responses were omitted from data analyses. Finally, anecdotally, it is frequently difficult to enlist Japanese males as participants in research, even when it is paid. This likely serves to explain why out of an original five male volunteers, two were acculturated outside of the Japanese context.

Keio SFC places English students into three levels of classes using the Educational Testing Service's (ETS') Test of English as a Foreign Language Institutional Testing Program (TOEFL ITP), ensuring that all participants had an external measure of general English proficiency. The TOEFL ITP is a standardized multiple-choice test of American academic English intended to assist institutions in placement, progress monitoring, and other such uses, covering listening comprehension, structure and written expression, and reading comprehension (Educational Testing Service, 2016, 2016). It is administered at and by the institution in question (Educational Testing Service, 2016). Cut scores for the Common European Frame of Reference (CEFR) levels are published by ETS (Tannenbaum & Baron, 2011) and are as follows (Table 4.1):

Table 4.1.

TOEFL ITP Total Cut Scores Mapped to CEFR Levels

TOEFL ITP Cut Score	CEFR Level
627	C1
543	B2
460	B1
337	A2

Recruitment of participants became more targeted for proficiency as data collection proceeded to prevent over-representation of some score ranges. However, it is important to note that English is not a required subject at Keio SFC, and the English courses are better understood as electives. As such, students are not explicitly expected to advance through the levels of the program, and the TOEFL ITP is therefore typically only taken, at student expense, in the first semester of study. For this reason, with older students, the scores may not be representative of current proficiency, as students may have improved or worsened in the intervening years. However, of the twelve final participants, six were first-year students, meaning the TOEFL score listed can be understood to be a fairly accurate estimate of overall proficiency for the purpose of sample targeting. Participant demographics are displayed in Table 4.2.

4.2.2. Equipment

The eye-tracker used for this study was the open source, head-mounted Pupil Dev eye tracker (Kassner & Patera, 2012; Kassner, Patera, & Bulling, 2014), assembled by the present researcher according to the device specifications. The Pupil Dev headset uses dark pupil detection via a camera mounted below the user's right eye, with a resolution of 640 × 480 pixels and a refresh rate of 30 Hz. Gaze points are mapped to video captured by the world camera mounted above the right eye, with a resolution of

Table 4.2.

Study I Participant Demographics

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

1280 × 720 pixels and a refresh rate of 30 Hz. The system is calibrated by a nine-point calibration routine available through the data capture program Pupil Capture, which also captures the gaze positions, world video, pupil video (optional), and audio (optional). Visualization of the data is achieved via the partner application Pupil Player.

The Pupil hardware and software were installed and configured on a 2013 Apple Macintosh Mini running Ubuntu Linux 14.04 LTS. The computer was outfitted with two monitors and two keyboards and mice, allowing the researcher to operate the eye tracking software outside the view of the participant. The experimental monitor was a 21-inch widescreen monitor on an adjustable mount, allowing for a wide range of heights to allow participants to view comfortably and with the entire screen in the view of the Pupil world camera. The researcher's monitor was occluded from view by a large cabinet next to the participant's desk. A threaded curtain also separated the experimental area from the researcher's area. The experimental monitor was surrounded by black foamboard panels obscuring the cables and computer from view, in order to prevent distraction (Goldberg & Whichansky, 2003). This foamboard

partition extended around to either side of the participant as well. The participant's view was thus limited to the monitor, the speakers, the keyboard, and the mouse (see Figure 4.1).

Mozilla Firefox displayed the test content, Kazam screen capture software recorded post-test interviews using the Pupil headset's microphone for voice recording and an Apple iPhone used as a backup voice recorder, and VLC media player software displayed the eye-tracking videos during post-test interviews. The Ubuntu menu bar clock was deactivated to reduce distraction during testing.

Figure 4.1. The participant's experimental desk.

4.2.3. Instrument

The instrument was comprised of six short videos, with one multiple-choice item per video. Video durations ranged from 38 seconds through one minute and 39 seconds. The videos were selected scenes from the American television comedy of manners *Curb Your Enthusiasm,* chosen for the fact that this program is improvised. Scene outlines are prepared and the actors improvise the final dialog on camera. This results

in pseudo-authentic language use, complete with hesitation, misspeaking, and false starts, common features of unplanned speech (Shohamy & Inbar, 1991). Furthermore, owing to the improvisational technique, nonverbal behavior such as gesture and facial expression use also closely resemble authentic language production. Finally, the series features no "laugh track" and most scenes have no background music, increasing the material's realism. Scenes were chosen wherein only a man and a woman appeared, although one scene, "Cheryl's Call," also featured a voice on a speakerphone. This allowed items to reference "the man" and "the woman" rather than expecting participants to remember character names.

The instrument was administered via the quiz module in Moodle 2.6, and began with an informed consent item prior to the test content, in compliance with the Lancaster University ethical guidelines, the project having been granted full ethical approval by the Lancaster ethical review board. Each video was preceded by a preview of the item stem in Japanese. The participant then viewed the video one time, and was then presented with the stem again, along with four answer options, also in Japanese. After answering, the participant advanced to a preview of the next item. A list of videos and task types is displayed in Table 4.3. The following subsections detail the specific choices made in instrument design.

4.2.3.1. Item preview

The choice to preview items before engaging the listening text was in accordance with the recommendation of Buck (1991), who, in his introspective study of listening test examinees, found that item preview gave the listeners enough information to frame the questions. It is rare that one listens to anything, least of all a television program, without context or purpose. By providing the purpose in the form of the question stem, the examinees did not have to struggle to remember every detail

Table 4.3.

Study I Video Content and Task Types

Name	Length (m:ss)	Task	Content Summary
Shopping	1:35	Explicit	A man returns from an unexpected clothes-shopping trip; his wife complains that he does not go shopping with her.
Ted and Mary	0:45	Explicit	A man complains that some friends have stood he and his wife up for a night out; he opines that it would be better for them to lie about an excuse than to say nothing.
Football	0:38	Implicit	A woman returns home from a trip and wants to tell her husband about it; the man is more interested in watching an American football game on TV.
Dinner	1:39	Implicit	A man and a woman argue about whether to stop for a drink before going to a restaurant for dinner; the man does not understand why his wife wants to do so.
Salesgirl	1:28	Explicit	A man is browsing in a shop; a female salesperson follows him despite his repeated assurances that he is merely browsing.
Cheryl's Call	0:53	Implicit	A couple is riding in a car when the woman receives a phone call from another man on the car's speakerphone; the man finds it strange that his wife informs the caller that her husband is in the car.

of the conversation, only that which was helpful for the task at hand. This choice likely resulted in easier items, as those who have investigated item preview for audio-mediated listening multiple choice tests (Berne, 1995; Chang & Read, 2006; Elkhafaifi, 2005; Gries & Wulff, 2005; Yanagawa & Green, 2008), and video-based tests (Koyama, Sun, & Ockey, 2016) have repeatedly found. Furthermore, removing the items from view as the examinee watches the video reduces the likelihood that he or she will come to consider the video as a distraction, as attention is flipped repeatedly between the video content and the item, as was reported by participants in several of the studies reviewed in Chapter 2 (Coniam, 2001; Cubilo & Winke, 2013; Suvorov, 2013). It also matches the Wagner and Batty methodologies (Batty, 2015; Wagner, 2007, 2010a, 2013). Unlike these studies, however, examinees did not have access to the questions or options during the video, ensuring that their attention was focused only on viewing the material.

4.2.3.2. Number of items per video

The decision to limit the number of items to one per video was made for two considered reasons. In the cases of the Wagner (2007, 2010a), Ockey (2007), Batty (2015), and Suvorov (2013, 2015) studies, which used "testlet" designs, with multiple items per video text, it is impossible to fully investigate the items individually, as they are not locally independent. The item responses will be confounded, especially since 33% of item variance in such testlet designs has been attributed to the content of the input alone, as opposed to any feature of the items themselves (Freedle & Kostin, 1999). By ensuring that each item's response is attributable to only one text, the items and the viewing behavior associated with them can be understood to be independent of each other, similar to the ETS Test of English for International Communication (TOEIC), but in contrast to the "testlet" design of the TOEFL Internet-Based Test (iBT). A further benefit of this design, which is especially important for an eye-tracking study, is that it obviates the need for note-taking, which would result in the difficulty in interpretation of gaze behavior noted by Suvorov (2015), as examinees' gaze moves back and forth between the video and the notepad, rather than staying fixated on the video input only.

4.2.3.3. Language of items

The choice to present item stems (as well as all instructions) in Japanese is also a considered one. Filipi (2012) demonstrated that, although examinees do not perceive a difference, items presented in the L2 are harder than their L1 counterparts. Although it is clear that this increased difficulty is indeed related to the overall language proficiency of the examinees, it is outside the construct of foreign language listening comprehension. This may be of little concern when listening is measured as a part of overall proficiency, or may be unavoidable in a test intended for speakers of

multiple L1s. However, as the present study seeks to investigate the behavior of examinees under the best possible and purest circumstances, the choice was made to eliminate any ambiguity arising from the possibility that examinees misunderstood the items, and therefore the purpose of listening. Notably, the same decision was also made in several of the studies reviewed in Chapter 2 (see Baltova, 1994; Chung, 1994; Parry & Meredith, 1984).

4.2.3.4. Task types

In order to investigate any difference in viewing behavior when confronted with different item task types, three explicit items and three implicit items were selected from a list of possible items written by the present researcher for each video. For the purposes of the present research, the task types are operationally defined as below:

- *Explicit items.* Any item whose answer appears explicitly in the conversation. An example would be the first item of the test, "Shopping," whose item was "What did the man buy?" and the conversation included one character saying "You got a jacket?", to which the other character replied, "Yeah. You like it?"
- *Implicit items.* Any item whose answer does not appear explicitly in the conversation, but which can be surmised by the conversation. An example would be the fourth item on the test, "Dinner," which was an argument between a husband and wife about dinner plans. The wife wanted to stop somewhere to have a drink and socialize before moving to another restaurant for dinner, and the husband argued that she could have a drink at the same restaurant where they ate dinner. The question was "Why doesn't the man like the woman's suggestion?" and, although he never explicitly states it, the correct response was, "He thinks it is pointless."

It is important to note that no items were written in such a manner that they required the visual content in order to answer correctly. For example, in the third item, "Football," the word "football" never appears in the scene, but the male character is watching an American football game. Any item specifically referencing the sport he was watching would be understood to require the visual channel, and as such, no such items were written.

4.2.4. Assistant

To ensure expedient data collection, an assistant aided in testing and interview sessions. Two assistants were used, depending on schedules. Both were adult female native speakers of Japanese. One was a graduate student at Keio SFC, and was paid according to the rates set by the university at the time (¥900/hr.; approximately £6.80 currently), and the other was the researcher's wife, who donated her time.

4.2.5. Procedure

The data collection procedure was developed according to the recommendations of Pernice and Nielsen (2009), and piloted on the assistant described above, and one other volunteer participant. Based on these experiences and the recommendations of the volunteers, a final procedure was developed, described below. The data collection procedure checklist can be seen in Appendix A.

The participant was first introduced to the present researcher and the assistant, making some "small talk" to put the participant at ease. The participant was seated at a desk in a plain chair with no tilt or swivel functions and the Pupil headset placed on him or her (see Figure 4.1). The Pupil headset was disinfected with an alcohol wipe in view of the participant, in order to allay any fear that he or she may have had about sharing a device that touches the face, and placed on the participant. USB cables were routed behind the participant's back and he/she was instructed to lean back to restrict

movement and anchor the cables for comfort. In some cases, owing to differences in facial geometry, risers constructed of Sugru-brand molding rubber were placed under the nosepieces of the headset to ensure visibility and good pupil detection.

4.2.5.1. Eye-tracking data collection

As per Pernice and Nielsen's (2009) recommendations, the participant was then instructed to click the link to complete a practice task without the eye-tracker running. The task opened with a brief introduction to the main characters of the videos as a form of schema setting (Buck, 2001; Imhof, 2010; Rost, 2011; Wolvin, 2010; Wolvin & Coakley, 1996), and then presented a single practice item identical to the format of those described in the Instrument section. The purpose of a practice item is to familiarize the participant with the format and the equipment before actual data collection, and to allow the participant some time to settle into his or her natural comfortable position at the desk. It also offered the opportunity for the researcher to adjust settings in the Pupil Capture software before initial calibration. The most common of these were settings determining the maximum and minimum pupil sizes, which are highly individual.

After finishing the practice item, the assistant returned to the experimental area, showed the participant how to close the activity, and addressed any adjustments necessary (e.g., monitor height, chair distance). At this point, the Pupil nine-point calibration routine was commenced. If an adequate lock did not occur, the calibration was run again. As per Pernice and Nielsen's recommendations, care was taken to phrase this without any implication of fault on the part of the participant. The researcher would simply say, "Okay, here it comes again" in a cheerful tone, without revealing that it was normally intended to be run only once.

Once calibration was complete, the participant was instructed to begin the test, and the Pupil Capture software was set to record. The participant then agreed to the onscreen consent form and began the test. As the participant watched the videos, the present researcher watched his or her gaze positions (transitions and fixations) on his monitor, making note of behaviors to address in the post-test interview.

Following the test, the assistant entered the experimental area, thanked the participant, and helped remove the headset. She would then engage in more small talk about the test (e.g., "So, had you ever taken a test like that before?", "Is this your first time doing eye tracking?") as the present researcher exported the scanpath-overlaid video for use in the upcoming retrospective interview.

4.2.5.2. Retrospective verbalization data collection

Retrospective verbalizations were collected via semi-structured interviews based on gaze displays, as described by Holmqvist et al. (2011), Pernice and Nielsen (2009), and Ehmke and Wilson (2007). Once the video was prepared, the researcher returned to the experimental area and explained that they would be watching the eye tracker video together and that he would stop it periodically to ask what the participant remembered about his or her behavior, following the general guidelines described by Pernice and Nielsen (2009) and by Ehmke and Wilson (2007). Prior to watching each video, the researcher placed in front of the participant a laminated card with the question that was asked and the correct answer, to help guide the conversation. All interviews were conducted in Japanese. Care was taken to use similar phrasing of questions between participants. See Appendix A for the basic interview script. Example questions include the following:

1. Please try to recreate your thought processes here. What are you looking at and why?

これを見たときの思考のプロセスを振り返ってみてください。何を見ていました？それはなぜですか？

2. That's interesting! Why did you look there at that time?

   面白いですね！なぜこのとき、そこを見たのですか？

3. How does looking there help you answer the question?

   そこを見ることが、質問を答えるうえで、どんなヒントになったと思いますか。

4. Was there anything in particular you were looking for to help answer the question?

   質問を答えるために何かを特別に探していましたか？

Once the interview was complete, the participant was thanked for his or her time, paid, and offered some wrapped Belgian chocolates as a further sign of gratitude. Interview audio and video were captured on the computer for later analysis.

4.3. Data Analysis

4.3.1. Data Transformation

Prior to analysis, the data required considerable preparation. The data were of two types: quantitative eye-tracking data and qualitative interview data.

4.3.1.1. Eye-tracking data

The scanpath-overlaid videos exported from Pupil Capture were imported into the Apple Final Cut Pro X video editing software package (*Final Cut Pro*, 2015), and each full session video was cut into six short videos of the participant's interaction with each of the six videotexts. Due to the fact that the Pupil eye-tracker is a head-mounted device, much of the captured video area was unnecessary for analysis, and was frequently crooked, owing to the participants' natural postures. For this reason, the video was cropped and straightened before being exported for further analysis.

The scanpath-overlaid videos were imported into QSR International's NVivo 10 qualitative data analysis software (*NVivo qualitative data analysis software*, n.d.) for manual scanpath analysis. Manual analysis is well-established in the literature in cases where human perception remains superior to that of software (Holmqvist et al., 2011), and was required here due to two factors:

1. The AOIs were not fully defined prior to data collection, and were determined through realtime viewing of the scanpath video.
2. Head-mounted eye-trackers without head-tracking functionality typically require manual coding of AOIs, as even the best eye-tracking visualization software is inadequate to track dynamic AOIs whose position is constantly in motion (Holmqvist et al., 2011).

Each eye-tracking video was coded in NVivo for the participant, the videotext, and the task type. The list of oculomotor events of interest (e.g., looking at eyes, looking at mouth, scanning face, looking at gesture) was used to code the videos, adding new events as deemed necessary, and those codings added to previous videos. The final list of behaviors deemed frequent enough to warrant coding can be seen in Table 4.4. Events were coded at NVivo's maximum resolution of one-tenth of a second.

The resulting matrices of events and other codes (referred to as "nodes" in NVivo) were exported to Microsoft Excel tables in the form of counts and durations. As simple durations are difficult to interpret, these durations were divided by the total video lengths to produce percentages of the video duration represented by the events, following the methodology of Suvorov (2013, 2015). These percentages can be understood as measures of total dwell time. The NVivo exports were collected into datasets suitable for statistical analysis.

Table 4.4.

Visual Cues Coded in Qualitative Data Analysis

Name	Description
Face	Any period of fixation on the face; may include those not captured by levels below; aggregates levels below
Speaker's Face	Any period of fixation on the speaker's face; aggregates levels below
Speaker's Face Regions	Any period of fixation on the speaker's eyes or mouth; aggregates levels below
Speaker's Eyes	A period of fixation on speaker's eyes
Speaker's Mouth	A period of fixation on speaker's mouth
Speaker's Face Scan	Rapid transitions with short fixations around the speaker's face
Listener's Face	Any period of fixation on the listener's face; aggregates levels below
Listener's Face Regions	Any period of fixation on the listener's eyes or mouth; aggregates levels below
Listener's Eyes	A period of fixation on listener's eyes
Listener's Mouth	A period of fixation on listener's mouth
Listener's Face Scan	Rapid transitions with short fixations around the listener's face
Alternating Between Faces	Alternating transitions between the faces of the two people in the scene
Hands	A period of fixation on the hand; may include those not captured by levels below; aggregates levels below
Illustrative Gestures	A period of fixation on gestures described as "illustrators" in the previous section
Emblematic Gestures	A period of fixation on gestures with set, lexical-like meanings
Body	Any period of fixation on a part of the body aside from the face or hands
Objects	A period of fixation on objects being interacted with by a character, or which is otherwise related to the discussion (e.g., a head of lettuce in someone's hand during a discussion of what to eat for dinner)
Setting	A period of fixation on other parts the frame which are unrelated to the discussion (e.g., a lamp in the background)

4.3.1.2. Interview data

The screen-capture videos with interview audio were transcribed by a professional transcription company in Tokyo and the time-coded transcripts were imported into NVivo for qualitative analysis. Interview videos were coded for participant, and sections were coded for videotext and task type. 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 videotext. Following the same iterative approach to node/theme creation as above, the list of nodes displayed in Table 4.5 was arrived at and applied to all interviews.

Table 4.5.

Reasons for Oculomotor Events Coded in Qualitative Data Analysis

Name	Description
Determining Affect	The explanation for the event is related to determining the mood of the character in the scene
Supplementing Comprehension	The explanation for the event is related to supporting the linguistic stream for improved comprehension
Cultural Explanation	The explanation for the event is related to a perceived cultural difference between Japanese people and foreigners/Americans
Related to Item	The participant has specifically cited the item in question as a reason for the event
Unrelated to item	(An aggregate category of those below)
Determining Who Will Speak Next	The explanation for the event (usually alternating between faces) is that the participant was waiting for a character to begin speaking, but was unsure of which would begin
Habit	The explanation for the event is personal habit
Unconscious	The explanation for the event is that it was an unconscious behavior
Other	(Any other explanation)

4.3.2. Analysis Methods

Study I data were analyzed quantitatively with the qualitative (interview) data aiding in the interpretation of the quantitative (eye-tracking) data. An overview of the comparative events and dwell times for the total sample, the participants, the videotexts/items, and the task types is achieved through descriptive statistics. To determine whether viewing behavior is predicted by overall proficiency, non-parametric Spearman's ρ correlation coefficients (owing to the small sample size) are employed to investigate the relationship between TOEFL scores and total dwell time on the visual cues. Genders have often been found to attend to nonverbal cues differently (see Burgoon, 1994; Burgoon et al., 2016; Costanzo & Archer, 1989; Noller, 1985; Rosenthal et al., 1979); however, due to the very small sample of male participants (n = 2), any investigation of gender-based viewing differences would be inconclusive and, as such, will be omitted here. Similarly, although a multiple analysis of variance (MANOVA) was considered as a means of demonstrating a difference between the task types' vectors of mean dwell times, the current sample size is

insufficient, requiring a minimum of ten values per cell (Tabachnick & Fidell, 2007). Differences are instead sought via paired *t*-tests of each event of interest, similar to the methodology of Suvorov (2013, 2015). All quantitative analyses were completed in SPSS Statistics 22. Qualitative reports from participants on the reasons for their viewing behaviors are incorporated into the reports of the results.

4.4. Results

4.4.1. Overview

Overall descriptive statistics, irrespective of videotext or task type, for the manually coded oculomotor events are displayed in Table 4.6. Events coded can be seen in the upper lines, and dwell times (percentage of videotext length coded) can be seen in the lower lines. Table 4.7 breaks out the percentages for the face events only; values are not percentages of the total dwell time, but of the face dwell time. Reasons given for the events in interview can be seen in Table 4.8. The reason, "Related to Item" has been removed, as it necessarily overlapped with other reasons given by the respondent. It will be addressed separately in later sections. Cell values represent the row percentages.

The visual cue accounting for the largest amount of dwell time was, not surprisingly, the face, with an average of 81.23%. The most common reason given for any facial watching was "Determining Affect," (38%) i.e., determining the person's mood by facial expression, which aligns with nonverbal communication theory. The speaker's face comprised most of the total face dwell times (75.01%) and 60.76% of the dwell time overall. Time spent oriented toward faces was comprised almost equally by the speaker's eyes (24.92%), mouth (25.46%), and scanning the speaker's face (22.56% of facial dwell time); these represented percentages of the total dwell

time of 20.19%, 20.41%, and 18.56% respectively. Reasons given for viewing the eyes or scanning the face were, predictably, largely for determining affect, but this

Table 4.6.

Overall Descriptive Statistics for Oculomotor Events and Dwell Time Percentages

	Min.	**Mdn.**	***IQR***	**Max.**	**Mean**	***SD***	**Skew**	**Kurt.**
	Event	Event	Event	Event	Event	Event	Event	Event
Visual Cue	Dwell	Dwell	Dwell	Dwell	Dwell	Dwell	Dwell	Dwell
Face	3	14	9.75	29	14.75	6.02	0.264	−0.675
	53.37	81.74	12.51	95.86	81.23	8.81	−0.637	0.486
Speaker's Face	4	11.5	8	23	11.90	4.77	0.276	−0.986
	32.64	60.98	16.67	82.89	60.76	11.62	−0.286	−0.478
Speaker's Eyes	0	4	7.75	15	4.93	4.50	0.517	−1.042
	0	18.23	31.51	68.11	20.19	19.63	0.677	−0.543
Speaker's Mouth	0	4	6	16	4.46	4.16	0.767	−0.301
	0	16.97	33.02	69.21	20.41	19.69	0.755	−0.345
Speaker's Face Scan	0	3	3	10	3.25	2.34	0.841	0.276
	0	19.85	18.57	67.33	18.56	13.20	0.743	1.088
Listener's Face	0	3	3	8	2.85	2.03	0.371	−0.675
	0	5.00	7.46	36.79	6.16	5.65	2.309	10.816
Listener's Eyes	0	1	2	6	1.29	1.61	1.350	1.072
	0	0.70	4.77	13.58	2.57	3.49	1.358	0.944
Listener's Mouth	0	0	1	4	0.39	0.74	2.414	7.318
	0	0	0.89	6.42	0.73	1.48	2.333	5.107
Listener's Face Scan	0	0	2	5	0.81	1.16	1.399	1.576
	0	0	3.41	24.34	2.29	3.98	2.898	12.342
Alternating Between Faces	0	3	3	11	3.22	2.58	1.008	1.008
	0	9.73	18.67	46.60	12.79	10.96	0.640	−0.190
Hands	0	1	3	8	1.65	1.79	1.070	0.927
	0	1.00	2.84	9.56	1.64	2.00	1.625	3.114
Illustrative Gestures	0	0	1	4	0.76	1.07	1.276	0.580
	0	0	1.02	6.89	0.79	1.35	2.292	6.013
Emblematic Gestures	0	0	0	1	0.01	0.12	8.485	72.000
	0	0	0	1.02	0.01	0.12	8.485	72.000
Body	0	0	1.75	13	1.31	2.32	2.715	9.089
	0	0	2.58	13.52	1.86	3.10	1.807	2.556
Objects	0	3.50	6	13	3.69	3.44	0.661	−0.235
	0	4.53	11.39	30.11	6.11	6.44	1.110	1.479
Setting	0	1	2	8	1.32	1.40	1.958	6.394
	0	1.15	4.60	13.58	2.97	3.85	1.279	0.292

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 6 videotexts (N = 72).

Table 4.7.

Breakdown Descriptive Statistics for Facial Dwell Times

Visual Cue	Min.	Mdn.	*IQR*	Max.	Mean	*SD*	Skew.	Kurt.
Speaker's Face	39.05	78.71	16.54	100.00	75.01	12.72	-0.754	0.192
Speaker's Eyes	0	20.61	39.19	90.11	24.92	24.16	0.664	-0.513
Speaker's Mouth	0	19.89	40.53	83.49	25.46	24.66	0.713	-0.598
Speaker's Face Scan	0	24.55	22.81	70.79	22.56	15.40	0.482	-0.068
Listener's Face	0	6.84	9.30	44.42	7.60	6.80	2.236	10.730
Listener's Eyes	0	0.85	5.56	15.97	3.13	4.23	1.330	0.745
Listener's Mouth	0	0.00	0.76	9.19	0.93	1.91	2.419	5.925
Listener's Face Scan	0	0.00	4.76	29.38	2.82	4.84	2.848	11.935
Alternating Between Faces	0	12.39	21.21	52.78	15.50	13.19	0.590	-0.491

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

Table 4.8.

Reasons Provided for Oculomotor Events

	Reason Given						
Visual Cue	Determining Affect	Supplementing Comprehension	Cultural Explanation	Determining Who Will Speak Next	Habit	Unconscious	Other
Face	38%	25%	1%	8%	7%	5%	16%
Speaker's Face	41%	26%	2%	-	11%	7%	12%
Speaker's Face Regions	37%	33%	2%	-	12%	6%	9%
Speaker's Eyes	54%	23%	3%	-	16%	1%	4%
Speaker's Mouth	14%	47%	2%	-	7%	13%	18%
Speaker's Face Scan	65%	12%	-	-	-	18%	6%
Listener's Face	53%	16%	-	5%	-	9%	17%
Listener's Face Regions	39%	40%	-	-	-	-	21%
Listener's Eyes	43%	44%	-	-	-	-	12%
Listener's Mouth	-	-	-	-	-	-	100%
Listener's Face Scan	100%	-	-	-	-	-	-
Alternating Between Faces	26%	28%	-	27%	-	-	19%
Hands	4%	25%	10%	-	13%	37%	12%
Illustrative Gestures	5%	32%	13%	-	17%	22%	11%
Emblematic Gestures	-	-	-	-	-	100%	-
Body	-	-	-	-	-	55%	45%
Objects	-	6%	-	-	4%	20%	71%
Setting	-	24%	-	-	-	6%	70%

was the reason given 65% of the time for scanning the speaker's face. Notably, almost half (47%) of the reasons for watching the speaker's mouth was to supplement comprehension. In all such cases, respondents (n = 6) reported that it was easier to understand what the character was saying by watching the mouth.

The listener's face drew gaze for considerably less time than the speaker's face. It is for this reason, most likely, that both skewness and kurtosis for the means associated with these visual cues were all very large. For that reason, it is more appropriate to refer to the non-parametric statistics of median and interquartile range (*IQR*) for the purposes of description. The median dwell time on all cues associated with the listener's face was 5%; it represented 6.84% of facial dwell times. In interview, 53% of the reasons given for watching the listener's face were related to affect. For listener face scanning behavior, it was 100% of the reasons given.

The second most common facial visual cue was "Alternating Between Faces" (12.79% of the total, an average of 15.50% of the facial orientation). The reasons for this are fairly evenly split among "Determining Affect," "Supplementing Comprehension," and "Determining Who Will Speak Next." The last of these is of little theoretical importance, as it typically occurred in moments of silence, and was likely unconscious; it also tended to occur at the beginning of scenes before either character began speaking. This behavior has been long observed in the literature with still photos, and typically occurs as the viewer attempts to discern the overall gist of the scene presented (Yarbus, 1967 as cited in Duchowski, 2007).

All hand events only accounted for a median of 1.00% of the total dwell time, lower, even, than either body or setting events, despite the theoretical importance of gestures. The most oft-given reason for looking at a character's hands was that it was unconscious (37%). 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," resulting in the "Unconscious" coding of those responses.

Illustrative gestures accounted for a mean of 0.79% of the looking (although the large skew indicates that the median would be a more appropriate statistic of centrality, but the median is 0), and the most oft-given reason for this was to supplement comprehension. Two respondents gave explicitly cultural explanations for their behavior. Respondent 4, who has spent some time studying in the United States said:

> なんか、イメージですけど、外国人のその独特の動きとかあるじゃないですか。そういうのも関係してるかもしれない、なんか、これでも伝わるっていうこと、ちょっと分かってるから。
>
> *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].*

Emblematic gestures had only one event in the entire administration of the test over the twelve examinees and six videos. This is likely, however, partly a function of the fact that there was but one emblematic gesture in the entire test. The single respondent 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.

Body and setting visual cues were looked at only an average 1.86% (although this distribution was heavily skewed) and a median of 1.15% of the time, respectively. 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 24% were related to supplementing comprehension, this was only one respondent, who spoke of it twice during the interview.

Objects were looked at at roughly the same percentage of time as listener's faces. Although Table 4.6 shows that 71% of the mentions of object viewing behavior were simply classified as "other," this obscures the relatively large number of times that the reason given was directly related to the question asked. It is worth noting that the average number of events is fairly high in comparison to other non-facial visual cues. The disparity between number of events and percentage of time indicates that objects, when looked at, are not watched for long.

Distribution statistics for the events and dwell times indicate some extremely right-skewed distributions, indicating a high degree of variability in individuals' interactions with the videotexts. This is unsurprising due to the large number of variables and the relatively small sample size, as smaller sample sizes are inherently more variable.

4.4.2. Individual Differences

The participants' gaze behaviors were compared via descriptive statistics for the individuals and non-parametric correlations between TOEFL scores and dwell times.

4.4.2.1. Comparisons of individuals

Recognizing that individual differences may play a large part in the observed patterns of gaze behavior, descriptive statistics were also prepared for the participants, which can be found in Table 4.9. The clearest difference between the participants' viewing behavior is that individuals appear to prefer watching either speakers' eyes or mouths, with only Participants 5 and 11 seeming to split time between them.

Ideally, a repeated-measures ANOVA of the dwell times for each of the visual cues of interest for each of the participants could be used to determine if these differences were statistically significant, but several of the participants' behaviors had extreme outliers and/or departed from normality according to Shapiro-Wilk tests for the grouped data, violating key assumptions of the repeated-measures ANOVA. Once again, this is most likely due to the small sample size for each of these behaviors.

4.4.2.2. Proficiency

An interaction between participant English proficiency and viewing behavior was also sought via Spearman's ρ correlation coefficients between TOEFL scores and total dwell time percentages (See Table 4.10). Total dwell percentages represent the percentage of time spent oriented toward the visual cue in question across all six of the videos, calculated by dividing the sum of times in orientation by the total sum of the video lengths. Although both TOEFL scores and dwell times are continuous variables, and most met the more common Pearson product-moment correlation assumptions of Normality, linearity, homoscedasticity, and absence of outliers, the small sample size ($N = 12$) could impede interpretation of these statistics (Tabachnick & Fidell, 2007). For this reason, the non-parametric Spearman's ρ, useful in instances wherein the Normality assumption is violated (Bachman, 2004), was employed. This revealed only four significant correlations. The first was a negative correlation between TOEFL score and total percent dwell time on the listener's eyes ($\rho = -0.61$, $p = 0.036$). The next was a positive correlation between TOEFL score and time spent watching the listener's mouth ($\rho = 0.61$, $p = 0.034$). A significant correlation was also observed between TOEFL score and total dwell time oriented toward the body

Table 4.9.

Descriptive Statistics for Dwell Time Percentages by Participant

	Participant											
	1 (F)	2 (F)	3 (F)	4 (M)	5 (F)	6 (F)	7 (F)	8 (F)	9 (F)	10 (F)	11 (F)	12 (M)
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Visual Cue	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*	*SD*
Speaker's Eyes	36.7	3.6	35.9	10.2	15.4	32.2	6.8	38.8	8.4	30.4	22.3	1.7
	15.3	*8.9*	*21.1*	*12.2*	*11.6*	*17.3*	*12.1*	*23.0*	*9.0*	*6.2*	*26.8*	*2.2*
Speaker's Mouth	9.8	32.6	7.1	34.1	20.4	0.8	33.6	4.6	21.4	10.0	25.1	45.4
	15.6	*20.1*	*9.3*	*21.9*	*11.3*	*1.3*	*18.1*	*9.9*	*13.3*	*9.9*	*20.6*	*20.1*
Speaker's Face Scan	14.3	22.2	14.3	16.2	19.1	27.2	15.6	20.1	30.0	18.5	9.3	15.8
	9.6	*13.7*	*12.3*	*10.2*	*7.1*	*22.4*	*13.5*	*17.3*	*7.3*	*15.3*	*9.5*	*9.4*
Listener's Face	7.1	6.1	10.2	6.7	6.1	6.8	4.2	4.2	5.3	4.6	8.6	4.1
	5.3	*3.6*	*13.4*	*4.3*	*4.4*	*7.0*	*4.4*	*4.1*	*4.4*	*4.0*	*4.3*	*3.0*
Alternating Between Faces	13.2	12.7	6.0	7.1	14.0	17.2	9.5	15.2	13.0	14.9	16.9	13.8
	6.8	*9.8*	*5.3*	*9.7*	*12.1*	*17.1*	*10.2*	*13.9*	*11.5*	*8.4*	*14.3*	*11.4*
Illustrative Gestures	0.4	1.3	0.5	1.3	1.2	0.3	2.3	0.1	0.6	0.7	0.1	0.6
	0.6	*1.3*	*0.8*	*1.5*	*2.8*	*0.8*	*2.0*	*0.2*	*0.7*	*0.9*	*0.3*	*1.5*
Objects	4.8	7.6	6.7	7.8	4.8	5.3	10.0	6.5	4.6	6.6	4.7	4.0
	5.1	*5.5*	*7.1*	*9.3*	*5.8*	*5.8*	*10.8*	*5.6*	*6.4*	*7.4*	*5.9*	*3.4*

($\rho = 0.85$, $p = 0.000$). Finally, a significant correlation was observed between overall proficiency as measured by the TOEFL ITP and dwell time on objects ($\rho = 0.64$, $p = 0.024$).

Table 4.10.

*Spearman's ρ Correlations between TOEFL Score and Dwell Time Percentages (*N *= 12)*

	TOEFL Score	
Visual Cue	Spearman's *ρ*	*p* (2-tailed)
Face	−0.57	0.051
Speaker's Face	−0.27	0.391
Speaker's Eyes	−0.23	0.471
Speaker's Mouth	0.36	0.255
Speaker's Face Scan	−0.12	0.713
Listener's Face	−0.01	0.983
Listener's Eyes	−0.61*	0.036
Listener's Mouth	0.61*	0.034
Listener's Face Scan	0.32	0.313
Alternating Between Faces	−0.44	0.151
Illustrative Gestures	0.36	0.255
Body	0.85*	0.000
Objects	0.64*	0.024
Setting	−0.11	0.745

* Significant at the 0.05 level.

4.4.3. Videotexts

Differences in dwell time for the visual cues of interest were examined by videotext via descriptive statistics and one-way ANOVA. Although the behaviors were the product of the interaction between both the video content and the item stem, these will be referred to simply as "videotexts" for clarity. It is important, however, to bear in mind that the item itself was an important motivating factor for the viewing behaviors.

Table 4.11 displays descriptive statistics for the dwell times by videotext. These data are also visualized in Figure 4.2. Throughout, standard deviations are quite high, usually equaling or surpassing the mean, which further demonstrates the impact of participants' own peculiarities on these data, as well as the limits of a small sample size. However, some patterns do emerge.

Table 4.11.

Descriptive Statistics for Dwell Times by Videotext

Visual Cue	Videotext					
	Shopping	Ted and Mary	Football	Dinner	Salesgirl	Cheryl's Call
	Mean *SD*	Mean *SD*	Mean *SD*	Mean *SD*	Mean *SD*	Mean *SD*
Speaker's Eyes	26.87 *25.86*	12.69 *17.03*	25.79 *23.34*	22.92 *16.15*	21.83 *17.21*	11.07 *13.43*
Speaker's Mouth	19.11 *21.04*	34.59 *19.83*	18.57 *21.47*	20.56 *20.31*	17.76 *19.00*	11.86 *10.92*
Speaker's Face Scan	18.10 *13.55*	19.46 *17.66*	11.10 *10.57*	29.31 *7.95*	9.62 *8.13*	23.77 *9.10*
Listener's Face Regions	4.64 *3.99*	2.22 *1.53*	2.19 *3.24*	5.74 *3.79*	2.09 *2.17*	4.07 *3.97*
Listener's Face Scan	2.68 *3.29*	0.06 *0.19*	0.22 *0.76*	4.87 *3.14*	0.96 *1.88*	4.97 *7.07*
Alternating Between Faces	0.40 *0.82*	16.96 *6.13*	19.45 *8.66*	4.51 *3.08*	9.34 *6.54*	26.10 *9.70*
Illustrative Gestures	0.45 *0.80*	2.02 *1.95*	0.02 *0.08*	0.61 *0.89*	1.65 *1.61*	0.00 *0.00*
Objects	14.39 *6.71*	0.37 *1.04*	10.68 *3.32*	2.21 *1.84*	8.82 *3.29*	0.20 *0.71*

NOTE: Dwell times listed as percentages of their respective videotexts. (*N* = 12).

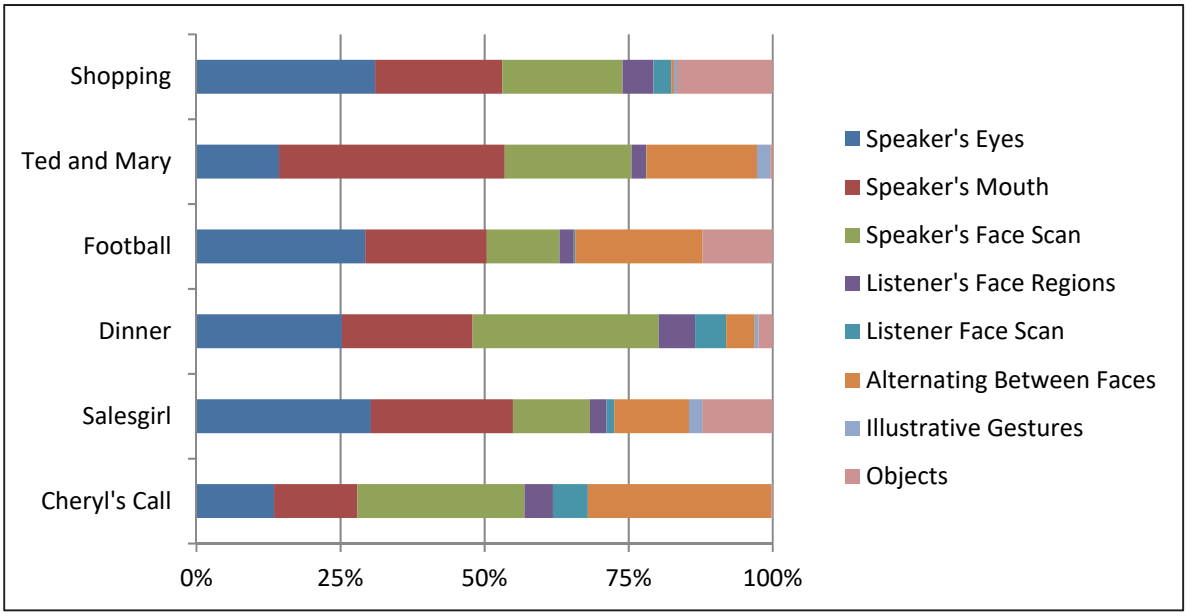

Figure 4.2. Stacked bar chart illustrating the comparative dwell times for each visual cue on each videotext/item.

The highest dwell times for “Speaker’s Face Scan,” “Listener’s Face Regions,” and “Listener’s Face Scan” were observed on the videos “Dinner” and “Cheryl’s Call.” Both of these videotexts have strong emotional components and items referring to the characters’ attitudes or unstated thoughts. Long dwell times were also observed for “Alternating Between Faces” on “Ted and Mary,” “Football,” and “Cheryl’s Call.” In the case of “Ted and Mary,” most of this appears to be of little comprehension importance, as the most common coding for much of it was “Determining Who Will Speak Next,” as the scene opens on the couple sitting in chairs, silently, for several seconds. “Cheryl’s Call” has already been discussed, and “Football” also featured a question about a character’s feelings.

Gaze behavior corresponding to illustrative gestures seemed to appear mostly in “Ted and Mary” and “Salesgirl.” These two scenes have the most full-body camera angles, allowing the viewer to see the natural movements of the actors’ hands to a greater degree than in the other scenes. This is the likely reason for these results.

Objects were viewed the most in the “Shopping,” “Football,” and “Salesgirl” videos. In the case of the “Shopping” and “Salesgirl” videos, the item referred directly to objects. The item for the former was:

男性は何を買いましたか。

What did the man buy?

The male character enters the house carrying shopping bags and he and his wife discuss his shopping trip. During this time, most examinees looked at everything he was holding and everything he was wearing. The “Salesgirl” item asks:

男性は何を探していますか。

What is the man shopping [searching] for?

In interviews, these two items resulted in the majority of responses coded "Related to item," reserved for comments wherein the respondent specifically referenced the item as a reason for their gaze behavior. In the case of the "Shopping" video, it is explained early on that the man has bought a jacket, and most (nine of the twelve participants) answered the item correctly, but examinees continued to look at his belongings throughout. The "Salesgirl" item was the easiest on the test, with all respondents answering correctly. The man repeatedly says that he is not looking for anything, but examinees continued to look at many items in the store.

As with the participants, although a repeated-measures ANOVA of the visual cue dwell times for each of the videotexts would reveal whether the observed differences were significant, the assumptions of that statistical procedure were not met by these data.

4.4.4. Task Type

The final set of analyses is of the dwell times associated with the two task types investigated in Study I: explicit and implicit items. Descriptive statistics are available in Table 4.12 and Figure 4.3 graphically presents the comparative dwell times for the task types.

4.4.4.1. Descriptive statistics

Descriptive statistics continue to reveal the trends discussed in the videotext section above. Generally speaking, implicit items, whose answers are not explicitly stated in the verbal channel, and which frequently address motives or attitudes of the characters in the scenes appear to be more strongly associated with facial visual cues, the seat of affect displays. The effect is most pronounced in the "Listener's Face Scan" and "Alternating Between Faces" categories, which both involve checking the facial reaction of the character who is not currently speaking.

Table 4.12.

Descriptive Statistics for Dwell Times for Explicit and Implicit Task Types

	Explicit	**Implicit**
	Mean	Mean
Visual Cue	*SD*	*SD*
Speaker's Eyes	22.12%	20.19%
	16.29%	*14.59%*
Speaker's Mouth	21.64%	17.74%
	15.26%	*14.87%*
Speaker's Face Scan	15.10%	24.12%
	7.88%	*5.17%*
Listener's Face Regions	3.18%	4.57%
	2.10%	*2.77%*
Listener's Face Scan	1.50%	3.97%
	1.41%	*2.64%*
Alternating Between Faces	7.12%	13.52%
	3.07%	*4.27%*
Illustrative Gestures	1.22%	0.32%
	0.98%	*0.47%*
Objects	9.47%	3.35%
	3.36%	*1.30%*

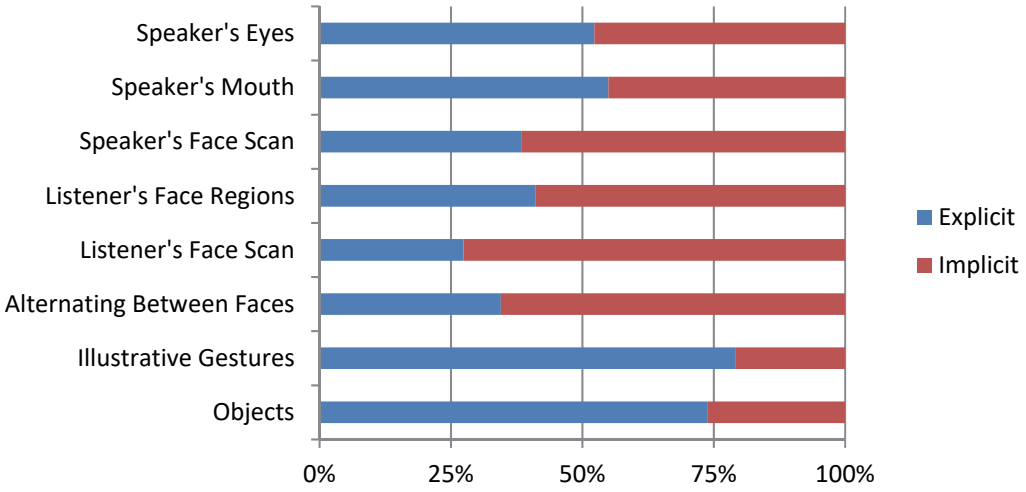

Figure 4.3. Stacked bar chart illustrating the proportion of dwell time for the visual cues of interest for the task types.

Respondent explanations for the "Alternating…" behavior typically involve determining affect. This comment from Participant 6 about the "Cheryl's Call" scene, is a good example:

> 二人はどんな気持ち、こっちは、「何で入ってくんのよ」みたいな感じで、こっちは、「何で俺、入れないんだよ」みたな、ちょっとなんか。
>
> What are these two people feeling? It's like this one [the wife] is all, "Why are you butting in [to the conversation]???" and he's all, "Why won't you let me in???"

It is typically during moments when knowledge of both parties' moods or attitudes would be advantageous to comprehension or to answering the item in question that this behavior arose.

"Illustrative Gestures" and "Objects" both occupied more dwell time in explicit items than in implicit. These results may be spurious, however, for the videotext-based reasons discussed above. They may be more of an artifact of the scenes used than the items themselves. There simply were more gestures in two of the three explicit items than in the implicit items, and two of the explicit items had some relationship to inanimate objects.

In interview, after every behavior explanation the participants gave, they were asked whether an observed behavior was related to the item they were trying to answer at the time. Building a matrix of the viewing behavior events and task type further demonstrates the patterns described above (see Table 4.13). Eighteen instances of looking at the face when watching videotexts for implicit items were reported to be related to the items in question, as opposed to only two for explicit items. "Alternating Between Faces" events were explained as being related to implicit items eight times, as were seven events of orienting toward the speaker's face. Although dwell times associated with the listener's face were low overall, in three instances participants

explained their orientation toward the listener's face, no matter how brief, as being related to an implicit item; no such responses were collected for explicit items.

Table 4.13.

Contingency Table of Occulomotor Events Reported to be Related to Item by Task

Visual Cue	Explicit	Implicit
Face	2	18
Speaker's Face	0	7
Listener's Face	0	3
Alternating Between Faces	0	8
Hands	1	0
Emblematic Gestures	0	0
Illustrative Gestures	0	0
Body	0	0
Objects	32	7
Setting	0	0

NOTE: Counts are of comments; some instances may have generated more than one comment.

No orientations toward specific gestures were identified as being related to the items in question, although one comment was made on a general hand code that was attributed to an explicit item. It occurred in the "Shopping" videotext, whose question asked, "What did the man buy?" The participant looked at the man's hand in order to determine what he was holding, as it was unclear. He was holding a jacket or suit on a hanger over his shoulder, obscuring it from view. Thirty-two comments on object viewing attributed the behavior to explicit items, as opposed to seven to implicit; the overwhelming majority of these occurred in the "Shopping" and "Salesgirl" videotexts (both explicit items).

4.4.4.2. Paired t-tests

To investigate whether the differences observed in the descriptive statistics discussed above were statistically significant, paired *t*-tests were computed comparing examinees' viewing behavior on explicit and implicit items. As the paired *t*-test assumes a nearly-Normal distribution of differences, Shapiro-Wilk tests were run on

each of the differences of interest and Q-Q plots were examined. No problems were found, so the test was completed. Results and effect sizes are displayed in Table 4.14.

Table 4.14.

Paired t-*Test Between Visual Cue Dwell Times for Explicit and Implicit Task Types*

	Paired Differences						*p*	Cohen's
Visual Cue	Mean	*SD*	*SE*	95% *CI*	***t***	**df**	**(2-tailed)**	**|*d*|**
Speaker's Eyes	1.935	13.033	3.762	[−6.346, 10.216]	0.514	11	0.617	0.24
Speaker's Mouth	3.908	9.845	2.842	[−2.347, 10.163]	1.375	11	0.196	0.64
Speaker's Face Scan	−9.028	7.863	2.270	[−14.024, −4.032]	−3.977	11	0.002	1.17
Listener's Face Regions	−1.386	3.403	0.982	[−3.548, 0.776]	−1.411	11	0.186	0.29
Listener's Face Scan	−2.471	2.427	0.701	[−4.013, −0.929]	−3.526	11	0.005	0.79
Alternating Between Faces	−6.402	4.271	1.233	[−9.116, −3.688]	−5.193	11	0.000	1.33
Illustrative Gestures	0.901	0.649	0.187	[0.489, 1.313]	4.810	11	0.001	0.65
Objects	6.124	3.289	0.949	[4.034, 8.213]	6.450	11	0.000	1.15

Significant task-based differences in “Speaker’s Face Scan,” “Listener’s Face Scan,” “Alternating Between Faces,” “Illustrative Gestures,” and “Objects” were all observed. Cohen’s *d* measures of effect size were calculated with correction for the correlated variables inherent to paired *t*-tests (Cohen, 1988). Cohen *d* statistics are considered “large” at any value above |0.8|; three of the significant differences surpass this by a wide margin (up to 1.33), indicating that the differences in examinee watching behavior between the two formats is extremely large.

4.5. Discussion

This section discusses the results of Study I. It addresses the results in subsections dedicated to participants, the videotexts, and finally task types.

4.5.1. Participants' Interaction with Visual Cues

As reported above, it appears that examinees interacting with a video-mediated test of foreign language listening comprehension tend to focus mostly on the face of whomever is speaking, with only small departures from this to look at gestures, objects, the setting, and so on, and that this was fairly true of the entire sample. Although this finding is commonsensical, a surprising caveat was that individual participants tend to favor either eyes or mouths when watching someone speak, at least in the context of Study I. The reasons given by respondents in interview for the former almost all referred to the expression of the speaker; in fact, the word *hyōjō* (表情; "expression" or "countenance") appeared twenty times among nine of the participants when asked why they were looking at the speaker's eyes. This seems to agree with the findings of Coniam (2001), whose respondents reported that facial expressions were useful for determining the speakers' attitudes and predicting what they were likely to do. The reasons given for watching the mouth mostly centered on an increased facility for comprehension.

Suvorov (2013) noted the importance of lip reading to his respondents as well, and Ockey (2007) points out that the interaction between watching lip movements and perceiving sound is hardly limited to speakers of a second or foreign language. McGurk and MacDonald (1976) famously discovered the so-called "McGurk" effect, wherein a video of a speaker pronouncing the syllable /ga/, but with a soundtrack of the speaker saying /ba/, will result in people reporting that they heard /da/, whereas the same soundtrack played without video will result in them reporting the correct sound, /ba/. Perhaps future research can investigate the specific impact of lip reading on video-mediated listening tests, perhaps by blurring out the lips of speakers.

A surprising finding was how little gestures seemed to draw examinees' eyes during viewing of the videotexts, given their prominence in the nonverbal communication literature, and their importance to previous studies investigating examinee interaction with video-mediated listening tests. When asked about their looking at a gesture, only two of the respondents 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 respondents mentioned the gestures appearing in his videos repeatedly. However, it is important to note that the gestures that were most impactful were those made to facilitate an explanation of famous American Western lawman Wild Bill Hickock's distinctive and unintuitive gun draw. Furthermore, Wagner interpreted his respondents' use of the same gestures to explain the content back to him as an indication of their importance to comprehension, rather than the perhaps-more-likely interpretation that the content was difficult to describe in words, and that his respondents also lacked sufficient vocabulary to do so. The gestures that seemed to affect Cubilo and Winke's (2013) respondents the most were topic organizing gestures in academic lectures, which do not appear in Study I.

Only one respondent viewed the single emblematic gesture in the six videotexts, and she did not remember doing so. This perhaps illustrates the difficulty of incorporating these 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 orient toward them. It is, of course, possible that they unconsciously glanced at that and other gestures in saccades too fast for the eye-tracking hardware to register, or that they were detected via parafoveal vision. The latter possible explanation would conform to the "spotlight" model of attention, wherein only that

which is consciously chosen for attention is oriented towards in foveal vision, with detection of any other signal being context-sensitive (Posner, Snyder, & Davidson, 1980).

However, even when respondents were specifically questioned whether they had noticed a gesture, those who had not looked, or looked too quickly for the eye tracker to register, 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.

Although the sample size was small, the two largest significant correlations observed between TOEFL ITP scores and visual cues viewed indicated that those with higher proficiency spent more time overall oriented toward the bodies of the characters in the scenes and objects appearing alongside them. Perhaps this is an indication that close attention did not need to be paid to the speaker in order to comprehend what he or she was saying, freeing the participant to look around the scene as they wished, having already determined the answer, or simply able to listen and look elsewhere with comparative ease. This interpretation is somewhat supported by several of the answers given by Participant 3, who had a TOEFL score of 530. She tended to look around the scenes at anything she found interesting, for example, in the "Shopping" videotext:

Q: あの、このさっき、あのー、オーブンを見ましたね？

Um, just before, um, did you look at the oven?

A: はい。

Yes.

Q: それ、何か理由がありますか？

Was there some reason for that?

A: いや、うーん、これは、どれ、いつ頃の物なのかなと思って、あんまり見たことがなかったので。

No, uhhh, that... what... when is that thing from, I wondered. I haven't really seen [an oven like that].

Q: あー、オッケー。

Ah, okay.

A: 多分、関係ないですね、質問には。

That is probably unrelated to the [test] question.

She continued this behavior. The following interchange took place only approximately thirty seconds later, after a brief conversation about her looking at the character Cheryl's necklace, wondering what it was:

Q: 髪型も見てます。

You're looking at [her] hairstyle.

A: あー、確かに。

Ah... Clearly.

Q: そ、そんなに聞いてない？[...]

Are you maybe not listening that closely?

A: なんか、大体もうここで分かって、 [Explains the content of the scene.] はい、それでもう、あとは自分の興味あるの、見てたのかな。

Um, basically, I already understand it. [Explains the content of the scene.] *Yeah, because of that, afterward I just looked at whatever interested me.*

Q: なるほど。もう、あ、頭の中にもう答え、決まったから。

I see. Because you've already settled on the answer in your head.

A: 決まった、はい。

Yeah, I've already decided.

Q: 楽に何でも見てて。

You're just comfortably looking at anything.

A: はい。

Yeah.

This pattern of viewing behavior continued throughout the test, with the participant offering prosaic, "I just thought it was interesting" answers when asked about her scanpaths.

Participant 4, who had a TOEFL ITP score of 533 gave a similar explanation for an observed pattern of behavior during the first videotext, "Shopping":

Q: すごくはげてる頭、見てますね。

You're really looking at his bald head, aren't you?

A: そうですね。こんとき、でも、面白くて、ちょっと笑いながら、いろいろ見てたんだと思います、こう、ちょっと余裕が出たみたいな。

Yeah. At this time, it was funny, and I was kind of laughing and looking at all sorts of things, I think. I had already gotten some freedom [to do that].

These answers may indeed provide an explanation for the correlations observed in the proficiency data analyses.

Finally, although males were found to watch mouths more than women, the sample size was simply too small to generalize from. Based on these data, there is not sufficient evidence to suggest a real gender-based difference in viewing behavior.

4.5.2. Videotexts' Interaction with Visual Cues

Perhaps the most important finding of Study I is the large impact videotext has on viewing behavior. Although it was not possible to determine whether the observed differences in behavior between the videotexts were statistically significant, it was clear from the descriptive statistics that examinees interacted differently with them. This finding, however, must be treated with caution. The design of the study was such that it is impossible to extract the predictive effect of the item from the content of the video. Furthermore, as Suvorov (2015) also points out, the editing, camera angles, focal depth and so on all have a profound effect on how viewers view video. In fact, that is a major point of cinematography: to control and direct the audience's attention. This was most apparent in Study I with the "Football" item, which included several shots where the television displaying the football game was placed in the frame in such a way as to remind the viewer that the husband was watching the game. Given this observation, it becomes unclear whether the large number of oculomotor events associated with "Objects"—the television—in that particular videotext can be attributed to the item (although understanding that the man was watching the game was important to determining the answer to the question), or simply to the craft of cinematography.

The impact of cinematography may be playing an additional role here, with respect to the surprisingly low number of gesture events. As discussed briefly in the results section, only two of the videos, "Tom and Mary" and "Salesgirl," had enough wide, distant camera shots to capture the hands sufficiently to display enough gestures to make an impact on counts. This will have implications for test design in Study II.

4.5.3. Task Types' Interaction with Visual Cues

Significant differences between examinees' viewing patterns on implicit and explicit items were discovered with respect to scanning the speaker's or listener's face, alternating quickly between the speaker's and listener's faces, illustrative gestures, and objects. The caution regarding the impact of cinematography above notwithstanding, this is an important finding. The increase in dwell times for visual cues associated with affect for items pertaining to attitude, unstated opinions, or emotional state confirms Ockey's suspicions that video would have more of an impact with items of that nature (2007). Such items are a common fixture of tests such as the TOEFL (Alderson, 2009; Bejar et al., 2000; Nissan et al., 1996), and if video were to be integrated into such tests, they may be more influenced by the video content than is desired. If video has a facilitative effect on such items, it is important to detect and describe it sooner rather than later.

Another interesting finding of Study I was the significant increase in dwell times associated with objects in explicit items. Although two of the questions make mention of objects, in only one of them ("Shopping," wherein the man is holding his new purchases) do characters interact with any of them. Simply the suggestion of them may have been enough to change behavior. Respondent 9 had this to say when asked why she was looking at objects in the "Salesgirl" video:

「まだ何か探してるんだろうな」っていうのがあったからかな。

I guess I was thinking, "Maybe he's still looking for something."

This behavior was in spite of the fact that the male character states repeatedly and pointedly that he is not looking for anything, and the fact that every examinee answered this item correctly. Although references to objects may not be necessarily limited to explicit items, it seems reasonable to believe that implicit items would be

less likely to do so. Furthermore, references to objects in the stimulus have been found to be a poor predictor of item difficulty (Nissan et al., 1996); however, it seems possible that item response may change for such items if delivered via video.

4.5.4. Implications of Study I for Study II

The first major implication of Study I for Study II is related to the finding that the visual cues with the highest dwell times were the seat of affect display nonverbal cues: the face. In previous comparative video listening test work, it was unknown what information was missing when an item was administered in an audio-only format, and what, therefore, was the cause of any effect observed. By establishing that viewers spend most of their time focusing on faces, especially the speaker's face, Study II can be designed to compare a video-mediated version of a test to its audio-only counterpart with the knowledge that the major difference between the formats is likely to be the presence or absence of facial affect displays.

The second major theoretical implication of Study I is a hypothesized difference between the task types, with examinees encountering the implicit items in the video-mediated format benefitting from a facilitative effect of the facial affect nonverbal cues. This effect should be more pronounced in items related to the emotional or attitudinal condition of characters in the video.

Although gestures did not draw much conscious attention from the participants of Study I, there are two implications for Study II test design regarding gestures. For gestures, which figure prominently in the nonverbal communication literature, to play any part in the videotexts, scenes must be framed with wide angles with little to no camera movement or cuts in order to capture gestural use, as cinematography was observed exerting a great deal of influence on participants' viewing behavior in Study I. This will necessitate relatively large filming spaces and a stable tripod. More

significantly, it will make actor line memorization all the more important, since mistakes will require scene retakes. With longer scenes and inexperienced performers, this can be difficult; it will therefore be important to locate reliable performers. The second implication for gestures in the Study II instrument and videotexts is that emblematic gestures do not need to be incorporated unless they occur naturally, as they are unlikely to be looked at by examinees.

4.6. Chapter summary

This chapter has reported on Study I, which sought to quantify and qualify examinee watching behavior in video-mediated listening tests. It began with an explanation of the participants, the equipment, a detailed description of the instrument and its development, and the data collection procedure. It then elucidated the transformation of the collected data for analysis, followed by a presentation of the analysis methods used. Next, it reported the results of Study I, beginning with the overall trends in the data and moving on to individual participant differences. It then explored the differences in examinee viewing behavior between the six videotexts used in the study, and concluded the presentation of results with a description of the influence of task type on watching behavior. Finally, the results of the analyses were discussed, along with implications for instrument development prior to Study II.

The next chapter describes the process of instrument development for Study II, drawing from the observed results of Study I.

CHAPTER 5. STUDY II INSTRUMENT DEVELOPMENT

This chapter details the development of the instrument used in Study II. It begins with a description of the test specifications, then moves to the item design and translation process, and then continues with a description of the video production. The chapter then describes the item moderation stage, including the construction of the item moderation instrument using the Moodle course management system, and reports the results of the item moderation. Finally, the chapter details the development of the main data collection instrument, also with Moodle. The chapter concludes with a brief summary.

5.1. Test Specifications

Prior to developing the instrument, test specifications were drawn up. As previous work has found that examinees tend to watch videos of academic content less than those of conversations (Wagner, 2007, 2010a), and the effect of video on scores associated with academic content is smaller or nonexistent in comparison to conversational material (Batty, 2015; Wagner, 2010b), the decision was made to focus attention only on that text type. As such, the target construct of the instrument was described in the test specification as below (See Appendix B for the full test specification document.):

> The present test is intended to measure informal communicative listening proficiency in English. This can be broadly defined as incorporating the following:
>
> - *Grammatical knowledge* encompassing phonology, lexis, and syntax.
> - *Pragmatic knowledge* required to comprehend the gist of utterances based on an understanding of the illocutionary force of speech acts.

- *Sociolinguistic knowledge* of the meanings of common features of informal conversation such as idiom, figures of speech, and register.

5.2. Item Development

Thirty scripts were written or adapted from situations found in other listening tests, notably those developed by the present researcher for the Kanda English Proficiency Test (KEPT) at Kanda University of International Studies in Chiba, Japan, allowing some content overlap with the 2015 *Language Testing* article by the present author. Other material was adapted from scenarios presented in practice TOEIC (Test of English for International Communication) and TOEIC Bridge items (Exam English Ltd., 2014; Rilcy, 2008; "Short business conversations / TOEIC® listening," 2016), and still others were original works based on the present author's own conversations. All of the scenes featured informal conversational English in both strictly social as well as general business situations.

Each script was written to be performed in thirty seconds, and each featured one man and one woman in a conversation, obviating the need for character names. Each had at least one instance where a character communicated something without explicitly stating it, whether it be simply implied or contained in an idiom. Care was taken to ensure that the language used was as close to authentic as could be expected for scripted dialog, drawing from the present researcher's experience both with writing such material for tests and with co-writing a radio drama series some years prior. Finally, all pertinent nonverbal cues (facial expressions, gestures, posture) were explicitly scripted to ensure that the actors' performances matched exactly the intentions of the researcher, thereby ensuring that the accompanying items functioned as imagined (the successes and failures associated with this approach will be discussed in the following chapter). An example script can be seen in Figure 5.1, and

8. ATTIRE

FADE IN

INT. HALLWAY

MAN and WOMAN meet in hall. Both are wearing conference name badges and carrying papers.

WOMAN

(reading)

Hey, did you see that the invitation for the reception tonight says to dress "casually?" To me, "casually" means jeans and a T-shirt.

(looks at MAN)

What do you think it means here?

MAN

(uneasy)

Yeah, what *does* that mean? Jeans? Shorts? Polo shirt and khakis?

(palms upturned)

What?

WOMAN

Right? So what do we do?

MAN

You know what I'm going to do?

(pointing gesture)

Shirt and a tie.

(palms up)

That way, I can take the tie off if it's too formal.

WOMAN

Well that's fine for you, but what about me?

MAN

Yeah, I guess you're on your own there.

WOMAN

(sarcastically)

Gee, thanks.

FADE OUT

Figure 5.1. Example videotext script.

the full set of scripts can be found in Appendix C. A full list of the scenes and their descriptions can be found in Table 5.1 below.

Table 5.1.

Full List of Scenes, Settings, and Content

Scene	Title	Setting	Summary
1	Licorice	Office	A woman who doesn't like licorice is persuaded to eat some.
2	Coffee or Tea	Hallway	A man is worried that he won't be able to buy tea at a coffee shop.
3	Notes	Classroom	A man coaxes a woman into giving him her notes.
4	Concert	Foyer	A woman convinces a man to call in sick to see a concert.
5	Double Booking	Foyer	A woman has to resolve an accidental double booking.
6	Old Flame	Café	A woman confides that she is falling for a former boyfriend.
7	Smoking	Café	A woman reveals her secret to quitting smoking.
8	Attire	Hallway	A man and woman puzzle over what to wear to a reception.
9	Presentation Consternation	Foyer	A man is nervous about giving a presentation.
10*	She's Late Again	Park Bench	A woman is annoyed that a coworker is always late.
11	Friend Calling	Café	A woman declines a call from an annoying friend.
12*	Allergies	Park Bench	A man blames his sneezing on allergies.
13	Lost Item	Classroom	A man is having trouble describing his lost briefcase.
14	Registration	Classroom	A woman is angry about her registration fees.
15	The Party	Hallway	A man and woman discuss what to bring to a party.
16	Mysterious Key	Foyer	A man discovers a mysterious key in his pocket.
17	Whose Number is This?	Café	A woman finds a suspicious phone number on her husband's phone.
18	Crafts Fair	Hallway	A man chides a woman for pretending to be interested in crafts.
19	Party Aftercare	Foyer	A woman is worried that a party guest is now avoiding her.
20	Book	Foyer	A woman asks a man to return a book she lent him.
21	Int'l Man of Mystery	Foyer	A woman flirts with a well-traveled conference attendee.
22	The Ride Back	Classroom	A man bothers a woman as she prepares to give a talk.
23	The Play	Café	A woman scolds her husband for his poor grasp of time.
24	Job Prospects	Café	A mentor encourages a man to apply for a new position.
25	Travel Costs	Office	A man and woman attempt to mend a company travel budget.
26*	Fax Machine	Office	A man cannot find a fax machine.
27*	Meeting Time	Office	A woman deals with a troublesome client's schedule.
28*	Transferred	Office	A man cannot find the person he is supposed to meet with.
29*	"Right-sizing"	Office	A woman advocates massive layoffs.
30*	Job Interview	Office	A man performs poorly at a job interview.

* Indicates scenes that were not filmed due to inclement weather.

For each scene, at least two multiple-choice items were written (one "explicit" and one "implicit"), and in some cases each scene had three or four corresponding items in the initial pool. The items were written simultaneously in English and Japanese, with only the Japanese to be eventually administered. The English versions were prepared only for the benefit of the item moderators (see following sections) and eventual write-up of the study. A total of 87 item candidates were written for the thirty scenes. Following completion of basic item writing, the Japanese items were checked by a native speaker of Japanese, and compared to the English equivalents, with the present researcher and native speaker collaborating on edits to either or both versions to bring them as closely in line as possible, both in propositional content as well as nuance.

5.3. Video Production

As one of the author's concerns with previous work on this topic is the low quality of the videotexts used, the production of the videos for Study II required a great deal of preparation, practice, and funding. The video production process is detailed below.

5.3.1. Actors

To avoid the acting problems noted in the Batty (2015) study, professional actors were sought to perform the scenes. Many such experimental tests (e.g., the aforementioned KEPT; Suvorov, 2009) overlook the importance of skilled performers in the video listening content, even whilst touting the benefits of the presence of nonverbal cues. However, the inclusion of realistic nonverbal behavior relies on the performers' experience with the craft; acting is not easy work, and its importance should not be underestimated. Trained, experienced actors are adept at quickly learning lines, taking direction, and attempting different versions of scenes.

A talent agency representing foreign actors and models was contacted and the requirements of the project explained. As most non-Japanese actors working in Japan need only look foreign, as they will mostly serve as models in advertisements, etc., and are frequently dubbed over in Japanese, such talent agencies are unaccustomed to meeting demands such as an extensive acting résumé or native-English-speaking ability. Specific requirements for the project were as follows:

- Two men and two women of young adult to middle age, race/ethnicity irrelevant.
- Native speakers of North American, British, or Australian / New Zealand English.
- Post-secondary training in, and an established résumé of, acting, especially in film/television.

Although the company representative assured the present researcher that she could provide a list of such people, the list produced was mostly made up of very young female models and men from non-English-speaking European countries. Furthermore, the desired shooting schedule was impossible for most, and the price was higher than budgeted. For these reasons, the author contacted an Australian acquaintance who has appeared in several Japanese dramas and who is a university theatre arts instructor, and enlisted her help to locate and hire quality performers. Using her connections, a cast was assembled of four performers (including herself): An Australian man in his forties, an American man in his early thirties, the aforementioned Australian woman in her early fifties, and an American woman in her early thirties. All four members of the cast had extensive experience within the Tokyo theatre and improv communities, and, as such, were unfazed by the tight shooting schedule or the requirement that they follow the script and direction exactly. Each was

paid ¥30,000 (approximately £230 as of this writing) for one full day of acting, except for the Australian woman, who was paid ¥40,000 (approximately £300) for casting the project and coordinating the actors' schedules and commutes.

5.3.2. Locations

The scripts called for a wide range of settings (See Table 5.1). Locations for filming were scouted during two trips to the Mita campus of Keio University in Tokyo, due to its central location for the performers. Permission for use of the locations on the day of filming was secured from the facilities management. Care was taken to enable the largest number of locations possible to be filmed in the same building. In addition to this was one outdoor location and two others in adjacent buildings.

5.3.3. Equipment

A Sony professional digital high-definition (HD) video camera with a shotgun microphone and pre-amplifier was checked out from the campus media center, along with a heavy-duty tripod and a Tascam DR-40 portable digital audio recorder. Both of these were configured to record to low-latency SD cards. In addition, a consumer digital HD video camera was also available in case of a failure with the professional unit.

Audio recording presented a challenge in that a professional video production team would normally include a sound technician and boom operator, but these would be unavailable for the current study. For this reason, a pair of Shure wireless lapel mics and receiver, as well as a Behringer two-channel microphone pre-amp were purchased. Thus, the audio capture process consisted of three simultaneous recordings:

1. The audio stream along with the video using the camera's attached shotgun microphone, captured as center-panned stereo

2. The lapel mics, through the external pre-amp, onto one 24-bit, 48 kHz dual-monoaural WAV file on the digital audio recorder
3. The digital audio recorder's internal mics onto a separate 24-bit, 48 kHz dual-monoaural WAV file on the recorder

This redundancy allowed for mixing the various recordings in post-production for the best possible sound quality, and prevented the failure of a single piece of equipment from stopping production.

To ensure that there was adequate lighting for the scenes, two battery-operated adjustable LED lighting panels, light stands ("sticks"), and diffusion paper were also purchased. Additionally, a fifty meter, crank-spooled waterproof extension cable was purchased for outdoor filming. Finally, props for the thirty scenes were purchased, packaged, and tagged with their scene numbers to facilitate properties management on the filming day.

5.3.4. Filming

Prior to filming the scenes, each was storyboarded for a static camera (see Figure 5.2), single-shot take. Scenes were framed to show as much of the characters' bodies as possible, with no close-ups or cutaways. Scenes were lit with available light, the LED panels being employed to eliminate harsh shadows. An assistant monitored audio levels from the lapel microphones while the author directed the performers, operated the camera, and ensured the performances did not deviate from the written script. Heavy rain prevented the use of one outdoor location; these scenes were not filmed. The rain and the delays associated with it also prevented expeditious movement to the two locations outside of the main building, and alternates were located and arranged on the day of filming, made possible by the large collection of versatile props transported to the location, and the performers' own experiences with other projects.

Of the thirty scenes, 23 were successfully filmed, 22 of which were featured on the final instrument either as scored items or example videos. See Appendix D for the filming checklist.

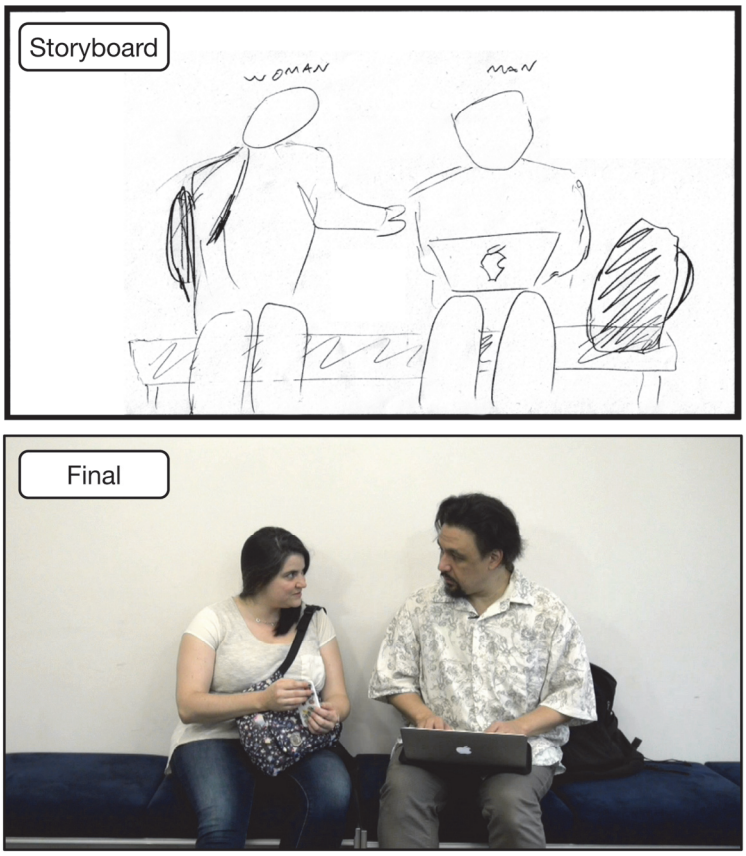

Figure 5.2. Example storyboard and final shot of a listening video.

5.3.5. Video Editing

Video was edited in Apple Final Cut Pro 10 (*Final Cut Pro*, 2015). First, the best take of the scene was chosen, then the audio from the lapel microphones and/or that from the internal microphones on the digital audio recorder for the relevant scene was imported. The audio sources were automatically synchronized to the video using Final Cut's "Synchronize clips" feature. Each audio track was evaluated for clarity, volume, interference, and background noise, and the best was chosen as the final version, or a mix of multiple sources were used for a composite clip. The audio was processed

using Final Cut's various audio processing tools, in addition to other professional digital signal processing tools from Apple's Logic Pro X digital audio workstation package (*Logic Pro X*, 2015), including filters to remove air conditioning hum, compressors to narrow dynamic frequency, and gain controls to remove rustling or compensate for poor microphone reception. In one case, a sound effect was added to indicate that a character had cancelled an incoming call on her mobile phone, as modern smartphones do not give any audio feedback for this action.

Once the audio issues had been corrected and the highest-quality composite vocal track had been assembled, it was grouped into a single video object in Final Cut for any other editing. One performer had significant trouble with remembering lines, resulting in some unnatural gaps between characters' lines. These were shortened by increasing the speed of these sections of video/audio, imperceptibly reducing the gaps. Finally, a fade in from black and fade out to black was added to the beginnings and endings, respectively, of the video scenes. Audio was also faded from and to silence during these transitions.

The videos were then exported to web-compatible 854 × 480 pixel, MP4 video with 48 kHz monoaural audio (effectively 24kHz; monoaural was used in order to lower file size), and also to MP3 audio at 44.1 kHz, 128-bit monaural audio for the audio-only versions. Because MP4 video, though ubiquitous, is rather large, and could therefore threaten the smooth delivery of test content during data collection due to limited network bandwidth (especially if a group of participants was accessing the site from a single location), the MP4 videos were then converted to much-smaller WebM-video/Ogg-Vorbis-audio files using the video format conversion program Adapter (*Adapter*, 2014). This resulted in video files that were sometimes as much as two-thirds smaller than the original MP4 files, with no obvious degradation in quality.

However, as less-sophisticated web browsers (Internet Explorer, Apple's Safari) lack WebM video support, both versions would be necessary in the final instrument (see "Main data collection version" section below). YouTube versions of the videos can be found at the link below:

https://www.youtube.com/playlist?list=PLU2WdxKk5xR9a0ftN9esqfXscqPDd6eOW

5.4. Item Moderation

To ensure that all potential items had clear, unambiguous answers, correct translations, and were properly classified as either "implicit" or "explicit," item moderation by expert judges was carried out, and the items with the most favorable results were used in the final data collection instrument. Although this is not considered a full study, it nonetheless employs a fixed embedded mixed-methods design, insofar as the uses of the quantitative and qualitative data were determined at the outset, and the quantitative data are of primary concern, with the qualitative data called upon for support (Creswell & Plano Clark, 2010). This section describes the process and results of this preliminary step prior to completion of the final instrument for Study II.

5.4.1. Participants

Twelve (12) language testers were consulted, two of whom were native Japanese speakers. Of the twelve, two were established, published listening assessment researchers. The two Japanese speakers were highly-published language assessment researchers. A further two of the remaining were responsible for language testing at their respective institutions, and the remainder were PhD candidates studying language assessment topics at Lancaster University.

5.4.2. Instrument

An instrument was developed specifically for the item moderation step using the quiz module on a Moodle 2.6 server hosted by a large webhost in the United States, and maintained by the present author. It began with a brief introduction to the purpose of the item moderation, including the following explanation of the task types:

Explicit: the answer was stated in the conversation

Implicit: the answer must be inferred

Following the instructions page was an informed consent page, and following that was the instrument proper.

Each of the 23 listening passages was presented in audio-only format, using the MP3s exported from Final Cut after editing. The use of the audio-only condition was to ensure no items were rendered un-answerable without the video. Microsoft Excel was employed to create a Moodle-compatible XML file to bulk-upload the items, and the audio was embedded into the relevant items using the Moodle media embedding feature. An audio player user interface was included with each listening passage to enable the respondents to listen again, etc., if necessary. Each was followed by the associated item candidates in both English and Japanese. The options for each multiple-choice item were shuffled by the quiz module. Following each item was a radio button question asking the respondent to judge the item either “explicit” or “implicit” and a comment field to leave any further comments about the item (e.g., double-keys, suggestions for translation changes). Sixty-five items accompanied the 23 listening passages. See Figure 5.3 for a representation of a listening passage and one item followed by the task type and comment questions.

Figure 5.3. Example of item from the item moderation instrument.

5.4.3. Procedure

An account on the author's Moodle server was created for each of the respondents, and login information was emailed to them. The respondents then logged in and completed the task at their leisure over the course of two weeks.

5.4.4. Results

One Japanese respondent did not have time to answer the questions, but did make translation suggestions on each item. The remaining respondents' responses were used to calculate item facility (IF) statistics, and percentages of agreement with the intended task type. Because of a great deal of rater variation among the respondents having less experience with listening assessment test design and research, the agreement scores for the two respondents with the most published work on the topic

were double-weighted. The problems associated with "expert" judgments of skills tested by items, however, are well-documented in the literature (e.g., Alderson, 1990, 1993; Alderson & Kremmel, 2013; Bachman, Davidson, & Milanovic, 1996; Bejar, 1983). The IF and agreement statistics were summed (expected value = 2.00) for easy comparison. The results of these calculations can be seen in Table 5.2.

Table 5.2.

Tabulated Results of the Item Moderation

Scene	Item	Type	IF	Type Agreement	Sum
1	01.1i	Implicit	1.00	1.00	2.00*
1	01.2e	Explicit	1.00	0.55	1.55
2	02.1e	Explicit	1.00	1.00	2.00*
2	02.2i	Implicit	1.00	1.00	2.00
2	02.3i	Implicit	1.00	0.82	1.82
3	03.1i	Implicit	1.00	1.00	2.00
3	03.2e	Explicit	1.00	0.64	1.64*
4	04.1i	Implicit	0.91	0.00	0.91
4	04.2e	Explicit	0.91	1.00	1.91
4	04.3i	Implicit	1.00	1.00	2.00*
5	05.1e	Explicit	1.00	0.36	1.36
5	05.2e	Explicit	1.00	1.00	2.00
5	05.3e	Explicit	1.00	1.00	2.00*
6	06.1e	Explicit	1.00	0.91	1.91
6	06.2i	Implicit	1.00	1.00	2.00
6	06.3e	Explicit	1.00	0.91	1.91*
7	07.1e	Explicit	1.00	1.00	2.00
7	07.2i	Implicit	1.00	1.00	2.00*
8	08.1e	Explicit	1.00	1.00	2.00
8	08.2i	Implicit	1.00	1.00	2.00*
9	09.1e	Explicit	1.00	1.00	2.00
9	09.2i	Implicit	1.00	1.00	2.00*
11	11.1i	Implicit	1.00	0.64	1.64
11	11.2i	Implicit	1.00	1.00	2.00
11	11.3e	Explicit	1.00	1.00	2.00
11	11.4e	Explicit	1.00	1.00	2.00*
13	13.1e	Explicit	1.00	1.00	2.00*
13	13.2i	Implicit	1.00	1.00	2.00
13	13.3i	Implicit	1.00	1.00	2.00

Scene	Item	Type	IF	Type Agreement	Sum
14	14.1i	Implicit	1.00	0.73	1.73
14	14.2e	Explicit	0.09	0.27	0.18
14	14.3i	Implicit	1.00	0.36	1.36
14	14.4e	Explicit	1.00	1.00	2.00*
15	15.1e	Explicit	1.00	1.00	2.00*
15	15.2e	Explicit	1.00	1.00	2.00
15	15.3i	Implicit	1.00	0.73	1.73
16	16.1i	Implicit	1.00	0.91	1.91*
16	16.2e	Explicit	1.00	0.91	1.91
17	17.1i	Implicit	1.00	1.00	2.00*
17	17.2i	Implicit	1.00	1.00	2.00
17	17.3e	Explicit	1.00	1.00	2.00
18	18.1i	Implicit	1.00	0.91	1.91*
18	18.2e	Explicit	1.00	1.00	2.00
18	18.3i	Implicit	1.00	0.73	1.73
18	18.4i	Implicit	0.64	1.00	1.64
19	19.1e	Explicit	1.00	1.00	2.00
19	19.2i	Implicit	1.00	1.00	2.00
19	19.3e	Explicit	1.00	1.00	2.00*
20	20.1e	Explicit	1.00	0.45	1.45
20	20.2i	Implicit	0.55	1.00	1.55
21	21.1e	Explicit	1.00	0.36	1.36
21	21.2i	Implicit	0.91	1.00	1.91
21	21.3e	Explicit	0.55	0.91	1.45
22	22.1e	Explicit	1.00	1.00	2.00
22	22.2i	Implicit	1.00	1.00	2.00*
23	23.1e	Explicit	0.64	0.18	0.82
23	23.2i	Implicit	1.00	1.00	2.00*
23	23.3e	Explicit	1.00	1.00	2.00
24	24.1i	Implicit	0.64	1.00	1.64
24	24.2e	Explicit	1.00	1.00	2.00*
24	24.3i	Implicit	1.00	1.00	2.00
24	24.4e	Explicit	1.00	0.91	1.91
25	25.1e	Explicit	0.91	1.00	1.91
25	25.2e	Explicit	1.00	0.82	1.82
25	25.3i	Implicit	0.45	1.00	1.45

* Indicates an item used in the final data collection instrument.

IF values ranged from 0.09 through 1.00, with the majority (n = 54; approximately 83%) having an IF of 1.00. The agreement percentage ranged from 0.00 through 1.00, with the majority (n = 43; approximately 66%) with full agreement.

Sums ranged from 0.18 to 2.00, with a slight majority of items with the full score of 2.00 (n = 36; approximately 55%). Based on these scores and the moderators' comments, ten explicit and ten implicit items with favorable results were selected to be included in the final data collection instrument (denoted by asterisks in Table 5.2). Translation notes were incorporated into the final versions as well. Final items accompany the scripts in Appendix C.

5.5. Main Instrument

This section details the creation of the final data collection instrument. Due to the technical requirements of the online delivery method, which was necessary to facilitate bulk data collection, the instrument development process was somewhat involved. This section first describes the online content delivery system and preparations for network connectivity, then details the steps taken to configure the items on the Moodle server; lastly it describes the general layout of the instrument.

5.5.1. Online Content-Delivery Platform and Connectivity

As in the case of the item moderation instrument, the final instrument was to be delivered via the Moodle learning management system's quiz module. However, the need for fast, reliable video content hosting, with specific interface requirements, posed some challenges. In a small-scale study such as those carried out by Suvorov (Suvorov, 2009, 2013, 2015), which also delivered video content using the Moodle quiz module, it is possible to monitor each examinee to ensure that he or she does not pause, rewind, or replay the video content for a "listen once" test design such as the present study. However, with a larger group taking the test on their own personal devices, this is simply not possible. As such, it was necessary that the videos be presented with no control interface at all. A result of this requirement, however, was that the video would need to automatically begin playing as soon as it was loaded.

If the material was hosted with YouTube, a common solution to the problem of video delivery, the embed code could indeed be written in such a way as to autoplay and suppress the controls, but the YouTube branding would remain in the lower right-hand corner of the display. If a user clicks this logo, he or she is taken to the relevant page on YouTube proper, and the controls return. For this reason, it was necessary that the video be hosted on the same server as Moodle.

The requirements for autoplay and suppression of the control interface posed another challenge. Although the HTML5 `<video>` tag in widespread use on the worldwide web can accommodate this easily, the built-in video display features of Moodle 2.6 are not configurable, and the quiz module is not compatible with HTML5 tags. However, during instrument creation, Moodle 3.0 was released and does include HTML5 support within the quiz module. For this reason, the Moodle server was upgraded to Moodle 3.0 and reconfigured.

Another challenge of self-hosting video content is the large amount of bandwidth necessary to ensure that playback is prompt and smooth. To ensure this, the server's hosting package was upgraded from "shared" hosting, meaning the server's files and bandwidth were hosted on a single server along with many other sites, whose traffic could impact performance, to a virtual private server (VPS) with solid-state disk (SSD) storage, ensuring that file retrieval was fast and that memory and bandwidth were specifically allocated to the Moodle server, unaffected by any other sites hosted by the company. For a database-intensive web application such as Moodle, the performance improvement brought by a VPS is immediately obvious. However, even the fastest server will appear slow if the individual users are accessing it via a slow connection. For this reason, a Buffalo-brand commercial 802.11n/a, 600 Mbps wireless access point (WAP) was purchased and configured to create a fast,

dedicated, temporary wireless network connected to the Keio University campus network for data collection sessions, resolving any wireless signal strength issues in the room and shielding the participants from the effect of any other users in the area.

5.5.2. Item Layout

For each item on the test, four Moodle questions were required:

1. The question stem to be presented before listening (as in Study I).
2. The listening passage in audio format.
3. The listening passage in video format.
4. The full multiple-choice item.

As in the item moderation step, Microsoft Excel was used to generate Moodle XML files for bulk upload of the items, including the HTML5 code required to automatically serve the audio and video files *sans* controls, and to provide MP4 video file fallbacks for browsers that did not support the WebM video standard. Audio-only listening passage questions featured a speaker icon in place of a video, and were designated “audio” (音声) at the top of the page, followed by an instruction to listen and click “Next” (次へ) when the dialog finished, and to answer the question. The video versions were designated “video” (ビデオ), and included the same instruction as that found on the audio versions, with the replacement of the verb “to listen” (聞く) with that for “to watch” (見る). See Figure 5.4 for an example.

Figure 5.4. Example of the audio and video versions of the same listening passage.

5.5.3. Layout of the Test

Two forms of the instrument were constructed in Moodle: ODD and EVEN. The ODD form featured video listening passages on odd-numbered items; the EVEN, on even. A list of the final items can be seen in Table 5.3.

An introductory page labeled “Begin the test here” (ここからテストを始めて下さい) was prepared on the Moodle course, which contained a brief overview of the test process, a request that participants take the test in Firefox, Chrome, Opera, or Microsoft Edge browsers, as they supported the WebM standard and would therefore have better performance, but also improve performance for everyone taking the test simultaneously, by reducing network congestion. More importantly, it included a short clip of unused video (“Int’l Man of Mystery”) for the purpose of allowing the participants to ensure that the video played and the sound was audible on their systems. At the bottom of the introductory page were two large buttons instructing the

Table 5.3.

Final Listening Passages, Task Types, and Formats for the Two Forms of the Main Data Collection Instrument

Item #	Title	Type	ODD Format	EVEN Format
1	Licorice	Implicit	Video	Audio
2	Coffee or Tea	Explicit	Audio	Video
3	Notes	Explicit	Video	Audio
4	Concert	Implicit	Audio	Video
5	Double Booking	Explicit	Video	Audio
6	Old Flame	Explicit	Audio	Video
7	Smoking	Implicit	Video	Audio
8	Attire	Implicit	Audio	Video
9	Presentation Consternation	Implicit	Video	Audio
10	Friend Calling	Explicit	Audio	Video
11	Lost Item	Explicit	Video	Audio
12	Registration	Explicit	Audio	Video
13	The Party	Explicit	Video	Audio
14	Mysterious Key	Implicit	Audio	Video
15	Whose Number is This?	Implicit	Video	Audio
16	Crafts Fair	Implicit	Audio	Video
17	Party Aftercare	Explicit	Video	Audio
18	The Ride Back	Implicit	Audio	Video
19	The Play	Implicit	Video	Audio
20	Job Prospects	Explicit	Audio	Video

participant to press one if his/her birthday fell on an odd day, and the other if even. The actual quiz activities were hidden on the Moodle course, ensuring that the only way to enter the test was to first read the introductory page.

The test proper began with a one-page consent form, adhering to the Lancaster University ethical guidelines, the project having received full ethical approval by the Lancaster ethical review board. This was followed by a short demographic section including gender, age, various standardized English test scores, and whether and how much time the participant had spent living in an English-speaking country. Following the demographic section was a practice item section that guided the participant through answering both an audio- and a video-mediated item following the same

pattern as that of the scored items to be presented later (i.e., stem, listening, multiple-choice question, on separate pages). After the practice items were the twenty scored items of the test. Following the scored items was a brief survey on the participants' perceptions regarding the test formats. The questions were as follows:

1. Which question format was harder?

   [Audio / Video / Same]

   より難しかったのはどちらの問題形式ですか？

   【ビデオの問題・音声の問題・ほとんど同じ】

2. Which question format do you think tested your listening ability better?

   [Audio / Video / Same]

   あなたのリスニング能力をより正確に測っているのは、どちらの問題形式だと思いますか？

   【ビデオの問題・音声の問題・ほとんど同じ】

3. On which question format do you think you scored higher?

   [Audio / Video / Same]

   あなたの正答率がより高いのは、どちらの問題形式だと思いますか？

   【ビデオの問題・音声の問題・ほとんど同じ】

4. How much did the videos help you understand the contents of the dialogs?

   [Not at all / A little / A lot]

   会話の内容を理解する上で、ビデオはどれくらい役に立ちましたか？

   【全く役に立たなかった・少しは役に立った・大いに役に立った】

These questions are identical to those used in the author's previously published study (Batty, 2015). The instrument concluded with instructions on how to save the answers and exit.

Each page of the test included only a "Next" button, preventing participants from returning to previous items to change answers or revisit listening passages. Furthermore, custom CSS code was employed to suppress Moodle's standard list of items, which appears on every page of a quiz in the upper right-hand corner, as it crowded the video content on smaller computer screens, and, as each item was spread over three pages (i.e., one for the stem, one for the audio/video, one for the full item), may have had a discouraging effect on participants as they moved through seventy separate pages in the instrument, despite the rather small number of items.

5.6. Chapter Summary

This chapter described the development of the instrument used in Study II. It began with a description of the construct to be addressed, according to the test specification. It then presented the process of passage construction and item design, then detailed the video production process, from location scouting and casting through editing and publishing to the web. It then reported on the item moderation step by which the items to be featured on the main data collection instrument were chosen from the item pool. Finally, it described the main instrument resulting from the steps outlined above. The considerable logistical challenges associated with the video format, in comparison to that of audio-only, will be discussed in the "Conclusions" chapter. Chapter 6 reports on the full Study II, utilizing the instrument described above.

CHAPTER 6. STUDY II

This chapter reports on Study II, which seeks to answer RQ2a, "How does the presence of visual cues interact with items on video-mediated listening comprehension tests," RQ2b, "How does the presence of visual cues interact with task types (explicit and implicit) to influence item responses on video-mediated listening comprehension tests," RQ3a, "How do individual examinee differences interact with the presence of visual cues on video-mediated listening comprehension tests," and RQ3b, "How do examinee perceptions of video interact with performance on video-mediated listening comprehension tests?" The study follows a fixed embedded mixed-methods design, wherein qualitative item analysis is used to aid in the interpretation of the quantitative results (Creswell & Plano Clark, 2010). The main data collection instrument described in the preceding chapter was administered to a large sample of Japanese students of English, and many-facet Rasch modeling was employed to investigate interactions between format of delivery (audio or video) and task type (explicit or implicit).

The chapter begins with an explanation of the method of Study II, including the university which served as the setting for the study, the participants, and a description of the data collection procedure. It then moves to the Data Analysis section, which presents and explains the methods of data analysis employed in the study, with a particularly explicit treatment of the Rasch models and related statistics used.

The results are then presented, beginning with those pertaining to the reliability of the instrument and item characteristics. The results of the item calibrations are then detailed. Comparative format difficulty (audio or video) is then explored, followed by the results of the analyses of format-item and format-task

interactions. Items displaying particularly large difficulty contrasts between the formats are then qualitatively analyzed. The results of this process then prompt the development of a corrected model which is used to investigate the format-task interactions more appropriately. Next, the results of the individual differences analyses are presented, followed by responses to the survey administered after the test.

Finally, the results are discussed with regards to the quality of the instrument, the comparative difficulty of the formats, and their interactions with the task types (explicit and implicit), followed by a discussion of the relationships between individual differences and performance on the two formats. The chapter concludes with a brief summary.

6.1. Method

This section describes the method of Study II, with the exception of the instrument development, which has been described in the preceding chapter.

6.1.1. Setting

The study was undertaken at Keio University's Shonan-Fujisawa Campus (Keio SFC) in Fujisawa, Japan. Keio University was the first Western-style university in Japan, founded in 1858 by Yukichi Fukuzawa, one of the architects of Japan's rapid modernization during the Meiji era (CE 1868 – 1912), beginning as a school of Western studies. It has grown to encompass many fields of study, and is spread across eleven campuses in Tokyo and Kanagawa prefectures. It is ranked by Times Higher Education's World University Rankings as the top private university in Japan ("World university rankings," 2015), and 9th worldwide on the same organization's Alma Mater Index, a ranking of universities by the number of Fortune Global 500 chief executive officers who are graduates ("Alma Mater Index," 2013). Two of its departments are listed in the top-ten university departments in Japan by standardized

test score (*hensachi*) of accepted students; seven of the remaining ten are departments at the nation's most prestigious university, Tokyo University, and Keio is the only private university to be listed in the top ten; the two Keio departments from which data were collected, Policy Management and Environment and Information Studies, are listed in the 22nd and 34th positions respectively ("Daigaku hensachi ichiran/rankingu 2017 [University hensachi list/ranking 2017]," 2016). It is therefore regarded as one of the most prestigious private universities in the country.

Keio SFC is mostly comprised of the two departments mentioned above: Environment and Information Studies (環境情報学部), and Policy Management (総合政策学部). In addition to traditional Japanese students, Keio SFC also includes a large number of "returnee" students—Japanese students who have grown up abroad, usually in English-speaking countries, including many of the students who attended the Keio high school located in New York City. Furthermore, Keio SFC features the GIGA Program, which is an English-mediated degree course open even to students with no Japanese proficiency at all. As such, it is a much more multilingual/mutiethnic university than most in Japan.

The English program at Keio SFC places students into three levels according to their scores on the TOEFL ITP (see the Method section of the Study I chapter for a description of this test) typically administered just prior to the students' first semester. The levels were as follows at the time of administration (CEFR levels in parentheses):

- *A Level:* Under 480 (low B1 and below)
- *B Level:* 480 – 524 (mid B1)
- *C Level:* 525 and above (high B1 and above)

Eight credits (four semester-long classes) of English are required in the first two years of study, but there is no overall curriculum or set of objectives. Instead,

students choose from a large selection of unique English-language electives, matching their schedule, language-learning goals, and personal preferences for content and/or instructor. Given this program design, teachers, most of whom are part-time or outsourced, have a high degree of flexibility and unchallenged autonomy in how they approach their classes. There are no external evaluations or even any coordination beyond that required to register students for the courses. This presents a rare opportunity for data collection, as the individual teachers possess the authority to volunteer class time to data collection without requesting approval from an administrative body. Class sizes vary widely, from as little as one through the low thirties.

6.1.2. Participants

In early December, 2015, the author contacted eleven Keio SFC English teachers by email, explaining the project, and requesting approximately 40 minutes of their class time to administer the test during one of the remaining days in fall semester, which runs until mid-January. Scores would be made available for classes if the teacher wished to discuss the test with the students afterward. Several teachers with listening-focused courses volunteered their classes, viewing the test as a review activity; most simply felt it would be an interesting break from the course content as the semester came to a close. Nine of the eleven teachers who were contacted volunteered class time to the project. Data were collected from the classes of seven of the teachers who volunteered, in addition to the author's own five English classes ($k = 19$). Care was taken to ensure a reasonable mix of proficiency levels, using the program levels (i.e., A – C) as a guide, and also to target larger classes for the sake of efficiency.

The final participants were 279 (168 male; 111 female) consenting, volunteer Japanese undergraduate students of English at Keio SFC. Ages ranged from eighteen

(18) through fifty (50) years old, but the median age was twenty (20) with an interquartile range of three (3) years. Each had a TOEFL ITP score available. TOEFL scores ranged from 330 (CEFR A2) through 597 (CEFR high B2). The original, full sample ($N = 291$) included twelve (12) participants with TOEFL scores over 600 (CEFR C1 begins at 627); these participants were cut from the data set, as most had grown up in English-speaking countries. The mean TOEFL score was approximately 465, with a standard deviation of approximately 52; skewness and kurtosis statistics indicate a fairly symmetrical, nearly-Normal distribution of TOEFL scores in the sample. Number of years of English study ranged from two (2) through eighteen (18), with a median of seven (7) and an interquartile range of three (3). See Table 6.1 for a list of participant demographics and Figure 6.1 for a histogram of the TOEFL score distribution.

Table 6.1.

Participant Demographics

	Age	TOEFL ITP*	Number of years of English
Valid *n*	277	279	269
Missing	2	0	10
Min.	18	330	2
Median	20	467	7
IQR	3	63	3
Max.	50	597	18
Mean	20.27	464.72	7.77
SD	2.846	51.859	2.662
Skewness	6.825	0.014	1.468
SE	0.146	0.146	0.149
Kurtosis	61.202	0.097	2.347
SE	0.292	0.291	0.296

* TOEFL ITP is administered in students' first year of university.

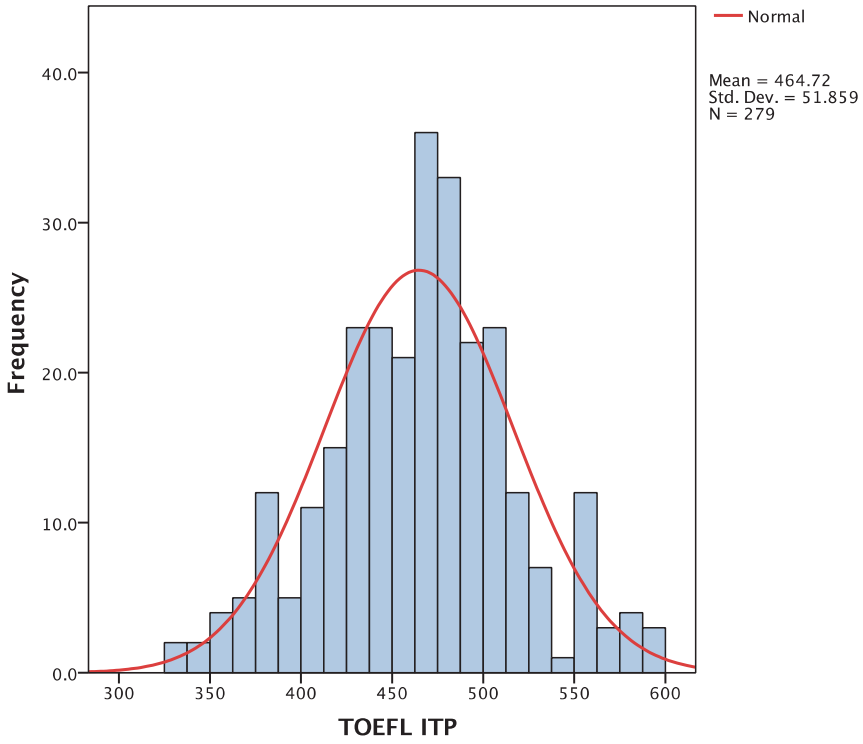

Figure 6.1. Distribution of participant TOEFL scores.

6.1.3. Procedure

In preparation for data collection, accounts for all students who would take part were bulk-uploaded to the Moodle server. On the day of the class, the author installed the WAP for the temporary high-speed wireless network prior to addressing the students. He then introduced himself and the research, explaining that it was for his PhD, and that it was comparing audio- and video-mediated foreign language listening tests. He stressed that participation was voluntary, and that the results would not be used to grade them in the current or any other course, so they need not feel pressure, but if they took part, he would appreciate it if they answered the questions to the best of their ability.

Students had been requested to bring their personal laptops and headphones on the day of data collection, although most students carry both every day already. To expedite logging into the Moodle site, a presentation slide was displayed using the

projector at the front of the classroom during the introduction. It instructed the participants in Japanese to refrain from using Internet Explorer or Safari browsers for the test, and gave the name of the temporary high-speed wireless network to which the participants should connect for optimal performance. Following the introduction, a single presentation slide was displayed with the URL for the Moodle site (moodle.aaronbatty.net), instructions in Japanese on how to log into the site, and animated GIF images demonstrating where to click to enter the introduction page. The author also brought four extra Windows laptops for participants who either neglected to bring their own or who had technical issues with theirs, as well as 100 pairs of "disposable" (i.e., airline-style) inexpensive earbuds purchased in bulk from a Chinese importer, for those who needed them.

After guiding the participants through the introductory page and assisting with any setup, the participants took the test at their own pace as the author observed to ensure that there were no technical difficulties or misunderstandings of the process. As participants finished the test, the author thanked them individually, once again reminding them that they could come to him with any further questions about the research or their participation. Once all participants were finished, the researcher once again thanked the class and the instructor, reminded them of how to contact him with questions, and the equipment was packed and removed from the class. A single data collection typically required approximately 45 minutes, or one-half of a class period, including equipment setup, introduction, administration, and disassembly. Depending on the teacher's preference, this took place at either the beginning or the end of class.

6.2. Data Analysis

This section describes the data analysis methods employed in Study II. Two general approaches to data analysis were used: many-facet Rasch measurement and the more

traditional methods of *t*-test, median test, and ANOVA. This section begins with a brief introduction to Rasch models, then describes the methods to be used in Study II, and finally explains the use of descriptive statistics, *t*-tests, median tests, and ANOVAs.

6.2.1. Rasch Models

Rasch models are a widely-used family of probabilistic models of item response. In order to describe the Rasch-based data analyses used in Study II, a brief description of Rasch models is in order. The present study employs many-facet Rasch measurement (Linacre, 1989); however, due to the nature of the instrument, in most cases the models used will be mathematically equivalent to the original dichotomous Rasch model (Rasch, 1960).

6.2.1.1. Dichotomous Rasch model

The original Rasch model was developed for standard tests with dichotomous vectors of item response, i.e., item responses were either correct (1) or incorrect (0). It places both persons and items on a single cline of person ability and item difficulty on the latent trait (construct) in question. It assumes—or some would argue requires—unidimensionality; i.e., all items measure the same construct (Bond & Fox, 2015; Engelhard, 2013). In addition to estimates of ability and difficulty, fit to the model can also be calculated and is usually reported alongside measures. Finally, test information function (TIF) is used to describe the test's precision of estimates at all levels of ability. These are explained in detail below.

6.2.1.1.1. Estimates

The Rasch model estimates the probability of a person answering any item correctly as a function of the difference between that person's ability on the tested construct and the difficulty of the item. This can be expressed algebraically as in Equation 1 below:

$$P_{ni}(x = 1) = f(B_n - D_i) \quad (1)$$

where

x = Response

B = Person ability

D = Item difficulty

n = Person

i = Item

Thus, person n scoring 1 (i.e., correctly, as opposed to 0 for an incorrect response) on item i is a mathematical function of the difference between his or her ability and the difficulty of the item (Bond & Fox, 2015).

Rasch estimates are presented in log-odds units, or *logits*. As such, the dichotomous Rasch model can be formally expressed as in Equation 2:

$$Ln\left(\frac{P_{ni1}}{P_{ni0}}\right) = \theta_n - \delta_i \quad (2)$$

where

Ln = Natural logarithm

P_{ni1} = Probability of person n scoring 1 (i.e., correctly) on item i

P_{ni0} = Probability of person n scoring 0 (i.e., incorrectly) on item i

θ = Ability parameter

δ = Difficulty parameter

Later extensions to the dichotomous Rasch model accommodate polytomous data, such as partial-credit items and rating scales (Bond & Fox, 2015; Eckes, 2015; Engelhard, 2013); however, the present study includes dichotomous data only.

6.2.1.1.2. Fit

In addition to ability/difficulty estimates, Rasch and other IRT models provide fit statistics to describe a person's or item's fit with the model. Infit and outfit mean-square fit statistics are chi-square statistics derived from the standardized residuals between observed difficulty or ability and the respective model-predicted values. The outfit statistic is unweighted, and is therefore more sensitive to outliers, exaggerating the values for responses with large residuals, and attenuating those for small residuals. Infit is information-weighted and is therefore more sensitive to inliers. These statistics are then averaged for each person or item to calculate their respective infit or outfit mean square statistics. The expected values for these statistics are 1, meaning that when the person ability and the item difficulty is equal, the fit statistics will equal one, indicating perfect fit to the Rasch model; incorrect responses to items for which the probability of success was low will have fit statistics less than 1, and correct responses for items of higher difficulty than the person's ability will have fit statistics greater than 1 (Bond & Fox, 2015; De Ayala, 2009; Engelhard, 2013; Linacre, 2002b).

As for interpretation, the mean-square fit statistics test the hypothesis that the data match the model usefully (Linacre, 2002b, 2003). Rather than having a particular critical value, the mean-square statistics are interpreted on a scale, according to conventions in the relevant literature (Engelhard, 2013). The most oft-cited guideline for interpreting the mean-square fit statistics states that only values above 2.0 are degrading to the measurement system, and that values between 0.5 and 1.5 indicate items productive for measurement (Wright & Linacre, 1994). A more principled

recommendation by Benjamin Wright, by way of Smith, Schumacker, and Bush (1998), is that cutoff values be sample-dependent, with critical values calculated according to Equations 3 and 4 below:

$$\text{Infit}MS = 1 + \frac{2}{\sqrt{N}} \quad (3)$$

$$\text{Outfit}MS = 1 + \frac{6}{\sqrt{N}} \quad (4)$$

Mean fit statistics can be misleading across persons or items, as they will always average to very close to the expected value of 1, even with wholly random data (Stewart, 2014), but small standard deviations in comparison to the mean indicate uniformity in the fit to the model. Item fit statistics are typically used to detect items which do not conform to the assumption of unidimensionality (i.e., the assumption that items appear to load on the same construct as the rest of the instrument).

6.2.1.1.3. Test information function

The final Rasch method of evaluating a test's quality to be discussed here is the test information function. Each item has an information function, maximized at its location on the ability/difficulty cline, where its precision of ability estimation is the highest, and which drops off on either side. Each item's information is rather small, but the total information of the items can be summed at every level of person ability, creating a test information function (TIF) that indicates the amount of information provided by the test at every level of ability (Baker, 2001; De Ayala, 2009). These are typically displayed as plots, visually representing the amount of information a test provides along the range of abilities.

6.2.1.2. Many-facet Rasch model

Linacre (1989) extended the Rasch model to incorporate other aspects of measurement, beyond person and item, with the method known as many-facet

(occasionally "multi-facet") Rasch measurement (MFRM). These aspects are called *facets,* and can be virtually anything the test developer/researcher believes may impact scores (Bachman, 2004; Linacre, 2002a). Individual cases within facets, e.g., individual test takers in a person facet, or individual items in an item facet, are referred to as *elements* of the facet in question. Many-facet Rasch models are most frequently associated with raters and tasks in rater-mediated judgments, allowing the measurement error associated with individual raters' own leniency/severity to be accounted for (Eckes, 2015; Engelhard, 2013). In such cases, a rating scale is typically used. For this reason, the many-facet Rach model is typically formally expressed as in Equation 5:

$$Ln\left(\frac{P_{nmik}}{P_{nmik-1}}\right) = \theta_n - \lambda_m - \delta_i - \tau_k \quad (5)$$

where

P_{nmik} = Probability of person n being rated *k* on item *i* by rater *m*

P_{nmik-1} = Probability of person *n* being rated *k–1* (i.e., one level down on the rating scale) on item *i* by rater *m*

θ_n = Judged ability of person *n*

λ_m = Severity of rater *m*

δ_i = Judged difficulty of item *i*

τ_k = Judged difficulty of rating category *k* relative to category *k–1*

In the case of the present research, however, the response vector is simply dichotomous, and only the second model, which seeks to estimate the difficulties of the two formats (although the person and item estimates are anchored), includes a facet beyond person and item in estimations. That model can be expressed as in Equation 6:

$$Ln\left(\frac{P_{nmi1}}{P_{nmi0}}\right) = \theta_n - \lambda_m - \delta_i \quad (6)$$

where

λ_m = Difficulty of format m

6.2.1.3. MFRM for interaction between facets

Facets in a model can interact with each other, such as when a rater is particularly strict when reading essays on a certain prompt, but not another. These interactions can also be investigated with MFRM, using the software package FACETS (Linacre, 2014) via the use of "dummy" facets—facets whose difficulty/severity estimates are anchored at zero logits, effectively removing them from estimation. This method has frequently been used for detecting rater effects (e.g., Bachman, 2004; Engelhard, 2002, 2007; Myford & Wolfe, 2003), but can be applied to any assessment situation in which multiple facets are theorized to contribute to scores (e.g., Batty, 2015; Engelhard, 2009). An example would be the model used in the next section to investigate the interaction of format with the test items, in which the format facet is dummied (i.e., anchored at 0) to test the hypothesis that the items are equally difficult regardless of format. In that case, in Equation 4, λ (format) is anchored at 0, therefore rendering the equation mathematically equivalent to that of the dichotomous Rasch model, expressed by Equation 2. For a demonstration of this, see Equation 7:

$$Ln\left(\frac{P_{nmi1}}{P_{nmi0}}\right) = \theta_n - 0 - \delta_i = \theta_n - \delta_i \quad (7)$$

Interaction terms between facets are then calculated after the main effects model, using the residuals of the two facets being investigated. This is accomplished by estimating the measures for the elements of each facet normally, then anchoring the measures and estimating a second-order model with the residuals included as a

new facet of measurement. Since only residuals remain after anchoring the first-order model's facet estimates, the interaction term can be estimated without affecting existing estimates. The residuals for the facets of interest (e.g., an item when delivered in either the audio or video format) are summed, and any nonzero sum indicates an interaction (Linacre, n.d.). Finally, the size of the interaction is estimated with the bias statistic via Equation 8 (following the example set forth in Equation 6):

$$t_{mi} = \frac{\hat{\varphi}_{mi}}{SE_{mi}} \tag{8}$$

where

t_{mi} = Bias statistic for the interaction between format m and item i

φ_{mi} = Interaction term between format m and item i

SE_{mi} = Standard error of the interaction term estimate

The bias statistic t can be interpreted as a Student's t for statistical significance, as its distribution is nearly equivalent. The degrees of freedom are the number of observations minus one (Eckes, 2015).

In addition to the bias statistics, the MFRM software package FACETS (Linacre, 2014) provides a set of "biased" measures for elements in an interaction analysis, which are the element's overall measure plus the interaction term for the context (i.e., the facet of interest). In the example above, this would take the form of an item difficulty estimate plus the interaction term for that item and the audio format, or, alternately, that of the video format. In a pairwise report, one of these biased measures is subtracted from the other, providing a *contrast value* that demonstrates the difference in measures for that element in the two contexts. As such, with the format-by-item example, it is possible to estimate the logit difference of difficulty for the item in question when administered as an audio item, versus when it is

administered in the video format, and the bias statistic can be used to determine statistical significance of the difference between the measures.

6.2.2. Rasch Measures in Study II

The dichotomous Rasch model is first used in Study II to investigate test reliability and item quality via separation statistics and item fit statistics. MFRM is used first to estimate the comparative difficulty of the two formats of delivery (audio and video), and then, using dummy facets, the interactions between format and task, and format and items. Biased estimates for the audio and video formats for each individual examinee are also used.

6.2.3. Descriptive Statistics, Median Test, *t*-Test, and ANOVA in Study II

A Mood's median test is used to investigate the equivalence of the ODD and EVEN forms, comparing the person measures between the two forms. To investigate individual differences' interaction with performance on the two formats (audio and video), *t*-tests and ANOVA of individual contrast values are employed.

This section has described the data analysis methods employed in Study II. It began with a brief overview of the dichotomous and many facet Rasch models, and explained how the latter is used to investigate facet interactions. It then described how each statistical method was applied to the data in question. The next section reports the results of these analyses of the test data.

6.3. Results

This section reports the results of the data analyses on the Study II data. It begins with the reliability and item fit of the instrument, then moves to a comparison of overall difficulty between the audio and video formats. Next, it reports the results of analyses

of format interaction with other features of the test itself. Finally, it reports on individual examinee differences pertaining to format and overall ability.

6.3.1. Reliability and Item Characteristics

Rasch summary statistics were consulted in order to ascertain whether the statistical properties of the test were sufficient to use for judgments regarding the interaction between delivery format and task type. Person measures and reliability will be discussed first, followed by the item characteristics.

6.3.1.1. Person measures and reliability

Summary Rasch statistics for the persons can be seen in Table 6.2. As the Rasch estimations were item-centered (i.e., the item difficulty average was scaled to zero, and all other estimates are in relation to that average), the average person estimate was 0.61 (*SD* = 0.15). A Wright variable map of these two distributions can be seen in Figure 6.2. The Wright map places the distribution of person ability estimates aside the item estimates for the purpose of comparison. The person estimates are positively oriented, meaning higher logit values indicate higher ability. The item estimates are negatively oriented, meaning that higher logit values indicate more difficult items.

Table 6.2.

Rasch Summary Statistics for the Person Facet

	Measures		Infit *MS*		Outfit *MS*			Reliability of			
N	Mean	*SD*	Mean	*SD*	Mean	*SD*	**Strata**	**separation**	χ^2	***df***	***p***
279	0.61	0.15	1.00	0.18	1.03	0.37	2.43	0.71	753.0	278	0.00

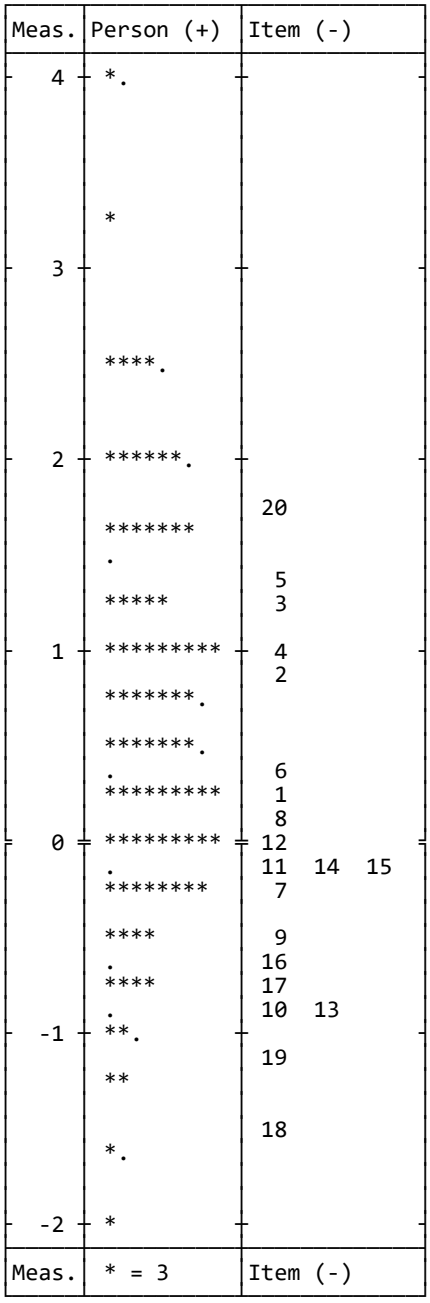

Figure 6.2. Wright variable map for persons and items.

Reliability of person ability estimates is maximized by a close match between the person sample abilities and the item difficulties (Bond & Fox, 2015). As can be seen by the Wright map, the items may have been very slightly too easy for the person sample, although there is something of a positive skew on the distribution of ability estimates. The top person estimates indicate full scores. This does not impact the estimations of others' ability estimates, however, as FACETS removes scores of 0%

and 100% from the estimation, and only assigns logit scores to them once the rest of the sample's estimates have been completed, based on the anchored values of those ability estimates. Overall, the spread of person estimates represents a wide range of abilities, as was intended with the choice of sample.

The persons can be separated into two strata with difficulty estimates three standard errors apart, with a reliability of person separation of 0.71, which can be interpreted similarly to a Cronbach alpha reliability coefficient (Fisher, 1992; Wright, 1996; Wright & Masters, 2002). Although these reliability measures are not terribly strong, they are not wholly unexpected with such a short test, and the fact that the items displayed a large format effect (to be discussed later), both of which negatively impact such estimates of reliability (Bachman, 1990, 2004). The chi-square statistic is the result of a test of the hypothesis that all the person ability estimates are equivalent; the significant statistic seen here indicates that they are statistically distinct (Linacre, 1992).

6.3.1.2. Item measures

For the present research, the item statistics are of more interest than the person measures. Table 6.3 displays the Rasch summary statistics for the items, and individual item calibration statistics for the items can be seen in Table 6.4. As the Rasch estimations were item-centered, the average item difficulty was 0.00 (*SD* = 0.86) logits, indicating a fairly wide range of difficulties, which is visualized by the Wright variable map in Figure 6.2. The items are fairly evenly spread from −1.49 through 1.70 logits of difficulty (higher values are harder), with the exception of a mode just below the average difficulty of 0.00. Another small mode can be found even lower, at just above −1.00 logits. The average difficulty of the explicit items was 0.29, and that of the implicit items was −0.29.

Table 6.3.

Rasch Summary Statistics for the Item Facet and Task Types

		Measures		Infit *MS*		Outfit *MS*			Reliability of			
Facet	***N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	**Strata**	**separation**	χ^2	***df***	***p***
Items	20	0.00	0.86	0.99	0.09	1.03	0.15	8.02	0.97	632.4	19	0.00
Explicit	10	0.29	0.97	0.98	0.09	1.02	0.15					
Implicit	10	−0.29	0.72	1.01	0.10	1.04	0.16					

Table 6.4.

Calibration of the Item Facet

Item	**Name**	**Type**	**Measure**	***SE***	**Infit *MS***	**Outfit *MS***
1	01.1i	Implicit	0.28	0.14	1.07	1.11
2	02.1e	Explicit	0.85	0.14	0.95	0.97
3	03.2e	Explicit	1.24	0.15	1.11	1.16
4	04.3i	Implicit	1.03	0.15	1.14	1.24
5	05.3e	Explicit	1.36	0.14	0.83	0.85
6	06.3e	Explicit	0.37	0.13	0.96	0.92
7	07.2i	Implicit	−0.19	0.14	0.92	0.96
8	08.2i	Implicit	0.13	0.15	1.16	1.17
9	09.2i	Implicit	−0.48	0.14	0.93	1.06
10	11.4e	Explicit	−0.85	0.15	1.02	1.17
11	13.1e	Explicit	−0.13	0.14	0.93	0.95
12	14.4e	Explicit	0.01	0.14	0.86	0.79
13	15.1e	Explicit	−0.88	0.15	0.99	1.11
14	16.1i	Implicit	−0.15	0.15	1.12	1.25
15	17.1i	Implicit	−0.17	0.14	0.97	0.93
16	18.1i	Implicit	−0.67	0.15	0.94	0.86
17	19.3e	Explicit	−0.80	0.15	1.05	1.01
18	22.2i	Implicit	−1.49	0.18	0.93	1.06
19	23.2i	Implicit	−1.18	0.16	0.89	0.76
20	24.2e	Explicit	1.70	0.15	1.08	1.25

The average fit estimates both demonstrate uniform fit to the Rasch model, with mean fit statistics very close to one, and small standard deviations in relation to the means. Average fit statistics for the task types (explicit and implicit) were also very close to 1, although the implicit items had slightly worse fit. Standard deviations on the task types' fits indicate little variation around the means, which is another indication of good overall model fit. The items can be separated into eight strata of difficulty with a reliability of 0.97, with estimates lying three standard errors apart,

and the chi-square test indicates that item difficulty estimates are significantly different from one another.

Table 6.4 displays Rasch measures and fit statistics for the items. Infit mean-square values range from 0.83 through 1.16, and outfit, from 0.76 through 1.25. These are well within the most-frequently cited set of cutoff values for fit statistics (i.e., Wright & Linacre, 1994), which deems values between 0.50 and 1.50 as "productive for measurement." Additionally, using the sample-dependent critical value calculation presented by Smith et al. (1998) results in cutoff values of 1.45 and 2.34 for infit and outfit, respectively. By either convention, these items display excellent fit to the Rasch model, which will facilitate the estimation of bias terms in the analyses following.

The TIF for the present instrument can be seen in Figure 6.3. As shown, the test renders the most information about examinees with abilities at just under zero logits, which was implied by the Wright variable map previously. The test continues to estimate ability with some precision all the way to the lowest-ability examinees, and does not drop to zero at any point among the observed person ability estimates.

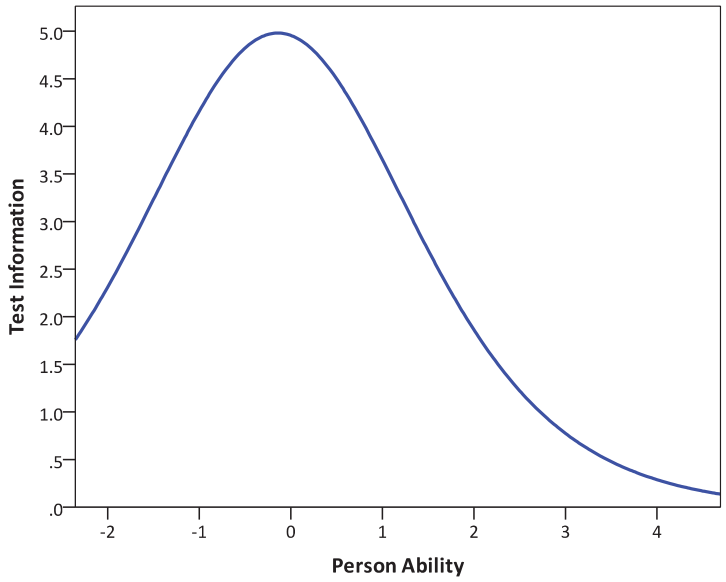

Figure 6.3. Test information function for the instrument. The horizontal axis is scaled to the person measures observed (−1.86 to 4.48).

6.3.2. Format Difficulty

A second many-facet Rasch model was constructed, including a facet for the format of delivery, as a means of comparing the two formats' difficulties. Such a model is ambiguous, since disjoint subsets arise from the fact that each person only encountered each item in one format or the other; i.e., the two formats themselves are disjoint. However, since the purpose of this model is merely to compare the difficulties of the formats to each other and place them on the same cline of ability/difficulty as that of the persons and the items, this problem is addressed by anchoring both the person and item estimates to their values in the previous, two-facet model, leaving only the format facet to be estimated in relation to them (Linacre, 2013). The difficulty estimates and other Rasch statistics for the format facet can be seen in Table 6.5, and a Wright variable map placing the formats alongside the person and item distributions can be seen in Figure 6.4. Because there are only two formats, and only the person estimates are allowed to "float," the difficulties will always "straddle" the mean facet difficulty of 0 logits, and therefore indicate their distance from each other. All other analyses following will only include the format as a dummy facet (i.e., with its location anchored at 0). As can be seen in Table 6.5, the two formats are 0.92 logits apart, with the audio format being more difficult than the video format. Both infit and outfit mean-square statistics are very close to the expected value of 1.

Table 6.5.

Rasch Measure and Fit Statistics for the Formats

Format	Measure	*SE*	Infit *MS*	Outfit *MS*
Audio	0.46	0.05	1.01	1.11
Video	−0.46	0.05	0.97	0.95

6.3.3. Format-Task and Format-Item Interactions

A third MFRM model was constructed, including both the format of delivery and the task type as dummied (i.e., estimates anchored at 0) facets to investigate the interactions between format and task type, and between format and individual items.

6.3.3.1. Format interaction with task type

The pairwise bias report for the interaction between format and task type can be seen in Table 6.6. In both the case of explicit and that of the implicit items, the video format was substantially easier. Explicit items delivered in the video format were an average 0.56 logits easier than when delivered in the audio format. The effect was even more pronounced with implicit items, which were found to be an average of 1.26 logits easier when delivered in the video format. In both cases, the contrasts were statistically significant, with *p* values lower than 0.000.

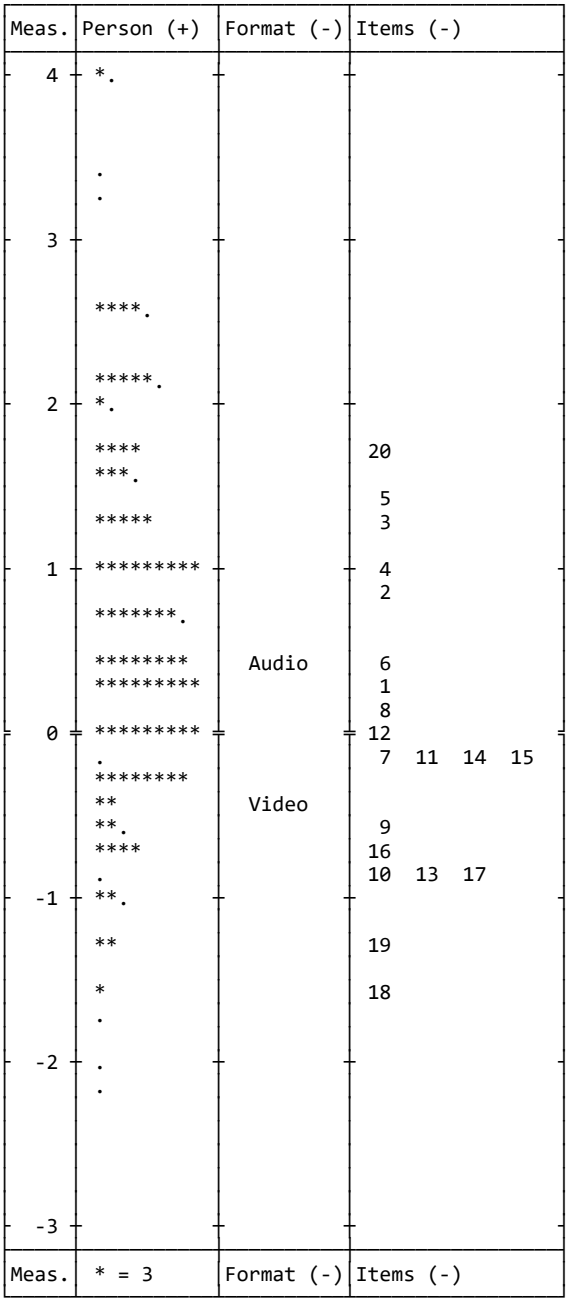

Figure 6.4. Wright variable map of persons, format, and items.

Table 6.6.

Pairwise Bias Report for Format Versus Task Type

	Audio		Video						
Type	**Measure**	***SE***	**Measure**	***SE***	**Contrast**	***SE***	***t***	***df***	***p***
Explicit	0.28	0.06	−0.28	0.06	0.56	0.09	6.20	2734	0.000
Implicit	0.58	0.06	−0.67	0.07	1.26	0.09	13.35	2721	0.000

6.3.3.2. Format interaction with items

To investigate how individual items interacted with the formats, the bias terms between the formats and items were calculated and a pairwise report prepared (Table 6.7). Those with particularly large contrast values were analyzed qualitatively.

Table 6.7.

Pairwise Bias Report for Format Versus Items

			Audio		Video						
No.	**Item**	**Type**	Measure	*SE*	Measure	*SE*	**Contrast**	***SE***	***t***	***df***	***p***
1	01.1i	Implicit	1.45	0.21	−0.88	0.22	2.33	0.31	7.57	270	0.0000
2	02.1e	Explicit	1.35	0.19	0.32	0.19	1.03	0.27	3.82	271	0.0002
3	03.2e	Explicit	1.04	0.20	1.42	0.19	−0.38	0.28	−1.36	269	0.1753
4	04.3i	Implicit	1.89	0.21	0.17	0.19	1.72	0.28	6.11	270	0.0000
5	05.3e	Explicit	1.56	0.22	1.21	0.19	0.35	0.29	1.23	268	0.2181
6	06.3e	Explicit	0.65	0.18	0.06	0.19	0.60	0.27	2.23	270	0.0264
7	07.2i	Implicit	0.54	0.19	−1.07	0.23	1.61	0.30	5.33	270	0.0000
8	08.2i	Implicit	0.65	0.18	−0.44	0.20	1.10	0.27	4.02	270	0.0001
9	09.2i	Implicit	0.06	0.19	−1.14	0.23	1.20	0.30	3.93	271	0.0001
10	11.4e	Explicit	−0.32	0.20	−1.49	0.24	1.17	0.31	3.72	266	0.0002
11	13.1e	Explicit	0.39	0.19	−0.69	0.21	1.08	0.29	3.78	271	0.0002
12	14.4e	Explicit	−0.05	0.19	0.07	0.19	−0.12	0.27	−0.44	270	0.6581
13	15.1e	Explicit	−0.49	0.20	−1.38	0.25	0.89	0.32	2.77	270	0.0060
14	16.1i	Implicit	0.20	0.19	−0.53	0.20	0.73	0.28	2.64	270	0.0088
15	17.1i	Implicit	0.65	0.19	−1.14	0.23	1.79	0.31	5.87	271	0.0000
16	18.1i	Implicit	−0.32	0.20	−1.06	0.22	0.74	0.30	2.50	269	0.0131
17	19.3e	Explicit	−0.69	0.21	−0.91	0.22	0.22	0.30	0.71	270	0.4798
18	22.2i	Implicit	−0.98	0.23	−2.11	0.30	1.13	0.37	3.03	262	0.0026
19	23.2i	Implicit	−1.21	0.23	−1.14	0.23	−0.07	0.33	−0.21	271	0.8301
20	24.2e	Explicit	2.16	0.22	1.21	0.21	0.95	0.30	3.20	271	0.0015

With the exception of items 3, 5, 12, 17, and 19, all contrasts were statistically significant at the 0.05 level, at least, and most were below the 0.01 level as well. Ten of the twenty items exhibited contrasts in excess of 1 logit, biased in the direction of the video format, and of these, seven are implicit items. Only items 3, 12, and 19, two of which are explicit items, show a bias in the direction of audio, and these effects are very small (|0.38|, |0.12|, and |0.07| logits). Detailed multiple-choice response

proportions for the two test forms, including format of delivery, are displayed in Table 6.8.

Table 6.8.

Response Proportions for the Items on Both Forms of the Instrument

Item	Response	ODD (*n* = 146)	EVEN (*n* = 133)	Item	Response	ODD (*n* = 146)	EVEN (*n* = 133)
1	Key	0.79	0.29	11	Key	0.77	0.50
	Dist. 1	0.08	0.16		Dist. 1	0.01	0.09
	Dist. 2	0.04	0.40		Dist. 2	0.16	0.35
	Dist. 3	0.09	0.14		Dist. 3	0.06	0.06
2	Key	0.38	0.51	12	Key	0.66	0.56
	Dist. 1	0.29	0.27		Dist. 1	0.01	0.02
	Dist. 2	0.05	0.04		Dist. 2	0.18	0.25
	Dist. 3	0.28	0.18		Dist. 3	0.16	0.17
3	Key	0.37	0.37	13	Key	0.86	0.67
	Dist. 1	0.12	0.26		Dist. 1	0.09	0.20
	Dist. 2	0.01	0.02		Dist. 2	0.02	0.05
	Dist. 3	0.49	0.35		Dist. 3	0.03	0.08
4	Key	0.29	0.54	14	Key	0.61	0.68
	Dist. 1	0.09	0.23		Dist. 1	0.18	0.11
	Dist. 2	0.10	0.07		Dist. 2	0.12	0.14
	Dist. 3	0.51	0.17		Dist. 3	0.09	0.08
5	Key	0.41	0.28	15	Key	0.83	0.44
	Dist. 1	0.25	0.29		Dist. 1	0.10	0.36
	Dist. 2	0.18	0.17		Dist. 2	0.02	0.09
	Dist. 3	0.16	0.26		Dist. 3	0.05	0.11
6	Key	0.52	0.56	16	Key	0.71	0.77
	Dist. 1	0.23	0.20		Dist. 1	0.12	0.16
	Dist. 2	0.22	0.17		Dist. 2	0.08	0.05
	Dist. 3	0.03	0.07		Dist. 3	0.10	0.03
7	Key	0.82	0.47	17	Key	0.79	0.71
	Dist. 1	0.05	0.26		Dist. 1	0.04	0.10
	Dist. 2	0.05	0.19		Dist. 2	0.08	0.11
	Dist. 3	0.07	0.08		Dist. 3	0.08	0.09
8	Key	0.52	0.66	18	Key	0.81	0.89
	Dist. 1	0.32	0.23		Dist. 1	0.08	0.05
	Dist. 2	0.02	0.02		Dist. 2	0.05	0.01
	Dist. 3	0.14	0.09		Dist. 3	0.07	0.05
9	Key	0.83	0.56	19	Key	0.83	0.79
	Dist. 1	0.10	0.28		Dist. 1	0.08	0.08
	Dist. 2	0.03	0.11		Dist. 2	0.06	0.10
	Dist. 3	0.03	0.05		Dist. 3	0.03	0.04
10	Key	0.71	0.83	20	Key	0.25	0.34
	Dist. 1	0.04	0.03		Dist. 1	0.34	0.23
	Dist. 2	0.18	0.09		Dist. 2	0.18	0.22
	Dist. 3	0.08	0.05		Dist. 3	0.23	0.21

NOTE: Underlines indicate video format.

For every item displaying a contrast greater than 1 logit or a bias toward audio, the corresponding video was re-examined for the purpose of identifying any extralinguistic information which may have exerted an inordinately large facilitative effect upon scores.

In addition to nonverbal cues, the videos provide some limited schematic context for the interaction, which was not presented to the audio-only group. However, the information provided is extremely sparse, as shooting locations were limited. For many of the scenes, the setting is an unremarkable bench placed against a simple white wall. Most schema-setting, therefore, was operationalized through costumes, although each actor only brought two or three outfits. Due to this fact, as well as the observation in Study I that participants neither oriented directly toward the setting with much regularity, nor did they mention it in their retrospective interviews, the largest source of visual information is understood to be that of the nonverbal cues of the characters appearing in the videos. As such, the descriptions of the high-contrast and audio-biased items below will primarily focus on the presence/absence of nonverbal cues.

6.3.3.2.1. Item 1

This implicit item had the largest contrast on the test, at 2.33 logits. In this scene, a man is sitting in an office break room snacking on black licorice. A woman enters in search of something sweet to eat. The man offers the licorice, but the woman declines because she thinks it tastes like medicine, and expresses a preference for donuts. However, the donut box is empty. The man slyly offers the licorice again, and the woman replies “Oh, what the hell…” The implicit question asked “What will the woman do next,” (女性は次に何をするでしょうか) with the credited response being “Eat some licorice” (リコリスを食べる). Although it was not scripted as such,

the actress in the video reached for, and took a piece of licorice as the video faded out (see Figure 6.5).

Figure 6.5. Unscripted extra-linguistic information in Item 1.

This was first noticed in video editing, but because the actress delivered the line as she moved, it could not be edited out, and it was hoped that the sound of her hand entering the bag of sweets would be sufficient indication to those who encountered the item in the audio format that she was changing her mind. Although the sound of the plastic bag was quite apparent in the audio version of the item, the extra information in the video format appears to have been enough to introduce a very large facilitative effect over audio, as 79% of those who encountered the item in the video format selected the credited response. The most common response for those who encountered the audio version was, “Buy more donuts,” (ドーナッツをもっと買う), which is somewhat surprising, as this is never mentioned in the dialog. It is worth noting, however, that despite the large video bias on this item, the fit statistics were relatively good, with an infit mean-square of 1.07 and an outfit of 1.11.

6.3.3.2.2. Item 2

This explicit item, wherein a man is worried that the coffee shop a woman has suggested will not serve tea, was the fifth-hardest item on the test, and the fifth-hardest item in the audio format. No extralinguistic information aside from nonverbal cues were identified in the video to explain the format contrast of 1.03 logits.

6.3.3.2.3. Item 3

This explicit item was the third-hardest item on the test at 1.24 logits, and was intended to be a particularly difficult item. It is one of the few items that was biased for audio, although the contrast is fairly small (0.38 logits). The item itself concerns a male student asking a female student if he can borrow her notes for a class that he missed. The female student complains that his excuses for missing class are never good, and lists some examples. The man then asks what the woman would consider a good reason for missing class. The woman thinks and replies, "For example, 'my train was delayed.'" The man then pauses and asks, mischievously, "Is it too late to say 'my train was delayed?'" The woman laughs and hands him the notes. The question asked why the male student missed class, which is never stated. The most frequently-selected answer (approximately 49% of examinees) for those who encountered the item in the video format was "His train was delayed" (乗っていた電車が遅れたから), but there was nothing the author could detect in the video that would have made that option more attractive in the video format.

6.3.3.2.4. Item 4

This implicit item had the third-highest video bias with a contrast of 1.72. The fit statistics are somewhat high, with an outfit mean square of 1.24, which, although still productive for measurement, is outside of the range usually recommended for

high-stakes tests (Wright & Linacre, 1994). The scene depicts a female coworker attempting to convince a male coworker to lie to their boss so that both of them can attend a concert together. The man implies that he is convinced and will call in sick, which is the information targeted in the item. Those who encountered the video version of this item selected the correct answer 54% of the time, whereas an almost equal proportion (51%) of those who encountered it in the audio format answered that he would report to work normally. The only discernable difference between the audio and video formats was the absence or presence of nonverbal cues. The actor concludes by first chastising his coworker, frowning, as he says, “That is very dishonest.” His facial expression then changes to a sly smile as he continues, “But it just might work” (see Figure 6.6).

Figure 6.6. Facial expression on the man’s line, “But it just might work” in Item 4.

Clearly, in the case of Item 4, examinees’ who answered the video version of the item were more likely to pick up on the implication of the man’s closing lines, despite the fact that his vocal intonation and phrasing would also have indicated this mood change. As an audio item, its difficulty is estimated at 1.89 logits, which would put it at the top end of difficulty for this test. However, as a video item, it was estimated at only 0.17 logits of difficulty.

6.3.3.2.5. Item 7

This implicit item exhibited the fourth-largest contrast, and was both a fairly easy item at −0.19 logits, and had excellent fit. However, this overall difficulty estimate includes the large 1.61-logit contrast between audio and video. As an audio item, its difficulty is estimated to be 0.54 logits, but when delivered in the video format, it dropped to −1.07. In this video, a man and a woman are sitting at a table in a café. The man produces a pack of cigarettes and pulls one out. The woman gently chides him for his habit. He admits he would like to quit, but has found it difficult. The woman shares her "secret" to quitting: "You have to *want* to quit." The man chuckles and responds "Yeah, there's no hope for me, then." The implicit question accompanying this video asked, "What did the man think at the end?" (最後に男性はどのように思いましたか). The credited response was "He will not be able to quit smoking" (タバコをやめられない). In the video, however, the man places the cigarette into his mouth just after delivering this line (see Figure 6.7; note the darkened frame, as the video fades out).

Figure 6.7. Unscripted extralinguistic information in Item 7.

Approximately 82% of those who encountered the video version of the item selected the correct answer. Those who encountered it as an audio item were still most attracted by the correct answer (47%), although the first two distractors ["He will take the woman's advice" (女性のアドバイスを受け入れる), and "He should start by smoking less each day" (毎日少しずつ吸う事を減らし始める)] also attracted fairly heavily. This item suffered the same problem, albeit to a lesser degree, as that of Item 1. However, unlike Item 1, the likelihood of a problem was not suspected by the present researcher during post-production, as it is subtle enough that it went unnoticed until re-inspection later. Although the contrast in Item 7 is not as large as that of Item 1 (1.61 versus 2.33 logits), this extra information was suspected to be the cause of at least part of the effect.

6.3.3.2.6. Item 8

This implicit item had a contrast of 1.10 logits in favor of video, and was the seventh-hardest item under both formats (tied with Items 6 and 15 in the audio format), and the eighth overall. The fit is reasonably good at 1.16 for infit, and 1.17 for outfit. Despite the relatively large contrast, however, no extralinguistic information beyond nonverbal cues was detected in the video.

6.3.3.2.7. Item 9

This implicit item also appeared to feature no noticeably facilitative visual information aside from nonverbal cues, despite having the fifth-largest video contrast at 1.20 logits. The scene depicts a man and a woman at a conference (they are wearing nametags and carrying schedules); the man is nervous about his first time presenting at a conference. He realizes that the woman is presenting today, and asks if she is nervous; she explains that on the contrary, she's mostly excited, as she enjoys talking

to people about her work. She assures the man that when he begins speaking at his presentation the following day, he will be fine. The man seems unconvinced.

The item was below average in difficulty, even in the audio format (−0.06 logits), but was very easy when encountered as a video item (−1.14 logits). Despite this discrepancy, fit was very near ideal at 0.93 for infit and 1.06 for outfit. No facilitative visual information was detected in the video beyond nonverbal cues.

6.3.3.2.8. Item 10

This explicit item was fairly easy in both formats (audio: −0.32; video: −1.49), and had reasonably good fit (Infit *MS* = 1.02; Outfit *MS* = 1.16). In both cases a large majority of examinees selected the credited response. An examination of the video revealed no visual information that could have been unintentionally facilitative beyond that of nonverbal cues.

6.3.3.2.9. Item 11

This explicit item has, by the standards of the present research, a fairly "small" video effect, with a contrast of 1.08 logits, and estimated format difficulties in the middle of their respective ranges. Fit is excellent at 0.95 for both infit and outfit. The scene portrays a man searching for his briefcase at a lost and found table, but the woman misunderstands him and offers two incorrect bags. The question simply asked what the man was looking for. Since the man's bag is never located, and thus never visible, however, there is no "unfair" advantage to the video. Half of the examinees who encountered the audio version of the item chose the correct response.

6.3.3.2.10. Item 12

This explicit item was one of the three items which were biased toward audio, however slightly. It was the fifth-hardest video item, but the twelfth-hardest audio item, with a contrast value of −0.17. Although the majority of both those who

encountered this item in the audio format, as well as those in the video format, selected the correct response, the second distractor seemed to be more attractive to the video group. The scene depicts a woman checking into a conference and complaining that she could not pre-register because the website would not accept her card. The male attendant at the reception desk assures her that she can register onsite, but that it costs more than the online rate. The woman complains, and the man relents and offers her the lower price. The question asked, “Why didn’t the woman pay online” (女性はなぜオンラインで払いませんでしたか), and the credited response was, “There was a problem with the website” (ウェブサイトに問題があったから). However, the video group seemed somewhat more attracted than the audio group to “She did not realize it was more expensive onsite” (現地で払う方が高いということを知らなかったから). No extra information in the video was detected, beyond nonverbal cues.

6.3.3.2.11. Item 15

This implicit item had the second-largest video bias, with a contrast value of 1.79. It was a fairly easy item with a difficulty estimate of −0.17; however, its estimate under the audio condition is a much-harder 0.64, whereas the video estimate is −1.14. Despite this discrepancy, the fit statistics are excellent, at 0.97 for infit and 0.93 for outfit. The scene depicts a woman finding an unknown phone number on her husband’s mobile phone, with whom he has been talking for a long time, and also at times when she was out of town. Her shock and anger rises as she finds it over and over, and the scene concludes with her thrusting the phone’s screen into her husband’s face, demanding, “Whose number is this?” The man looks mortified and guilty. The question asked, “Why is the woman upset” (女性はなぜ気分を害していますか), with the rather obvious answer, “She thinks the man is having an affair” (男性は浮気

していると思っているから). However, although that was still the most-common answer among those who encountered the item in the audio format (44%), a similar percentage (36%) chose the distractor, “She thinks the man forgot to call the restaurant” (男性はレストランに電話をかけ忘れたと思っているから). There is very little in the script to suggest this, as the restaurant is only mentioned in the first line, and the rest of the scene is the woman grilling the man about the number, finally demanding to know whose it is. The video group, however, chose the credited response 83% of the time. The only discernable difference between the two formats was the presence of the nonverbal behavior, although the behavior was particularly striking in this case.

6.3.3.2.12. Item 18

This implicit item was the easiest on the test, with a difficulty estimate of −1.49, and with good fit statistics of 0.93 for infit and 1.06 for outfit. The contrast value was the seventh-largest on the test, with a 1.13-logit contrast between audio and video. In the scene, a woman is struggling to set up a computer and projector before giving a talk while a man interrupts her, asking about her transportation wishes after the talk. She answers absentmindedly a few times, before snapping, “You know what? Can we talk about this later?” The video included the visual information of struggling, although the sounds were clearly audible on the audio version, and most of her frustration is expressed vocally. Furthermore, the video opens with the man saying, “I’m sorry to bother you while you’re preparing for your talk,” and ends with his apology. Since the question was very easy for both groups, with 83% of those who saw the video responding correctly versus 79% of those who did not, this item was deemed not to have sufficient extra information in the video as to unfairly bias responses.

6.3.3.2.13. Item 19

This implicit item was one of the three with a very slight bias toward audio, with a contrast of only −0.07 logits. It was the second-easiest item on the test, with a difficulty estimate of −1.18, and audio and video difficulty estimates of −1.21 and −1.14, respectively. Fit is extremely close to the model at 0.89 and 0.76 for infit and outfit. Strangely, it appears that a very slightly larger percentage of the ODD group, which encountered the item in the video format, selected the credited response than did the EVEN group, which encountered it in audio. Since the measures are all parameterized, however, this difference does not seem to be reflected in the difficulty estimates.

6.3.3.3. Equivalence of forms

In the analyses above, the videos for Items 1 and 7 were found to include extra, factual information that appears to have unintentionally benefitted those who encountered it in the video format. This presents a serious problem. Because both of the items were odd-numbered, the much-easier, video versions were delivered to the same subsample: those with odd-numbered birthdays. To determine whether this effect was large enough to affect the equivalence of the test forms, the person ability measures for those who sat the ODD and those who sat the EVEN forms were compared. A Levene's test for the equality of variances demonstrated that variances were not significantly different, but Q-Q plots and a Shapiro-Wilk test both indicated departures from Normality that would violate the Normality assumption of a *t*-test, so Mood's median test, a non-parametric test of centrality, was used to compare the medians of the two forms, revealing that there was a significant difference between them (see Table 6.9, "Initial Model").

As is evident from both the descriptive statistics and the median test, those who sat the ODD form, and who therefore encountered Items 1 and 7 in the video format, inadvertently sat a test which featured two items that were considerably easier than the versions encountered by those who sat the EVEN form. Removing only Item 1 from the Rasch model corrects this, restoring equivalence of the forms; however, out of an abundance of caution, Item 7 was also removed from the model specification and it was calibrated again. This reduced the total number of implicit items from ten to eight, but does not impact the feasibility of investigating the interaction between format and type, as they are parameterized as facets of measurement.

Descriptive statistics and the results of the median test between the ODD and EVEN forms in the corrected model are displayed in the "Corrected Model" section of Table 6.9, and demonstrate equivalence of these corrected forms. Rasch summary statistics for the corrected model are displayed in Table 6.10, and calibration of the corrected item facet can be seen in Table 6.11. The removal of Items 1 and 7 had little effect on the model fit of the items or the reliability of their estimates, and lowered the average difficulty of the implicit items from −0.29 to −0.37, likely due to the fact that the two items removed, especially Item 1, were substantially more difficult when delivered without the video. The reliability of person separation dropped with the loss of the two items from 0.71 to 0.69, but none of the other statistics were much impacted. The TIF for the corrected model is presented in Figure 6.8, displaying very slightly less information at all levels, which is to be expected as the total information is the sum of item information, and fewer items will always render less information (Baker, 2001). The implications of this will be further discussed in the Discussion section.

Table 6.9.

Descriptive Statistics and Mood's Median Test of the Test Forms

							Mood's Median Test*		
Model	**Form**	***n***	**Min.**	**Median**	***IQR***	**Max.**	χ^2	*df*	Asymp. Sig.
Initial	ODD	146	−1.97	0.71	1.60	4.55	5.774	1	0.016
	EVEN	133	−1.97	0.22	1.23	4.55			
Corrected	ODD	146	−1.86	0.52	1.56	4.48	0.278	1	0.598
	EVEN	133	−1.86	0.52	1.40	4.48			

* with Yates' continuity correction

Table 6.10.

Rasch Summary Statistics for the Item Facet and Task Types Under the Corrected Model

		Measures		Infit *MS*		Outfit *MS*			Reliability of			
Facet	***N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	**Strata**	**separation**	χ^2	***df***	***p***
Persons	279	0.63	1.20	1.00	0.20	1.03	0.43	2.34	0.69	692.1	278	0.00
Items	18	0.00	0.91	0.99	0.09	1.03	0.16	8.41	0.97	632.1	17	0.00
Explicit	10	0.29	0.98	0.98	0.09	1.03	0.16					
Implicit	8	−0.37	0.79	1.01	0.10	1.04	0.18					

Table 6.11.

Calibration of the Item Facet Under the Corrected Model

Item	**Name**	**Type**	**Measure**	***SE***	**Infit *MS***	**Outfit *MS***
2	02.1e	Explicit	0.86	0.14	0.93	0.95
3	03.2e	Explicit	1.26	0.15	1.12	1.19
4	04.3i	Implicit	1.04	0.14	1.10	1.20
5	05.3e	Explicit	1.38	0.14	0.87	0.87
6	06.3e	Explicit	0.37	0.13	0.95	0.91
8	08.2i	Implicit	0.13	0.15	1.15	1.14
9	09.2i	Implicit	-0.48	0.14	0.96	1.10
10	11.4e	Explicit	-0.85	0.15	1.01	1.12
11	13.1e	Explicit	-0.13	0.14	0.94	0.95
12	14.4e	Explicit	0.01	0.14	0.86	0.79
13	15.1e	Explicit	-0.88	0.15	0.99	1.26
14	16.1i	Implicit	-0.15	0.15	1.11	1.29
15	17.1i	Implicit	-0.16	0.14	1.00	1.02
16	18.1i	Implicit	-0.67	0.15	0.92	0.84
17	19.3e	Explicit	-0.8	0.15	1.04	1.01
18	22.2i	Implicit	-1.49	0.18	0.92	0.99
19	23.2i	Implicit	-1.18	0.16	0.89	0.75
20	24.2e	Explicit	1.72	0.16	1.09	1.24

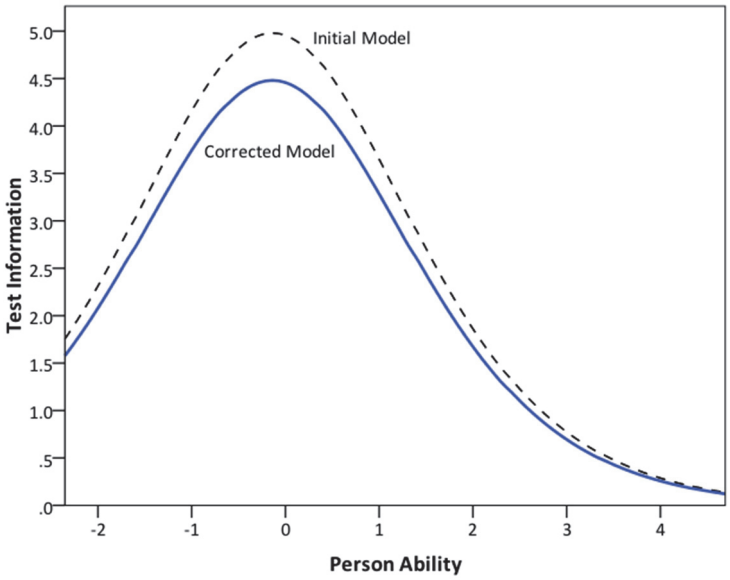

Figure 6.8. Test information function for the corrected model. The horizontal axis is scaled to the person measures observed (−1.86 to 4.48).

6.3.3.4. Corrected format interaction with task type

Using the corrected model, the interaction of format with task type was analyzed again. The results can be seen in Table 6.12, and are visualized in Figure 6.9. Even with the very large format effect observed in Item 1 and the less-pronounced effect of Item 7 removed, implicit items were still over one logit easier (1.02 logits) when presented in the video format, while the explicit items are unaffected by the change in the model. Implicit items are more difficult than explicit items in the audio format, but are easier than explicit items when presented with video. Both contrasts remain statistically significant with extremely low *p* values.

Table 6.12.

Pairwise Bias Report for Format versus Task Type with the Corrected Model

	Audio		Video						
Type	Measure	*SE*	Measure	*SE*	**Contrast**	***SE***	***t***	***df***	***p***
Explicit	0.28	0.06	−0.28	0.06	0.56	0.09	6.15	2724	0.000
Implicit	0.47	0.07	−0.55	0.08	1.02	0.11	9.64	2167	0.000

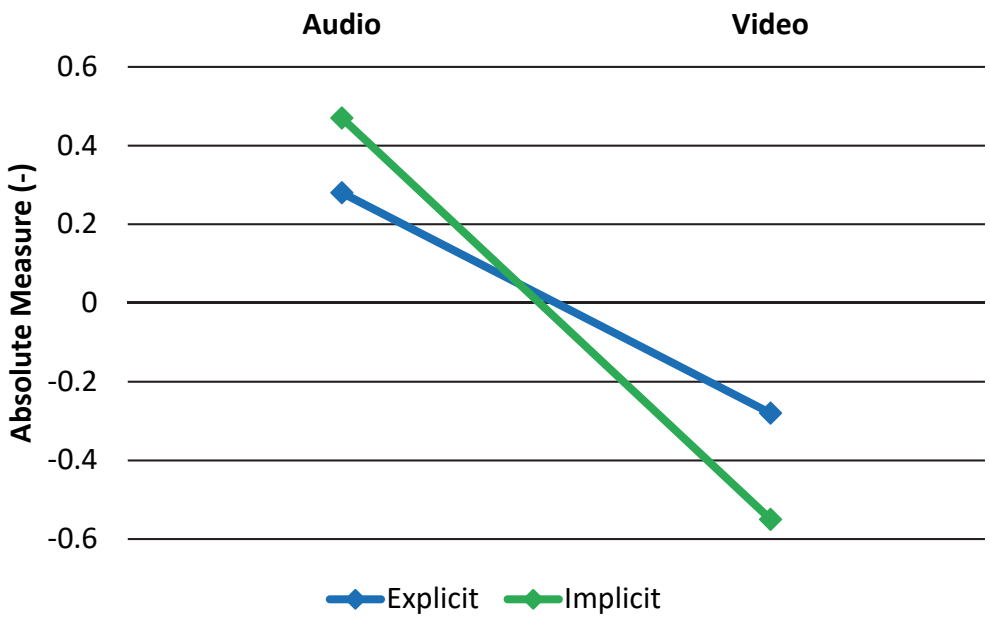

Figure 6.9. Comparison of average task type difficulty estimates under audio and video format conditions. Higher values are more difficult.

6.3.4. Individual Differences

To investigate RQ3a ("How do individual examinee differences interact with the presence of visual cues on video-mediated listening comprehension tests"), and RQ3b ("How do examinee perceptions of video interact with performance on video-mediated listening comprehension tests?"), proficiency, gender, and survey responses were compared to individual contrast values for the formats.

6.3.4.1. Proficiency

The interaction between proficiency and the facilitative effect of video was investigated via scatterplots comparing the external TOEFL ITP scores and person ability estimates from the test in question to mean contrast values (Figure 6.10 and Figure 6.11, respectively), revealing a high degree of variation in the facilitative effect of video at virtually all levels of proficiency. Despite the apparent lack of association between the variables, Pearson product-moment correlations were also calculated,

further demonstrating the lack of association between proficiency and facilitative effect of the video format (see Table 6.13).

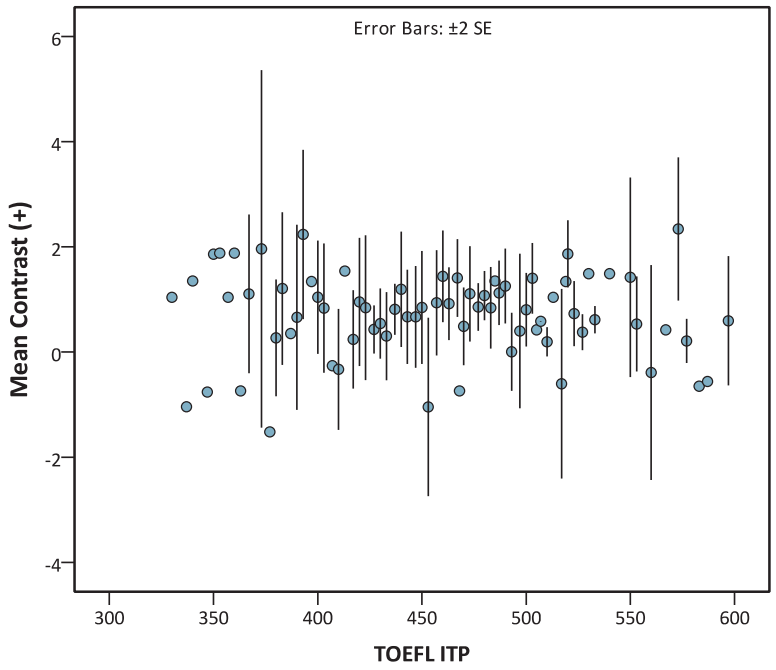

Figure 6.10. Mean contrast by TOEFL ITP score.

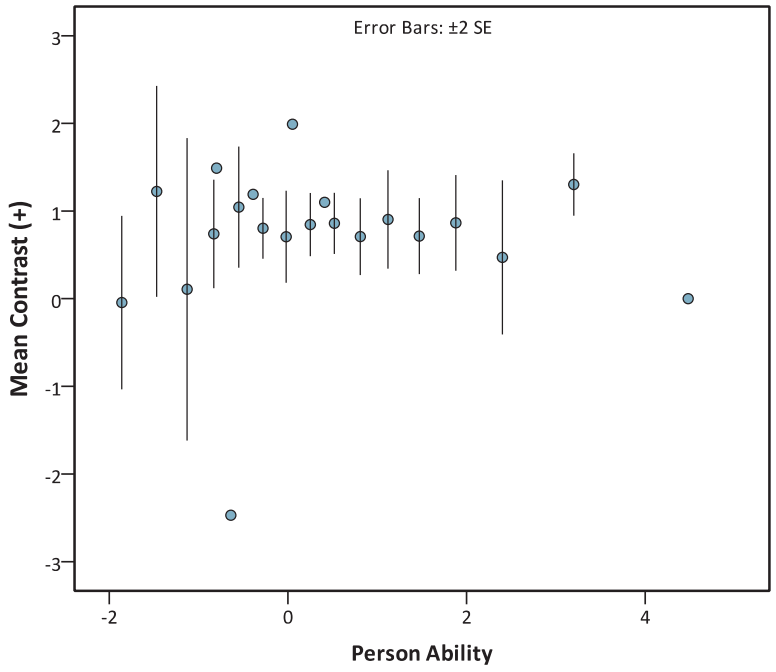

Figure 6.11. Mean contrast by person ability estimate.

Table 6.13.

Pearson Product-Moment Correlations between Proficiency Measures and Contrast Values

	Contrast	*p* (2-tailed)
TOEFL ITP	-0.023	0.703
Person Ability	-0.007	0.902

6.3.4.2. Gender

Contrast values for the genders were first checked with Shapiro-Wilk tests, finding no significant (<0.05) departures from Normality. Equality of variances was confirmed with a Levene's test. Having determined that the gender data violated no assumptions, a two-sample *t*-test of the contrast scores for the genders was completed. Descriptive statistics and results of the *t*-test for the genders' contrast values can be seen in Table 6.14. Although females benefited slightly more than men from the inclusion of video, standard deviations for average contrast values for both gender were quite large in comparison to the means, indicating a high degree of variability (see Figure 6.12 for a graphical representation of the comparative distributions). The difference between the genders in terms of the facilitative effect of video was non-significant.

Table 6.14.

Descriptive Statistics and Two-Sample t*-Test of Contrasts by Gender*

Gender	*N*	Mean Contrast	*SD*	*t*	*df*	*p*
M	168	0.75	1.21	−0.528	277	0.598
F	111	0.83	1.12			

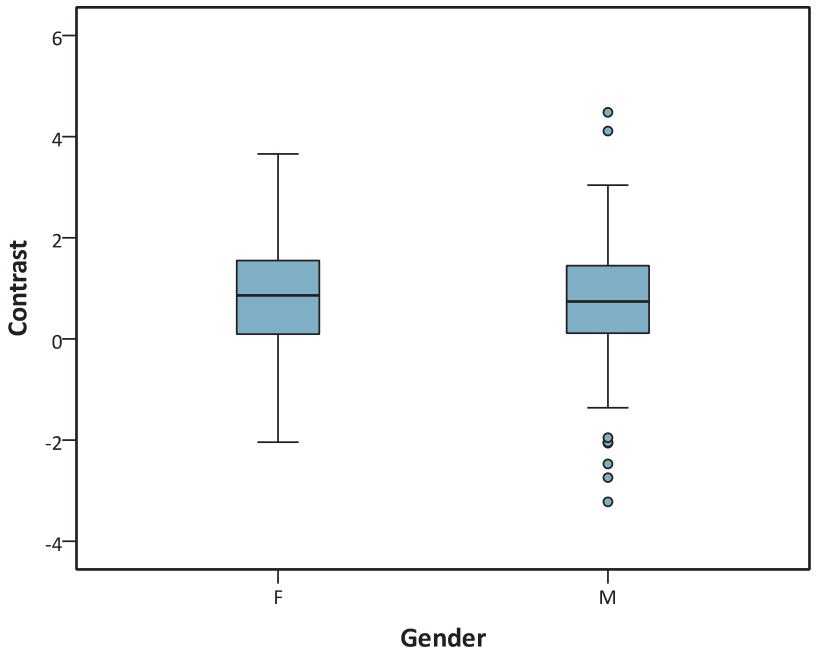

Figure 6.12. Boxplots of contrast distributions by gender.

6.3.5. Examinee Perceptions

This section reports the results of the post-test survey on perceptions of the two formats. Percentages of the selected answers can be seen in Table 6.15. Overall, examinees felt that the video format was easier, but that the audio format was a better test of their listening abilities.

Table 6.15.

Examinee Perception Survey Response Percentages

Question	Audio	Video	Same
Which question format was harder?	63%	11%	27%
Which question format do you think tested your listening ability better?	62%	20%	19%
On which question format do you think you scored higher?	10%	72%	18%
	Not at all	**A little**	**A lot**
How much did the videos help you understand the contents of the dialogs?	4%	54%	43%

Answers to the four survey questions were compared to the biased audio and video estimates, person ability estimates, and contrast values for the respondents. Each survey question will be addressed separately below.

6.3.5.1. Which question format was harder?

The first question asked respondents to identify which of the two formats was more difficult, or if they were equivalent. Descriptive statistics for the answers can be seen in Table 6.16, and are also represented graphically in Figure 6.13. Figure 6.13 includes not only the means, but bars representing two standard errors to indicate the spread of scores, which hinders interpretation in some cases. The majority of respondents identified the audio format as being more difficult than the video format, followed by those who claimed they were equal, and the remainder identifying video as more difficult. Ironically, those who identified video as the harder format benefitted most from it. Those who chose "Same" have the highest average ability estimates, while those who chose "Audio" have the lowest. Those who chose "Same" also experienced the least facilitative effect of the video format, but still performed better on those items. However, it is important to note that standard deviations are all larger than the means, indicating a great deal of variability among the sub-populations. One-way ANOVAs of the mean biased audio and video item measures, ability measures, and contrast values by the question responses revealed no significant differences between groups (See Table 6.17).

Table 6.16.

Descriptive Statistics for the Survey Question, "Which question format was harder?"

			Audio Meas.		**Video Meas.**		**Person Ability**		**Contrast**	
	N	***%N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Audio	175	63%	0.16	1.21	0.97	1.34	0.54	1.13	0.81	1.06
Video	30	11%	0.26	1.60	1.15	1.61	0.67	1.38	0.89	1.46
Same	74	27%	0.50	1.59	1.15	1.33	0.84	1.29	0.65	1.32

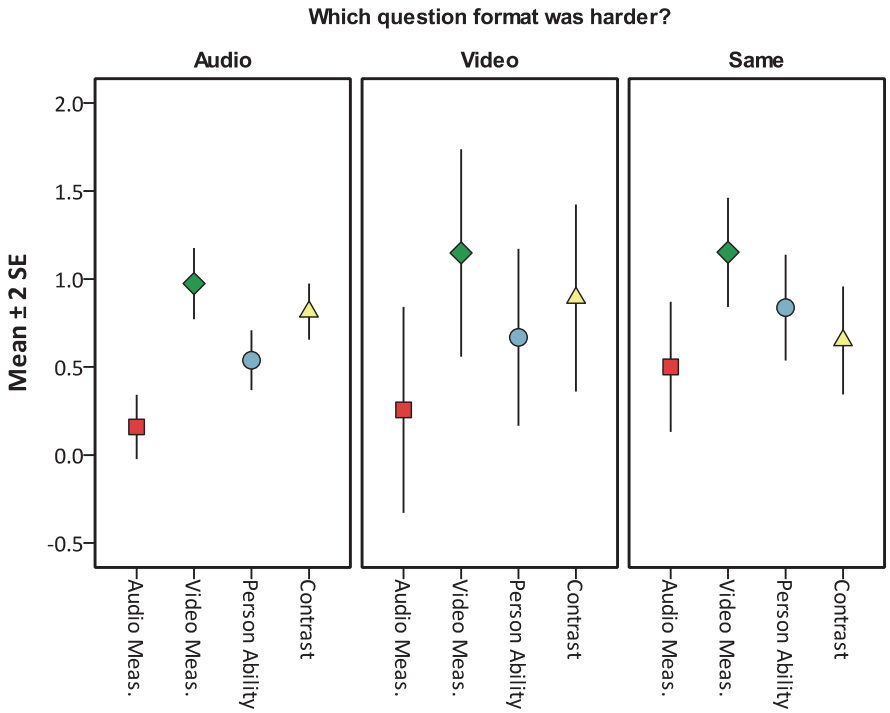

Figure 6.13. Graph comparing mean biased audio and video item measures, ability measures, and contrast values for the three possible answers to the survey question, "Which question format was harder?"

Table 6.17.

ANOVAs of Mean Biased Audio and Video Item Measures, Ability Measures, and Contrast Values by Responses to the Survey Question, "Which question format was harder?"

		SS	*df*	*MS*	*F*	*p*	$\eta^2_{partial}$
Audio Meas.	Between Groups	6.05	2	3.03	1.63	0.199	0.012
	Within Groups	513.97	276	1.86			
	Total	520.02	278				
Video Meas.	Between Groups	2.04	2	1.02	0.54	0.581	0.004
	Within Groups	517.63	276	1.88			
	Total	519.67	278				
Person Ability	Between Groups	4.69	2	2.34	1.63	0.198	0.012
	Within Groups	397.37	276	1.44			
	Total	402.06	278				
Contrast	Between Groups	1.81	2	0.91	0.65	0.522	0.005
	Within Groups	383.17	276	1.39			
	Total	384.98	278				

6.3.5.2. Which question format do you think tested your listening ability better?

The same analyses employed with the first survey question were applied to the second. See Table 6.18 and Figure 6.14. The majority (62%) of respondents believed that the audio format was a better test of their listening abilities, with the remainder of the sample split fairly evenly between the two other options. Those who reported that video was a better test of their listening abilities enjoyed the largest facilitative effect of the video format, while those who answered that the formats were the same witnessed very little contrast at all (0.37 logits). Standard deviations for the mean contrasts and abilities were once again quite large in comparison to the means. The person ability estimates for the three groups were extremely close to one another. This is further indicated by the very high *p* value (0.915) on the ANOVA of person estimates by groups (see Table 6.19), but a significant difference among the answer groups in terms of contrast was observed, although its effect size is extremely small ($\eta^2_{partial} = 0.033$). Post-hoc tests (Table 6.20) revealed that the contrast values for those who answered this question with "Same" are sufficiently low to differ significantly from both other categories.

Table 6.18.

Descriptive Statistics for the Survey Question, "Which question format do you think tested your listening ability better?"

			Audio Meas.		Video Meas.		Person Ability		Contrast	
	N	***%N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Audio	172	62%	0.28	1.25	1.09	1.30	0.65	1.12	0.82	1.13
Video	55	20%	0.05	1.60	1.10	1.48	0.58	1.40	1.05	1.13
Same	52	19%	0.43	1.47	0.80	1.45	0.61	1.27	0.37	1.28

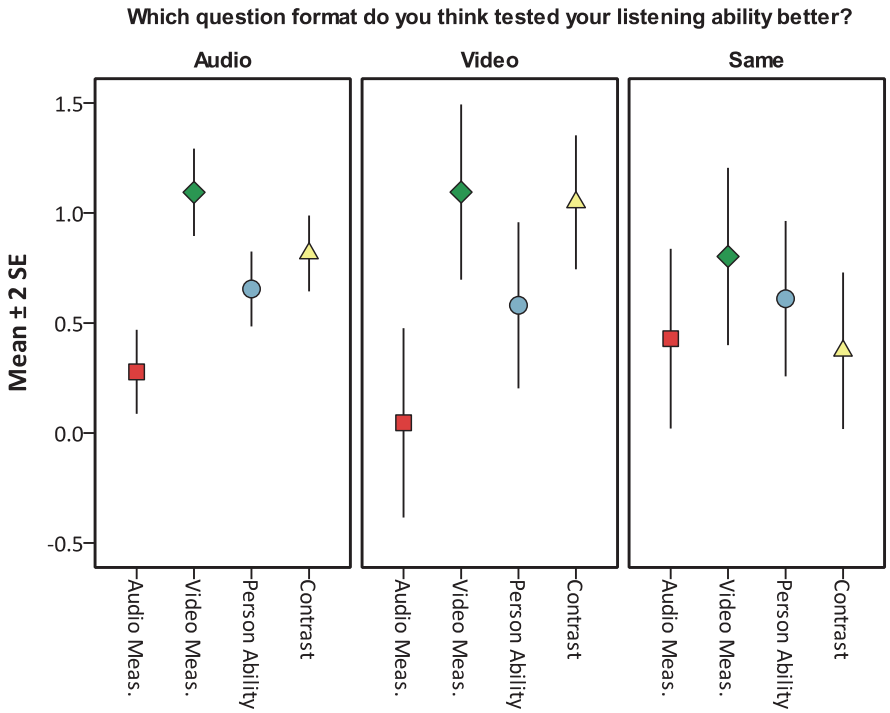

Figure 6.14. Graph comparing mean biased audio and video item measures, ability measures, and contrast values for the three possible answers to the survey question, "Which question format do you think tested your listening ability better?"

Table 6.19.

ANOVAs of Mean Biased Audio and Video Item Measures, Ability Measures, and Contrast Values by Responses to the Survey Question, "Which question format do you think tested your listening ability better?"

		SS	*df*	*MS*	*F*	*p*	$\eta^2_{partial}$
Audio Meas.	Between Groups	4.05	2	2.02	1.08	0.340	0.008
	Within Groups	515.97	276	1.87			
	Total	520.02	278				
Video Meas.	Between Groups	3.60	2	1.80	0.96	0.383	0.007
	Within Groups	516.07	276	1.87			
	Total	519.67	278				
Person Ability	Between Groups	0.26	2	0.13	0.09	0.915	0.001
	Within Groups	401.80	276	1.46			
	Total	402.06	278				
Contrast	Between Groups	12.77	2	6.39	4.74	0.010	0.033
	Within Groups	372.21	276	1.35			
	Total	384.98	278				

Table 6.20.

Significant Results of LSD Post-Hoc Multiple Comparisons of Contrast Values by Response to the Survey Question, "Which question format do you think tested your listening ability better?"

Dependent Variable	Response Group	Mean Difference	*SE*	*p*
Contrast	Same – Audio	−0.442	0.184	0.017
	Same – Video	−0.675	0.225	0.003

6.3.5.3. On which question format do you think you scored higher?

The same analyses were applied to the responses to the third question. Descriptive statistics are available in Table 6.21, a graphical representation is available in Figure 6.15, and the results of ANOVAs on contrast and ability by question response can be found in Table 6.22. A large majority (72%) of the sample responded that they thought they had scored higher on the video items, followed by those who responded "Same" at 18% of the sample. The mean contrasts for the three groups were very similar, but those who believed that they had scored higher on the audio items had substantially higher ability estimates than the other two groups. Once again, however, the standard deviations were quite large in comparison to the means, indicating a wide distribution of values. Despite the large ability difference of the audio respondents, however, no statistically-significant differences between groups on either contrast or ability estimates were observed.

Table 6.21.

Descriptive Statistics for the Survey Question, "On which question format do you think you scored higher?"

			Audio Meas.		Video Meas.		Person Ability		Contrast	
	N	***%N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Audio	27	10%	0.81	1.69	1.57	1.69	1.15	1.55	0.77	1.20
Video	201	72%	0.18	1.21	0.98	1.31	0.56	1.11	0.80	1.10
Same	51	18%	0.28	1.71	0.98	1.36	0.63	1.32	0.70	1.44

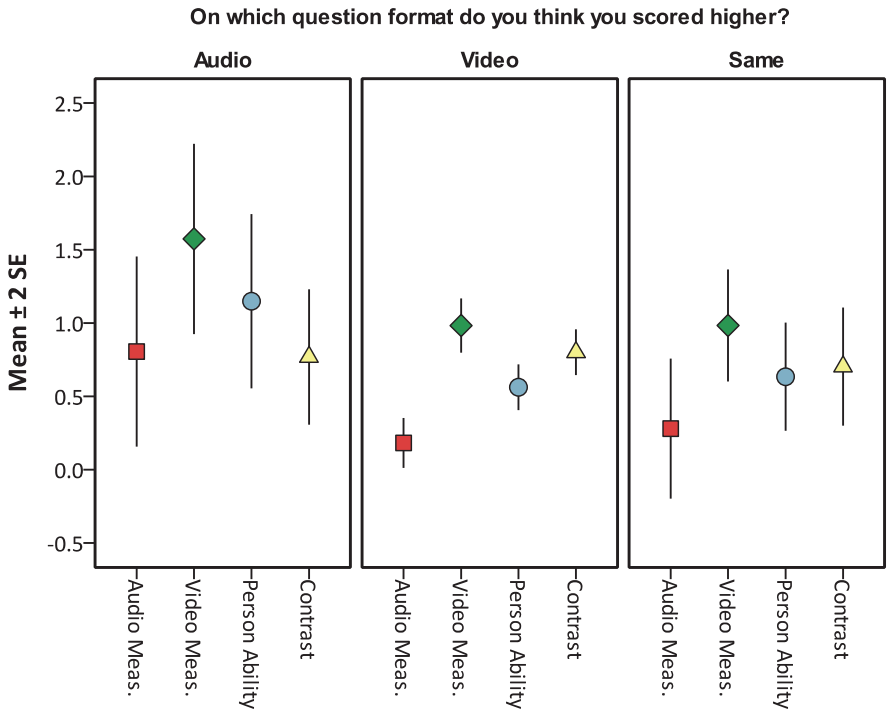

Figure 6.15. Graph comparing mean biased audio and video item measures, ability measures, and contrast values for the three possible answers to the survey question, "On which question format do you think you scored higher?"

Table 6.22.

ANOVAs of Mean Biased Audio and Video Item Measures, Ability Measures, and Contrast Values by Responses to the Survey Question, "On which question format do you think you scored higher?"

		SS	*df*	*MS*	*F*	*p*	$\eta^2_{partial}$
Audio Meas.	Between Groups	9.27	2	4.64	2.51	0.084	0.018
	Within Groups	510.75	276	1.85			
	Total	520.02	278				
Video Meas.	Between Groups	8.51	2	4.26	2.30	0.102	0.016
	Within Groups	511.16	276	1.85			
	Total	519.67	278				
Person Ability	Between Groups	8.20	2	4.10	2.87	0.058	0.020
	Within Groups	393.86	276	1.43			
	Total	402.06	278				
Contrast	Between Groups	0.39	2	0.20	0.14	0.869	0.001
	Within Groups	384.59	276	1.39			
	Total	384.98	278				

6.3.5.4. How much did the videos help you understand the contents of the dialogs?

The analyses were applied to the final survey question, asking respondents how much they believed the video format aided their comprehension. Descriptive statistics can be found in Table 6.23, a visualization in Figure 6.16, and results of one-way ANOVAs in Table 6.24. Just over half of the sample (54%) answered that it had helped "a little," while only 4% felt that it had not helped at all. There was no difference in average contrast between those who answered "not at all" or "a little," and only a small increase for those who claimed it had helped "a lot." The person ability estimates for the 10 respondents who claimed that the video had not helped at all were quite low (−0.04 logits), whereas the highest mean ability estimates were found in the "a little" responses. However, it is important to note that, once again, standard deviations are quite large in comparison to means, which makes confident judgments of difference difficult. No statistically-significant differences in the three groups of respondents' contrast values or ability estimates were observed.

Table 6.23.

Descriptive Statistics for the Survey Question, "How much did the videos help you understand the contents of the dialogs?"

			Audio Meas.		**Video Meas.**		**Person Ability**		**Contrast**	
	N	***%N***	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Not at all	10	4%	-0.40	0.99	0.30	1.13	-0.04	0.78	0.70	1.18
A little	150	54%	0.41	1.50	1.11	1.39	0.75	1.29	0.70	1.20
A lot	119	43%	0.13	1.19	1.02	1.34	0.55	1.10	0.89	1.15

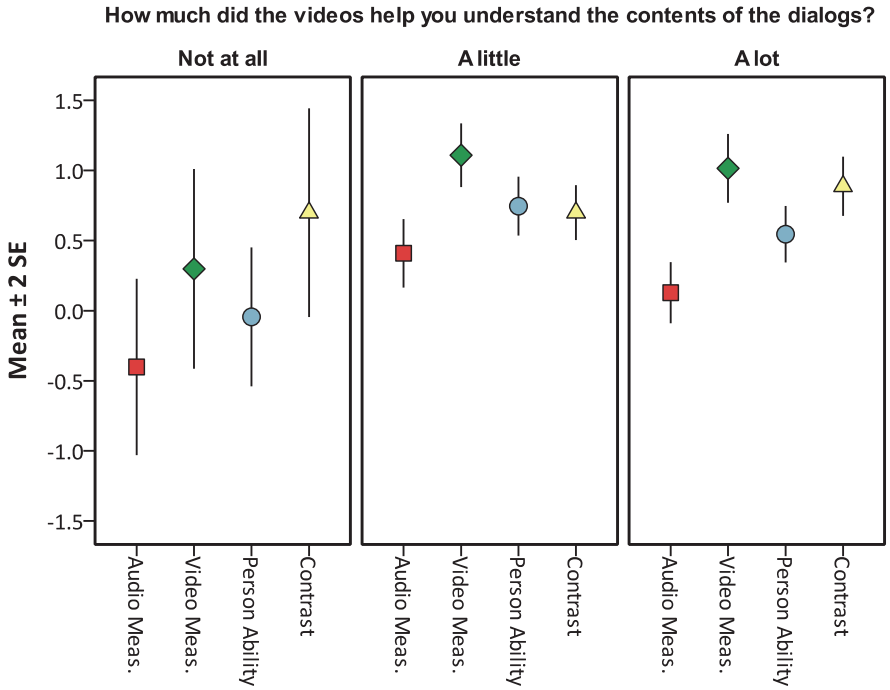

Figure 6.16. Graph comparing mean biased audio and video item measures, ability measures, and contrast values for the three possible answers to the survey question, "How much did the videos help you understand the contents of the dialogs?"

Table 6.24.

ANOVAs of Mean Biased Audio and Video Item Measures, Ability Measures, and Contrast Values by Responses to the Survey Question, "How much did the videos help you understand the contents of the dialogs?"

		SS	*df*	*MS*	*F*	*p*	$\eta^2_{partial}$
Audio Meas.	Between Groups	9.80	2	4.90	2.65	0.072	0.019
	Within Groups	510.22	276	1.85			
	Total	520.02	278				
Video Meas.	Between Groups	6.30	2	3.15	1.69	0.186	0.012
	Within Groups	513.37	276	1.86			
	Total	519.67	278				
Person Ability	Between Groups	7.39	2	3.70	2.59	0.077	0.018
	Within Groups	394.67	276	1.43			
	Total	402.06	278				
Contrast	Between Groups	2.40	2	1.20	0.87	0.421	0.006
	Within Groups	382.58	276	1.39			
	Total	384.98	278				

The preceding section reported the results of Study II. The following section will review, discuss, and interpret these findings.

6.4. Discussion of Study II

This section begins by addressing the overall evaluation of the quality of the instrument used in Study II. It then discusses the comparative difficulty of the audio and video formats as they pertain both to the individual items, but also to the two task types (explicit and implicit) investigated in this study. It closes with a test-focused discussion and a closer examination of the individual examinee differences and their interaction with the formats.

6.4.1. Evaluation of the Instrument

This section evaluates the quality of the instrument for the purpose of drawing conclusions regarding the relative difficulty of the formats and their interactions with task types. The person reliability of the instrument is discussed first, followed by a discussion of the item characteristics.

6.4.1.1. Reliability of the instrument

Overall, the reliability of the instrument was acceptable for the purposes of the present research. The reliability of the instrument is indicated by the reliability of the separation of the person measures, which was 0.69 in the corrected model (i.e., with Items 1 and 7 removed). It is important to note, however, that such metrics are test-length dependent, and given the relatively short length of 18 items in the corrected model, and the fact that there were no seriously misfitting items (discussed in detail below), these metrics of reliability are satisfactory (Alderson et al., 1995).

It would have been preferable, of course, to administer a longer test, but the logistics of this proved infeasible. Given the fact that the instrument was to be

administered to student volunteers in intact classes, time was at a premium. Class periods at Keio SFC are ninety minutes in length and only meet once a week (as is the norm at Japanese universities). Furthermore, the semester at Keio SFC is only fourteen weeks long, which includes an introductory session at the beginning of the semester, before the students have even officially registered, and (in the case of the English classes) any midterm or final tests. As such, most classes only have twelve full, regular class sessions. For this reason, it was extremely unlikely that any teachers would be willing to volunteer one-twelfth of his/her entire semester to a volunteer data collection. The first or last half of a single class period, however, was much easier to request from the teachers, and most who were asked saw it as an opportunity to work on a different language skill, basing an entire class session around it, usually with debriefing afterward. For this reason, great care was taken to design a test and administration procedure that reliably lasted no more than 45 minutes from setup to departure from the room. Furthermore, in the author's experience with other such data collections, 45 minutes seems to be the limit that a volunteer is willing to spend on such an activity, no matter how novel the task or how friendly the researcher. Even then, the participants' patience and interest seemed to wane, which is the likely reason for the fact that the last item on the test appears to be the most difficult. Finally, although it does not violate any campus policy to do so, the present author felt that it would be unethical to take up any more of the volunteer students' or teachers' time than one-half of one class session.

The design of the test also necessitated a rather short test, as the listening passages averaged a little over thirty seconds apiece, and each was accompanied by only one item. This decision was made, however, to ensure local independence of the items, as the present research is primarily concerned with their interaction with the

formats. This specification, however, came at an even greater time cost than most listening tests, which always suffer from the fixed temporal characteristics of the listening stimuli. Furthermore, the test design process, during which several item candidates for each passage were written, evaluated, and chosen to ensure equal coverage of formats and task types, necessitated slightly longer passages than were ultimately necessary for any one particular item. If the test had been audio-only, it may have been possible to cut unnecessary lines from the recordings in post-production; however, one of the challenges of video-based tests is the infeasibility of making such alterations after the material has been filmed. As such, the full length of each passage had to be used, even if some lines could have theoretically been cut. This decision, however, does provide a benefit: As multiple items have been prepared for each passage, future studies with different designs and different items remain possible, the most difficult and time-consuming process of video recording live actors having been already completed.

It is also worth pointing out that test reliability, within either the CTT or IRT traditions, assumes (or some would argue requires) strict unidimensionality of the test items. In the present study, this was deliberately manipulated by presenting each item in either the audio or video format, and the inclusion of both explicit and implicit items. The latter is conventional, of course—most tests of foreign language listening comprehension include at least these two broad task types—but the former is the main objective of the experiment. It is likely that this negatively impacted the model fit of the items, which damages the reliability of measures. If a longer test had been feasible, an anchoring section which was identical for both of the participant groups, and which comprised at least 20% of the total test length, would have allowed the

estimation of measures using one or the other format only, which would have made it possible to evaluate fit within only one or the other format conditions.

The TIFs of both the initial and the corrected models demonstrate that the test was most precise at just below zero logits of person ability, but that it provided information about all ability estimate levels present in the sample. It would have possibly been desirable for the test to feature several more difficult items, as was demonstrated by both the TIF as well as the Wright variable map. However, it is also important to note that, although care was taken to remove participants with the highest TOEFL ITP scores, or those whom the present author knew to have been raised in English-speaking countries, Keio SFC attracts an inordinately large number of bi- or multilingual students, and the sample therefore likely still included some very high-proficiency examinees, six of whom attained perfect scores. However, the final average person measure of 0.63 logits is very near the test's peak information, so although there could have been slightly better matching between the items and the persons at the higher end of the scale—which would have improved reliability of the person measures—the test was quite well-targeted to the population, even with two items removed.

6.4.1.2. Item characteristics

Overall, despite the issues discussed above pertaining to test length and dimensionality, the item statistics were very favorable. The range of difficulties, as indicated by difficulty estimates (−1.49 through 1.72), as well as the large number of strata (8.41) given the short length of the test, is adequate and the items appear to be fairly well-targeted to the person sample (a desire for a few more difficult items notwithstanding). The item fit is quite good throughout, conforming to the recommendations of the literature (e.g., Linacre, 2003; Smith et al., 1998; Wright &

Linacre, 1994), and the low standard deviations of the mean-square fit statistics indicate that the items' model fit is quite uniform. Given the experimental nature of the instrument, these results are encouraging.

One distinctive feature of the instrument is that the implicit items, which have frequently been found to be of greater difficulty than explicit in the literature (e.g., Bloomfield et al., 2010; Nissan et al., 1996; Shohamy & Inbar, 1991), were found to be easier in the present research (−0.37 versus 0.29 logits). However, these estimates include the facilitative effect of the video format in half of the cases, whereas the listening tests in the literature were audio-only. As such, the item difficulties as calculated without the video format bias are probably more comparable to those in the literature, and these difficulty estimates display the type of behavior seen elsewhere, with implicit items being more difficult than explicit (0.47 versus 0.28 logits).

6.4.2. Comparative Format Difficulty

As has often been found in the literature (e.g., Baltova, 1994, first study; Hernandez, 2004; Parry & Meredith, 1984; Shin, 1998; Sueyoshi & Hardison, 2005; Wagner, 2010b, 2013), the video format was generally easier than the audio format across both task types, with an average difference of 0.92 logits, with only three items being biased in the direction of video, and those biases being very small. In close examination of the items and the listening stimuli, some large differences were observed between the audio and video versions, but as in the previous Batty study (2015), many format-based difficulty differences are impossible to account for beyond mere conjecture.

One possible source of extra facilitative information in the video-format items is some manner of schema setting, providing contextual information at the outset, thereby freeing up time and cognitive resources for the process of comprehension. In

fact, it was first planned to include photos of the settings for the conversations in the audio-only format, but as most of the settings were extremely information-poor (e.g., a blue bench against a white wall), this was abandoned in favor of screencaps from the scenes. However, it simply was not possible to isolate any frames in most of the scenes which lacked any nonverbal communication cues. The presence of this information could either be facilitative or distracting, depending on whether the displayed nonverbal information supported or was at odds with the credited response. This would have contributed yet another possible source of score variation, which would be impossible to account or control for, so the photos were ultimately replaced by the speaker graphic presented in Chapter 5. This will be discussed further in the Conclusions chapter (Chapter 7).

Item 1 was the item with the clearest case of "unfair" influence in the video. However, if the video bias were not so apparent, there would have been nothing to indicate a problem with this item, its fit statistics being good overall. It was for this reason that the problem was not noticed until the very large contrast value of 2.33 was observed. Ultimately this item, along with Item 7, were removed from the final, corrected measurement model, due to concerns over whether the observed contrast was due to poor item design/production.

This issue, however, illustrates a central logistical challenge associated with video-mediated listening tests, which is not well-addressed by the literature. In the cases of these two items, the actors were merely plying their craft and acting in logical, realistic ways, given the script. It is entirely natural for someone to speak as he or she moves to complete an action, as the actress did at the end of the Item 1 scene, "Licorice." It was the author's responsibility as scene director to make note of the implications of this, and to request a retake of the scene. However, at the time of

filming, since the portrayal was realistic, its implication for the item went unnoticed. So, too, was the case in Item 7. Item writers are accustomed to carefully crafting the lines to be spoken for listening tests so as to provide opportunities for distractors without creating dialog that differs too much from authentic speech as to threaten construct validity. However, adding the silent, visual aspect to the test design process adds another layer of complexity. In the cases of both Items 1 and 7, silent visual cues crept into the final material that essentially disqualified the items from inclusion in the final analyses. From a logistical standpoint, the production of video for listening tests will always be much more involved than that of audio.

Item 4 also bears further discussion. It appeared that, for most examinees, the addition of the nonverbal cues of the man's facial expressions as he delivered his final lines were enough to eliminate any ambiguity the man's words may have contained. Clearly, when this item was presented with nonverbal cues, the examinees were able to draw upon their innate understanding of human nonverbal cues to infer the tacit intentions of the speaker. This is, of course, supported by the relevant psychological literature, as the visual channel is frequently used to reconcile ambiguity in the verbal channel (e.g., Burgoon, 1994; Folger & Woodall, 1982), and is more useful to affective, rather than cognitive, tasks (Noller, 1985). It is also supported by the viewing behavior observed in Study I, where facial expressions drew the most attentional orientation. A listener who had trouble reconciling the man's initial reaction (i.e., "That is very dishonest,") with his ambiguous follow-up ("But it just might work,") may have been able to augment his/her linguistic knowledge with the clear change in the man's expression between the two utterances. The question of whether this is construct-relevant, however, is one that remains open and unresolved, if not unresolvable. This will be discussed further in the Conclusions chapter.

Item 12 also gave rise to an interesting situation, in that it is an excellent example of when the presence of visual or nonverbal cues can serve to *distract* listeners, which is an issue that has been visited upon many times in the literature (e.g., Baltova, 1994; Coniam, 2001; Cubilo & Winke, 2013; Ockey, 2007; Sueyoshi & Hardison, 2005; Suvorov, 2013). Those who watched the video were much more likely to incorrectly believe that the reason the woman had not paid for her registration online was simply that she did not realize that the onsite price was higher, despite the fact that she complained about the website twice in the short dialog. However, in the video, it is clear that she is only mildly irritated when she explains that she couldn't pay online, and does not become angry until she hears the onsite rate, which is $30 more. Although the question was not "What was the woman angriest about?" it seems likely that it was this question that those who watched the video answered, as it was this point that provoked the strongest emotional response. This interpretation of the error is bolstered by the facts that both the question and the answers were presented in the examinees' L1, and that they previewed the question before seeing the video, eliminating the possibility that they simply misunderstood the stem.

Item 15, "Whose Number is This?" likely displays the opposite problem. This item was extremely easy for those who saw the video, but fairly difficult for those who did not. The scene opened with the pair gently holding hands atop a table at a café. The woman asks if the man has the number for the restaurant where they are to have their anniversary dinner. He hands her the phone and tells her that it is in the call history. This is when the wife discovers the unknown number. Rather than letting go of the man's hand, though, she grips it tighter, as though to prevent him from escaping, as she confronts him about the number, finally thrusting the phone in his

face and shouting. The strength of the nonverbal cues in this scene—the clasped hands, the expressions of anger and fear—may have been enough for even those who did not understand a word of the dialog to correctly guess that the topic of the discussion was infidelity, as male infidelity is a cause of inter-gender strife the world over. Once again, the question of construct validity is raised by this item. Is the ability to recognize pan-cultural nonverbal signals a worthy addition to the construct of foreign language listening comprehension, or is it better understood as error?

6.4.3. Interaction Between Format and Task Type

Even after two of the implicit items exhibiting the largest video biases were removed, implicit items were found to be 1.02 logits easier with video than with audio-only, and explicit items, 0.56 logits easier. Just as Ockey (2007) and Batty (2015) suspected, items which require the examinee to read into social contexts, resolving ambiguity in spoken language, are facilitated to a much greater degree by video than items that attempt to tap a purely linguistic construct. This was also observed to a certain extent by Wagner (2006), although his test and item format were not directly comparable to those of the present study. It seems that the presence of visual information, especially that of facial expressions (drawing from the findings of Study I) enhances the aural input sufficiently to allow many of those who would be unlikely to answer the same item correctly without it to select the credited response in both explicit and implicit multiple choice items.

As was briefly discussed previously, the average difficulty of the implicit items in the present study was lower than that of the explicit items, in contrast to much research on task types in such tests (e.g., Bloomfield et al., 2010; Nissan et al., 1996; Shohamy & Inbar, 1991). However, these estimates included both the audio and video versions of these items. When examining the biased estimates, it becomes clear that

the implicit items are indeed more difficult than the explicit items when presented in the audio format, but become easier than the explicit items when presented with video. That is to say, it appears that the addition of video *reverses* the traditionally-expected order of difficulty between explicit and implicit items. This finding has serious consequences for item design. If implicit items are written with the expectation that they will be the harder items on the test, but the test is delivered with video, tests may be mis-targeted to the population.

Moreover, this finding raises further questions regarding the construct-relevance of scores on video-mediated tests such as used in the present research. Does a correct response to such an implicit item without access to the nonverbal behavior via video mean the same thing as one that resulted from a testing session that included it? How does this impact judgements of examinees' language ability? Can nonverbal cues that are shared by all humans, the blind, and even other primates rightly be considered a component of language ability? —Or should the construct definition be widened to accommodate such ability? These questions will be addressed in greater detail in the Conclusions chapter (Chapter 7).

6.4.4. Individual Differences

This section discusses the interaction between individual examinee differences and the audio and video formats. Proficiency, gender, and examinee perceptions as indicated by responses to the four survey questions that followed the test will be addressed.

6.4.4.1. Proficiency

Some research has found more of a facilitative effect from video for those with lower proficiency (e.g., Parry & Meredith, 1984; Sueyoshi & Hardison, 2005). For this reason, the interaction between both an external measure of proficiency (TOEFL ITP scores) as well as the ability estimates from the present research were compared

with the size of the contrast between the audio- and video-biased measures, detecting no pattern whatsoever. This, too, is an encouraging result, as the present author has long suspected that video-mediated listening tests introduced a floor effect on scores. If that had been the case, larger contrasts would have been observed at the lower ends of the proficiency measure distributions, along with significant negative correlations between proficiency and contrast size. Absent this evidence, however, it is possible to conclude that video-mediated listening tests do not unfairly advantage the less-proficient, at least at the test level.

6.4.4.2. Gender

As many studies of sensitivity to nonverbal behavior have found women to be more perceptive of this information than men (e.g., Burgoon, 1994; Burgoon et al., 2016; Costanzo & Archer, 1989; Noller, 1985; Rosenthal et al., 1979), the interaction between gender and size of contrast between audio-biased and video-biased ability estimates was investigated. It was revealed that women indeed did experience a slightly larger benefit from the inclusion of visual and nonverbal communicative cues than men, on average, but that this difference was non-significant, with a very large *p* value of approximately 0.60, placing the likelihood of observing such a mean difference if there truly was none at 60%. The men were much more widely distributed in the difference between their audio- and video-biased measures than the women were, but the overall effect was essentially identical.

Despite the seeming anticlimactic nature of this observation, it is nonetheless a valuable finding, as it may eliminate any concern one may have about gender-based differential item functioning (DIF) of video-mediated listening tests. If a significant difference had been observed, such tests may have unfairly benefitted women, and damaged the construct relevance of the scores.

6.4.5. Examinee Perceptions

It would have been ideal to locate somewhat cohesive groups among the respondents based on their responses to the four questions, but attempts at cluster analysis of the responses were unsatisfactory, requiring more clusters than could be meaningfully interpreted to attain useful levels of cohesion and separation. Clearly, individuals differ a great deal in terms of their perceptions of the formats. Although the survey responses were compared with audio- and video-biased measures, ability estimates, and contrast measures, it is important to once again note that distributions tended to be so wide as to threaten interpretability, but some very general patterns do seem to emerge.

6.4.5.1. Perception of comparative difficulty

Responses to the first question, "Which question format was harder," demonstrated that those with lower overall ability found the audio items harder, implying a preference for video. This was also the most frequent response by a wide margin. These people were indeed helped quite a bit by the video, with an average contrast of 0.81 logits. Those who reported that there was no difference between the formats, however, tended to be higher-ability respondents, with an average person ability 0.17 logits higher than the next-most-proficient group. These respondents do not appear to have been helped much by the video format.

These results, taken with the finding that proficiency and contrast are not linearly associated (various non-linear methods were similarly explored, without success), indicate that if there is a relationship between proficiency and the effect of the video format on scores, it may be at an even smaller scale than proficiency measures, incorporating personality differences or preferences. However, again, it is

important to remind the reader that the distributions are so wide on any of the examined variables that if there is a signal lurking in the noise, it is very faint indeed.

6.4.5.2. Perception of comparative face validity

Responses to the question, "Which question format do you think tested your listening ability better," can be understood to indicate the respondents' view on which format was more likely to allow them to accurately demonstrate their ability. Those who claimed that video was the better test tended to have a much higher contrast between their audio- and video-biased measures, which is in line with the findings of Suvorov (2009). It is possible that this is due to a perception that the video format was more realistic, allowing them to demonstrate the skills they felt they possessed in the real world. However, once again, wide distributions threaten interpretability.

Those who answered that the formats were equally capable of assessing their abilities were essentially correct, given the small contrast values observed in this group. This group has a lower average person ability mean, however, than those who claimed that the formats were equal in difficulty in the first survey question. The low contrast values associated with this group are the source of the only significant difference between the three groups on the entire survey. Clearly, these respondents were fairly cognizant of their performance on the two formats.

Interestingly, although most identified audio as the more difficult format in the first question, an almost-equal number felt that it was a better test of their listening ability. It is possible that despite the respondents' awareness of the lower difficulty of the video items, they understood this facility to be the result of extralinguistic knowledge and skill, realizing that they were relying upon the facial expressions and other nonverbal cues of the actors to fill in the gaps in their comprehension, and felt

that this was illegitimate. If this is the case, it raises the question of whether the video format truly tests better, or is simply easier.

6.4.5.3. Perception of comparative performance

The third survey question asked respondents to identify on which format they believed they had scored higher. Overwhelmingly, they indicated that it was the video format. Once again, squaring this finding with the fact that the most people identified video as the easier format, and also as the format less-indicative of their listening ability, it may be necessary to conclude that the respondents—if not *preferred* video—at least felt that video was perhaps too easy; despite scoring higher than they did on the harder format, they nonetheless believe that the harder format is a better indication of their abilities.

Once again, it appears that those of higher proficiency associated themselves with the audio format, as the mean person ability of those who felt that they had performed better on the audio items was a quite-high 1.15 logits, although the standard deviation was larger than the mean. On the other hand, those with low ability estimates once again associated themselves with the video format. Regardless of the real size of the facilitative effect of video for those of varying proficiency, the perception of lower-proficiency people appears to be that the video format is essentially easier.

6.4.5.4. Perception of facilitative effect of video

Overall, it seems that respondents were aware of the degree to which the video format facilitated their item responses, as those who indicated that it had helped "a lot" had the highest contrast values. That being said, they were not much different from either of the other two groups of respondents. Very few people claimed no benefit from the video format, although the ability levels of those who responded as

such were quite low. The highest person estimates were found in the group claiming that the videos had helped "a little." It appears from these results that the overwhelming majority of respondents acknowledged the facilitative effect of the video format, but, once again, distributions prevent all but the most cursory of interpretations. Individuals do not fit into tidy groups in terms of their perceptions of the format differences, or their performances on them.

6.5. Chapter Summary

This chapter has reported on Study II, which sought to answer RQ2a, "How does the presence of visual cues interact with items on video-mediated listening comprehension tests," RQ2b, "How does the presence of visual cues interact with task types (explicit and implicit) to influence item responses on video-mediated listening comprehension tests," RQ3a, "How do individual examinee differences interact with the presence of visual cues on video-mediated listening comprehension tests," and RQ3b, "How do examinee perceptions of video interact with performance on video-mediated listening comprehension tests?" The study method was described, followed by a detailed explanation of the data analysis methods employed in the study. Next, the results were presented, followed finally by a discussion of said results. Overall, it was discovered that the presence of visual and nonverbal communicative cues has a fairly large (0.56 logits) average facilitative effect over the same items when presented as audio-only. This effect was even more pronounced with the implicit items, wherein the answers to the questions were not explicitly stated in the listening stimuli. Furthermore, implicit items were found to be more difficult, on average, than explicit items when presented with audio only, which is in line with the findings of other researchers into task type differences; however, when the same items were presented with video, they became easier than their explicit counterparts.

Finally, individual examinee differences and perceptions were examined with respect to the two formats, but results were largely inconclusive. Proficiency level, as determined by either incoming TOEFL ITP score or ability estimate, was found to be unrelated to the size of the video effect. Likewise, no difference between the genders in the size of the facilitative effect of video was observed. A significant difference in audio-video contrast size was observed among the responses to the question, "Which question format do you think tested your listening ability better," with those who felt that they were equivalent displaying the smallest contrasts. Although differences in scores and contrasts were observed among the groups of survey responses, no other questions yielded statistically-significant results.

The following chapter will summarize the findings of both Studies I and II, and discuss the implications of their combined findings in terms of construct validity, practicality, and recommendations for language assessors who may be considering the use of video in listening assessments. It will then briefly discuss the limitations of the studies herein, and close with recommendations for future research directions on this topic.

CHAPTER 7. CONCLUSIONS

This chapter presents the conclusions drawn from the preceding studies. It begins with a summary of the findings with regards to the three research questions, then moves to a discussion of the implications of these findings. The implication section first addresses the theoretical implications of the findings, followed by a discussion of the practical implications. Next, recommendations are made for language assessors who may be considering the video format for listening tests.

Following the recommendations section is a brief limitations section, which addresses methodological issues that arose during the studies. Many of these are reflected in the section immediately following, which outlines recommendations for further work on this topic. The chapter concludes with a brief summary and conclusion.

7.1. Summary of Findings

This section addresses each of the three research questions separately, reviewing the relevant findings for each.

7.1.1. RQ1a – 1c

As these three research questions are intimately related, they will be addressed together. The questions were as follows:

1a. *What are the specific nonverbal or visual cues to which L2 examinees attend when taking a video-mediated test of foreign language listening comprehension?*

1b. *How does viewing behavior of visual cues differ with respect to videotexts?*

1c. *How does viewing behavior of visual cues change with respect to task type (explicit and implicit)?*

RQ1a through RQ1c were investigated via eye-tracking methodology in Study I. As the twelve participants watched the six videotexts, each associated with a single item, it was observed that their attentional orientation (Posner et al., 1980) was directed at faces a large majority of the time (M = 81.23%, SD = 11.90%). Of that time, roughly 75% of the dwell time was oriented toward the speaker's face, in nearly equal parts the eyes, mouth, and scanning the speaker's face. Watching the listener's face was considerably less common, with a median (because watching the listener's face was fairly uncommon, the distribution was heavily skewed, necessitating the use of non-parametric descriptive statistics) of 5% of the time devoted to it, with the eyes making up the largest proportion of it. The most-common explanation for any facial watching was to determine the character's emotional state. However, this did not apply to watching of the mouth, which was most-commonly reported as a strategy to supplement comprehension, likely through lip-reading. Behavior wherein the participant scanned either the speaker's or listener's face occurred when examinees were listening to material in order to respond to an implicit question, likely due to the facilitative effect of understanding the characters' affective states when completing questions of this task type.

The behavior dubbed "Alternating between faces" by the present researcher, in which participants moved rapidly between the faces of the two characters on screen comprised almost 13% of the total dwell time, 15.5% of the facial watching, and occurred mostly during times when the participants were trying to track both characters' emotional states simultaneously, when it was not simply a waiting behavior at the beginning of scenes before anyone had begun speaking. Due to the more meaningful reason for this behavior, however, it was observed significantly

more in the context of implicit questions, which may have benefited from the added affective information.

Very little participant time was expended on watching the hands, whether they were engaged in illustrative gestures or emblematic. The test featured but one emblematic gesture (an "okay" hand signal), but only one participant oriented directly at it, and even she did not remember doing so. However, illustrative gestures (i.e., demonstrating or supporting the verbal channel with one's hands) were looked at significantly longer when responding to explicit items. This was also the case with objects, although it may have been an artifact of the particular items in the test, both of which referenced objects specifically ("What did the man buy?" and "What is the man looking for?").

The results of Study I demonstrated that the single most important visual information to test-takers on a video-mediated listening test, based upon what they consciously oriented toward and what they reported in interview, is affect displays (i.e., facial expressions). Although objects and the setting were viewed from time to time, the vast majority of the visual information that the examinees chose to attend to was related to determining the speaker's, and sometime the listener's, emotional or mental state, as a way to facilitate comprehension of the scene. It is certain that other information is also noticed in the parafoveal regions of vision, but the face is the visual cue that examinees find most informative. Given this finding, it can be stated that the presence of nonverbal behavior in the form of affect displays are the key difference between audio- and video-mediated listening tests. Other information is added with video as well, but its importance is dwarfed by facial expressions. This is what the video format adds; this is what the audio format lacks. The implications of this finding will be discussed in the Implications section.

7.1.2. RQ2a and 2b

Research questions 2a and 2b, like the three RQ1 sub-questions, will be treated simultaneously here, as items and task types are fundamentally related. They are as follows:

2a. How does the presence of visual cues interact with items on video-mediated listening comprehension tests?

2b. How does the presence of visual cues interact with task types (explicit and implicit) to influence item responses on video-mediated listening comprehension tests?

RQ2a and 2b were investigated in Study II by comparing equivalent audio- and video-mediated listening tests with many-facet Rasch modeling. The instrument was developed drawing on the findings of Study I, and included ten explicit items and ten implicit items. Participants were randomly placed into experimental groups based on birthdate, and had video-mediated items on every even-numbered item, or every odd.

Overall, the video format was found to be significantly easier than audio; however, a qualitative analysis of the items displaying the largest bias toward video revealed two whose videos included extra factual, visual information that may have reduced their difficulty for reasons unrelated to the research. In these cases, actors added single actions to the scenes which were not in the scripts, and the present researcher did not notice them until post-production or after administration. Unfortunately, both of these items fell on odd numbers, meaning those who sat the ODD form of the test (i.e., those who had video items on odd-numbered questions) inadvertently sat an easier test. In order to restore comparability, these two items were removed from the final, corrected model.

After removing the problematic items, the large contrast between audio and video difficulty estimates remained, but this contrast was not uniform throughout the items. explicit items were found to be an average of 0.56 logits easier when in the video format than audio, but implicit items were considerably easier still at 1.02 logits easier in the video condition. This was true despite the fact that the implicit items were more difficult than the explicit items in the audio condition, which has frequently been found in the literature. The implications of this finding will be discussed in the Implications section below.

7.1.3. RQ3a and 3b

Research Question 3a and 3b were focused on the examinees in video-mediated listening tests. Those questions were as follows:

3a. How do individual examinee differences interact with the presence of visual cues on video-mediated listening comprehension tests?

3b. How do examinee perceptions of video interact with performance on video-mediated listening comprehension tests?

In Study I, eye tracking was employed to address this question, however, although individuals tended to either focus on eyes or mouths, due largely to the sample size, no meaningful differences in viewing behavior were noted among different proficiency levels.

In Study II, RQ3a and 3b were investigated via ANOVA, correlation coefficients, and pairwise *t*-tests, but few significant interactions were observed. Neither gender nor proficiency were found to interact with the absence or presence of visual cues on the listening test. Grouping the respondents by their answers on a short survey of their perceptions of the formats and comparing their scores and contrast sizes was also largely inconclusive, although a significant difference in audio-video

contrast size was observed among the responses to the question, "Which question format do you think tested your listening ability better?" Those who felt that the two formats were equivalent in their ability to gauge listening skill were found to have significantly lower contrast values than those who identified either the audio or video formats as superior, indicating that for those who answered in this way, the test was equally difficult under the two format conditions. Overall, however, the facilitative effect of video does not seem to affect individuals differently in any systematic manner.

7.2. Significance of Contribution

This thesis has addressed critical gaps in the video listening test literature through the use of eye-tracking methodology and many-facet Rasch measurement, and expanded the knowledge on this topic several important ways.

Study I is the first study in the video listening literature to objectively track examinee visual attention in video listening tests. This has resulted in the following new findings:

- Examinees spend most of their time watching faces for the purpose of supporting the linguistic stream with information on the speaker's or listener's attitude or emotional state.
- Examinees spend more time watching listeners (i.e., non-speaking characters) when the question asked of them is implicit, as they believe that this information will facilitate selecting the credited response.
- Audio-only listening tests differ from video listening tests largely in the absence of facial expression information.

Study II was the second study in the video listening test literature (the first being the present author's own published work) to employ many-facet Rasch

modeling in order to estimate the per-item bias toward video or audio. It was also the first comparative study (among those published) to incorporate task as a variable. This resulted in the following new finding:

- Nonverbal and visual information delivered via video in video listening tests exerts a much stronger facilitative effect on items of the implicit task type, resulting in easier implicit items than explicit, even if that order of difficulty is reversed when delivered via audio-only.

These findings have both theoretical and practical implications, to be discussed in the following section.

7.3. Implications

This section discusses the implications of the findings of the present research. These are separated into two sections: theoretical and practical. The theoretical section will primarily focus on lingering questions of construct validity; the practical section will be devoted to discussing the logistical challenges associated with video-mediated listening tests.

7.3.1. Theoretical Implications

The discussion of the theoretical implications of the present research is divided into those related directly to the validity of the assessment method, and the test use case implications arising from the discussion of validity.

7.3.1.1. Construct validity

A major goal of this research was to quantify the impact of the presence of nonverbal communicative cues on the difficulty of items of two different task types: explicit and implicit. That goal has been attained, and those quantities are a 0.56-logit facilitative effect for explicit tasks, and 1.02 logits for implicit, even after two of the

implicit items exhibiting the largest video biases were removed. Just as Ockey (2007) and Batty (2015) suspected, items which require the examinee to read into social contexts, resolving ambiguity in spoken language, are facilitated to a much greater degree by video than those that attempt to tap a purely linguistic construct. It seems that the presence of visual information, especially that of facial expressions (based on the findings of Study I) enhances the aural input sufficiently to allow many of those who would be unlikely to answer the same item correctly without it to select the credited response in both explicit and implicit multiple choice items.

This finding raises further questions regarding the tested construct of video-mediated tests. Does a correct response to such an implicit item without access to the nonverbal behavior via video mean the same thing as one that resulted from a testing session that included it? How does this impact judgments of examinees' language ability? Can the comprehension of nonverbal cues that are shared by all humans, the blind, and even other primates (Eibl-Eibesfeldt, 1972, 1973, Ekman & Friesen, 1969, 1971; Ekman et al., 1972; Floyd, 2006; Fulcher, 1942; Matsumoto, 2006; Thompson, 1941) rightly be considered a component of language ability? —Or should the construct definition be widened to accommodate such ability? These are the questions revisited time and again in the video listening test literature, but the best answer to them remains equivocal at best: It depends on whether those abilities are relevant to the construct of interest.

The argument for the construct validity of video-mediated tests advanced most notably by Wagner, but also others, hinges largely upon the work of Messick and his warnings against construct underrepresentation—wherein the test is too narrow to include enough of the target aptitude or behavior (Messick, 1989, 1996). The argument is that a test which is not sufficiently authentic underrepresents the construct

and, as a result, its scores are not valid. “The major measurement concern of authenticity,” Messick writes, “is that nothing important be left out of the assessment of the focal construct” (1996, p. 243). As most interaction occurs face-to-face, the thinking goes, to deny the examinee the visual channel is to underrepresent the construct and fail to match the target language use (TLU) domain (Bachman & Palmer, 1996). To underrepresent is to render the measurements invalid.

However, Messick also writes about the problem of construct *over*representation (although he does not use that term), which threatens validity of scores because of their inclusion of construct irrelevant variance, which can occur when “the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct” (1996, p. 244). Such variance “constitutes a contaminant with respect to score interpretation” (1989, p. 34). Of particular importance here is whether the inclusion of nonverbal information results in what Messick called “construct-irrelevant easiness,” wherein “extraneous clues in item or test formats permit some individuals to respond correctly in ways irrelevant to the construct being assessed” (1989, p. 34). The central question, then, is this: Is the ability to decode facial expressions an important skill that assessors should incorporate into their measures, or is it a contaminant? This remains a difficult question, but perhaps there is a simpler way of thinking about it.

Study I demonstrates that examinees in video-mediated listening tests pay closer attention to faces, even of those who are not speaking, when viewing material for the purpose of responding to an implicit question, and Study II demonstrates that that behavior is associated with a sharp reduction in item difficulty for implicit items when compared to audio-only versions of the same items. Nowhere was this more clearly demonstrated as in Item 4 of Study II, wherein the male character implied that

he had agreed to fraudulently call in sick in order to attend a concert with his coworker. The scene concluded with him scolding his coworker, “That is very dishonest,” followed by a beat, then adding, “But it just might work.” The implication is clear, but the addition of his sly, conspiratorial facial expression in the video format resulted in the item becoming much easier than in the audio format. This is a good demonstration of the facilitative effect of nonverbal cues in such items, but it is not clear whether the item tests the same thing under the two formats.

One of the most important findings of Study II was that the average difficulty of the implicit items was lower than that of the explicit items, in contrast to much research on item task types in such tests (e.g., Bloomfield et al., 2010; Nissan et al., 1996; Shohamy & Inbar, 1991), although these estimates included both the audio and video versions of these items. When examining the biased estimates, it becomes clear that the implicit items are indeed more difficult than the explicit items when presented in the audio format, but become easier than the explicit items when presented with video. That is to say, it appears that the addition of video *reverses* the traditionally-expected order of difficulty between explicit and implicit items—a finding that has serious consequences for item design.

Implicit/gist items are typically intended to engage the examinee’s top-down processing of the target language and culture. L1 and other high-proficiency users of a language have little trouble teasing out implication from ambiguous statements because they have access to a vast trove of knowledge about idioms, social norms, pragmatics, schemata, and scripts in the target language/culture, even without access to the visual channel. It is this kind of somewhat nebulous language resource that we are attempting to measure when we ask examinees to interpret such statements. However, when the visual channel is engaged, granting access to nonverbal behavior,

it becomes unclear as to how much the examinee has drawn from his or her linguistic resources, and how much he or she has simply relied on pan-cultural nonverbal communication cues in order to answer.

The present research, by first observing examinees' watching behavior, then drawing from those observations to design a video-mediated listening test based upon those observations, suggests that the most likely difference between an audio-only version of an implicit item and an equivalent item featuring nonverbal-communication-rich video is that the examinees were able to rely on their *instinctual knowledge* of facial expressions to *sidestep* the very top-down language knowledge that such items are typically intended to assess. Is such an ability worth measuring? Furthermore, if it were, would it not be simpler to present the examinee with photos of facial expressions from the target language/culture group, which they must match to the appropriate emotional or mental state, in a manner similar to the research instruments found in the communication field, such as the FACS (Ekman & Friesen, 1978) or PONS (Rosenthal et al., 1979) tests?

Taking the idea further, should examinees on the autistic spectrum, such as those with Asperger's syndrome, receive lower listening comprehension marks due to their difficulty in discerning and decoding nonverbal behavior? The very same scholar of validity, Samuel Messick, who popularized the unified model of validity, which included admonitions against construct underrepresentation, is also he who popularized the concept of fairness as a component of validity, one which has since come to see wide acceptance. If one accepts that fairness is indeed a component of validity, the inclusion of facial expressions on language tests is somewhat problematic.

7.3.1.2. Test use

If one rejects the above reservations, however, there are indeed test use cases that can be imagined that would reasonably call for such nonverbal information to be included; however, they are not without their own qualifications.

It would seem that if the TLU domain (Bachman & Palmer, 1996) of a test is expected to primarily involve face-to-face conversations, and the examinee is therefore expected to always have access to nonverbal cues associated with the verbal channel, then it could easily be argued that a video-mediated test is more authentic, and that the variance observed in the scores as a result of this information is construct-relevant. What kinds of test might meet this criterion?

The present researcher's previous work on video-mediated tests of academic listening found no difference (Batty, 2015), and work by others investigating how much examinees view videos of academic talks has consistently found that it is less than what is observed in conversations (Suvorov, 2013, 2015, Wagner, 2007, 2010a). Simply put, there just is not much nonverbal information in academic listening stimuli such as lectures. In Suvorov's eye-tracking work, participants oriented toward "content visuals"—e.g., presentation slides—much more than "context visuals"—i.e., the speaking lecturer. Given that, tests of academic listening proficiency seem like a poor choice for the use of video, even if only from a logistical standpoint (see next section). Notably, ETS evaluated the prospect of incorporating video into the listening section of the TOEFL iBT, and opted, rather, for still visuals instead (Chapelle, Enright, & Jamieson, 2008).

If academic language tests are not a strong candidate for the addition of nonverbal communication as a component of the listening construct, what may be? Once can easily imagine a language for specific purposes (LSP) test for service-

people, such as those who will staff an information booth in a public place such as an airport, tourist information center, or large department store. The ability of sales staff who must support a foreign clientele to demonstrate some kind of conversational competence may also be worth measuring. However, are such tests a large part of the family of language tests? How frequently are such tests developed and used? For-profit business language tests such as BULATS or Versant, or even the nonprofit ETS' TOEIC, feature no video. The only business English test known to the present author which features video is Benesse's GTEC, which is almost entirely limited to the Japanese market. If there were a call from test users for the inclusion of nonverbal communicative competence in business-focused language assessments, would these testing companies not have responded?

Moreover, despite observations that most conversations take place face-to-face, and therefore include nonverbal information, it does not follow that those times in which it is unavailable are unimportant. Even in the customer-facing service roles imagined above, the ability to comprehend a phone call is likely just as important as a face-to-face conversation. Even with the ubiquity of the smartphone and mobile data technology well-established, distance communication is still largely an audio-only (or textual) affair. Video conferencing does indeed occur, but the conference call is still likely more common. Surely a business language test would need to incorporate an audio-only listening section in addition to video in order to closely match the TLU domain, but one wonders whether it would be worth the effort, as a proficient listener would perform comparably on both.

Even tourists need the ability to comprehend spoken language without the benefit of nonverbal cues, as airports, train stations, and buses make all announcements over a public address system, and rarely can we see them as they do it.

In fact, many of the most important messages a language user should comprehend are delivered aurally. As a resident of Japan, the present researcher has indelible memories of the frequent public safety announcements delivered from police cars and municipal public address speakers in the days and weeks after the March 11, 2011 Tohoku earthquake and subsequent nuclear disaster. To dismiss listening situations such as this due to their rarity could have grave consequences on test takers—consequences that validity theorists such as Messick and Kane (e.g., 2001) could easily be imagined calling into question the validity of tests which ignore these important language use cases.

As demonstrated above, although there are indeed contexts in which a video-mediated test would make sense from a validity standpoint, they are likely edge cases; there is nothing to suggest that test users are particularly interested; and even in such cases, the importance of audio-only listening comprehension should not be overlooked.

7.3.2. Practical Implications

This section details some practical implications of the present research. It first addresses challenges associated with attempting to mimic authentic language use, then those associated with the production and delivery of video-mediated listening tests.

7.3.2.1. Authenticity

One of the challenges of video-mediated MC listening tests is one that is shared with audio-only tests, but is nonetheless more difficult; that is ensuring that one's stimulus material is close enough to authentic language production as to represent the focal construct, while still ensuring that item distractors distract. If one uses purely authentic language, it can be difficult or impossible to write distractors for MC items which contain words or concepts from the listening passage, but which are

not the credited response. This is especially the case with conversational texts, whose propositional content is typically rather thin. Figure 7.1 displays an example of pseudo-authentic misdirection in such a script. In the conversation, four items are discussed in rapid selection, but one of them will not be brought to the party. This allows for several different questions to be written, using the other items as distractors. If a hypothetical question were to ask what the woman was going to bring to the party, “beer,” “fruit,” and “cocktails” would be viable choices for distractors, “cheese log” would not.

This difficulty is compounded, however, when the content is to be delivered via video, as not only the script, but the performances must be tended to very carefully. With an audio-only test, it is possible to seat the performers in a soundbooth with the scripts in hand, obviating the need for memorization. Missed lines, or poor line delivery can be addressed immediately, without stopping the recording, and the best performances can be edited together seamlessly. Speed, length of pauses, even line order can all be manipulated after the fact.

None of these are true with video, especially if the goal is to include as few cuts as possible. Cue cards off camera damage the pseudo-authenticity of the performances, as performers glance or even stare at unnatural angles, looking past or away from the other people in the scene. Lines must be memorized. A poor take requires the entire scene to be run again. Very little can be adjusted in post-production. The performance recorded is, for the most part, the performance that will be presented in the final test. It is much more time-consuming, and requires much better performers, than an equivalent audio-only test.

Furthermore, in order for the items to function as designed, the performers must adhere to all scripts exactly. A forgotten word which may not have changed the

MAN

(calling after WOMAN as approaching)

Hey! Wait up!

(catches up)

Hey, are you going to Yoko's party on Saturday?

WOMAN

I am.

MAN

(hands forward)

What're you bringing?

WOMAN

I was going to bring some homemade cookies. You?

MAN

I'm thinking fruit.

(pointing)

Oh, and beer. She never has any.

WOMAN

I was originally planning on bringing stuff to make cocktails, but I don't have time to go shopping.

MAN

Y'know, that's probably for the best; most of the people going don't even drink.

WOMAN

Oh! I didn't even think of that.

(nodding in emphasis)

Good thing I didn't spend any money on it, then.

Figure 7.1. Example of misdirection in a pseudo-authentic script.

meaning of the line, or even its believability, may negatively impact an item that relies upon it. This is further compounded when nonverbal information is deliberately scripted, as in the present research. The pressure placed on the performers and director

to maintain a balance between a sense of authenticity and the viability of items far outstrips that of an audio-only test.

7.3.2.2. Production and delivery

Despite the drop in prices for high-quality video production hardware and software in recent years, video production is still far more expensive than audio production. A reasonably-good quality condenser microphone can be had for £200 (Blue's Yeti Pro), with an integrated USB digital-analog converter (DAC), and a professional-grade digital audio workstation (DAW) software package, complete with more effects that would ever be necessary for such an application, for £150 (Apple's Logic X; although GarageBand would almost certainly be sufficient at £4), computer not included. These two purchases would provide a small test developer with a professional-sounding audio production system for little over £350 including any other incidental purchases.

The contrast with the equipment used for the production of the instrument in Study II is striking:

- Sony HXR-NX70U HD camera (~£2,000)
- Heavy-duty tripod (£200)
- Shure wireless microphone system (~£300)
- Behringer 2-channel microphone preamp (£75)
- XLR cables to connect the mic system to the preamp, and the preamp to the digital audio recorder (~£40)
- Tascam DR-40 digital audio recorder (£200)
- Tripod for the DR-40 (£100)
- Adjustable LED fill lights with stands (£100)
- Apple Final Cut Pro X video editing software with Compressor add-in (£270)

This comes to a total of £3,285, nearly ten times the price of the audio solution listed above, but does not match the quality of a true professional video production unit, and ignores the cost of a computer powerful enough to edit video reliably.

The equipment is not the only added cost of video production, however. Unlike the case of audio, any objects referred to in the scripts must be purchased. Sets must be dressed. Costumes must be arranged for. Rates for actors who will appear on screen are higher than those who do voiceover work. Once all the incidentals are included in the price, the difference is even clearer. It is partially for this reason that ETS, though originally considering video for the TOEFL iBT, ultimately abandoned it as a possibility (Bejar et al., 2000).

Although the challenges associated with maintaining pseudo-authenticity are discussed above, the opposite problem can arise during production. Actors are trained to create realistic portrayals of people in the scenes depicted on the script page, but this sometimes results in natural-seeming behaviors that ultimately damage items, as was clearly demonstrated by Items 1 and 7 in Study II.

The issues encountered with Items 1 and 7, which ultimately had to be removed from the final analysis, illustrate a central logistical challenge associated with video-mediated listening tests, one which is not well-addressed by the literature. In the cases of these two items, the actors were merely plying their craft and acting in logical, realistic ways, given the script. It is entirely natural for someone to speak as he or she moves to complete an action, as the actress did at the end of the Item 1. It was the present researcher's responsibility as scene director to make note of the implications of this, and to request a retake of the scene. However, at the time of filming, since the portrayal was realistic, its implication for the item went unnoticed. So, too, was the case in Item 7. Item writers are accustomed to carefully crafting the

lines to be spoken for listening tests so as to provide opportunities for distractors without creating dialog that differs too much from authentic speech as to threaten construct validity. However, adding the silent, visual aspect to the test design process adds another layer of complexity. In the cases of both Items 1 and 7, silent visual cues crept into the final material that essentially disqualified the items from inclusion in the study. From a logistical standpoint, the production of video for listening tests will always be much more involved than that of audio.

In addition to the cost and complexity, the technical requirements of delivering video, too, represent a significant challenge. When the present researcher began work on the video-mediated listening test originally designed by Paul Gruba at Kanda University of Foreign Studies, the video delivery required the use of a campus-wide AV system, allowing the testing committee to play the video from a central location on campus, which was wired to every classroom in use. For computer-mediated tests, the video must be stored locally on the computer in question, and the appropriate software installed, or streamed over the Internet to the examinee's browser. Given the myriad video formats and browser incompatibilities, the latter can be very challenging, requiring multiple versions of the files to operate as "fallbacks," should a browser report that it is unable to play the preferred file type. Furthermore, as described in the Instrument Development chapter, streamed video requires fast servers, sufficient bandwidth, and reliable connections to the client. The challenge associated with this is another reason for ETS' choice to instead stream still context photos, rather than video, for the TOEFL iBT (Bejar et al., 2000).

All told, the challenges associated with producing a video-mediated listening test are formidable, and may not result in scores that are any more valid or interpretable.

7.4. Recommendations for Language Assessors

Although the above takes a rather dim view of the use of video in foreign language listening tests, as is conceded in the "Test use" section, there are, perhaps, times when a video-mediated listening test, or section of a test, could provide benefit to a language assessor by more closely modeling the situations in which the examinee is likely to find him or herself. In such cases, there are a handful of lessons to be learned from the present project.

7.4.1. Professional Actors

Having produced video-mediated tests using language faculty as performers, the present researcher found that working with professionals was comparatively easy. As paid professionals, they arrived on time, and they arrived prepared. Their performances were believable, and contained natural-seeming nonverbal behavior, despite the fact that it was strictly scripted. Actors with experience in improvisational theatre are especially comfortable changing their performances quickly if a change needs to be made. Professionals can be expensive, but the difference in quality of performance is hard to overstate.

7.4.2. Adequate Production Crew

Although one may be tempted to video record performances with a very small crew, it is important to ensure that the director (likely a test developer) is supported by an assistant who reads the script as the scene is performed to catch deviations from it, allowing the director to focus his or her attention on the performances themselves. This also necessitates a dedicated camera operator, and likely a sound technician. Even for short, simple scenes, an adequately sized production crew allows for a clear division of labor and fewer mistakes. The problems observed in Items 1 and 7 are the clearest examples of problems that can arise if the director is not vigilant enough, but

Item 18, too, would have been improved if the present researcher had noticed that the woman did not release an exasperated sigh before her last line, demonstrating her frustration. This would have better signaled to those who encountered the item in the audio condition her emotional state, but as the performance was believable without it, it went unnoticed.

7.4.3. Sound Issues

In the video production for Study II, three separate audio streams were used to ensure that there were multiple sources in case of a problem with any one of them. However, in some cases, time-consuming post-production was still necessary to clean up the sound. The problem most difficult to address was air conditioner noise. The air conditioner was switched off whenever and wherever possible, but in one location, the control panel was inaccessible. When scouting locations, it is advisable to include this on one's checklist.

Another source of audio trouble were the wireless lapel mics used. Although they maintained reliable connection to the head unit, they picked up quite a lot of body movement, especially when women wore garments such as scarves. These mics were chosen because of the technical difficulty of operating a boom mic, but if a boom mic is an option, it will almost certainly deliver a more reliably-useful audio signal.

7.4.4. Scheduling

Video production is very time consuming. All the scenes in Study II were performed and recorded in one day, but the day was extremely tight, and went well over schedule, even though quite a few scenes were cut from the schedule due to inclement weather. Consulting with the actors or their agent(s) on a reasonable schedule may be an option if one has not produced such content before.

7.4.5. Outdoor scenes

Unless one has access to a soundstage or a quality green-screen room, outdoor scenes should be avoided. The weather's cooperation cannot be assumed.

7.4.6. Other Options

As discussed above, ETS opted against video in the TOEFL iBT, choosing, rather, to provide the examinees with context still images, and some content images (e.g., presentation slides). These provide the schema-setting benefit of video without the validity concerns presented by the nonverbal cues.

7.5. Limitations and Future Directions

This section describes some of the limitations of Study I and II, and recommends further work to address them and explore the topics of this research in more depth.

7.5.1. Eye-Tracking

Study I has some limitations which could motivate future eye-tracking research into examinee interactions with video-mediated listening tests. The first of these is the small sample size. A larger sample size would open up the possibility of more statistical analyses and generate more insightful quotes from respondents.

Another limitation of Study I was the Pupil eye-tracker. Although it met the needs of the exploratory study as designed, newer and more advanced hardware could reveal patterns undetectable by the Pupil unit, and could do so without as much labor. In the time since the study was completed, prices for eye-tracking hardware and software have fallen, increasing the number of options. Future work in this vein, therefore, should explore the possibility of using a newer eye-tracking apparatus.

In Study I, item factors were inextricable from videotexts, rendering it impossible to determine how viewing behavior may change on the same video, but a

different item. A future eye-tracking study may address this, likely replacing the *Curb Your Enthusiasm* videos with those from Study II, which have been created with the benefit of the insights gained in Study I.

Two final suggestions for future work in this vein are related to lip-reading and question preview. To attempt to determine the impact of lip reading on comprehension in video tests, item responses based on standard videotexts could be compared to versions wherein the mouth area of speakers is obscured (e.g., blurred or pixelated).

Finally, the effect of question preview on viewing behavior could be investigated by comparing behavior when examinees already know the question to be answered, and when they do not.

7.5.2. Increased Granularity of Gesture Classifications

Study I separated nonverbal cues into two broad categories: facial expressions and gestures (which were separated into illustrators and emblems), drawing from the classification model proposed by Ekman and Friesen (1969). However, the Ekman and Friesen model would itself allow for smaller subdivisions of gestural behavior, and the McNeill (1992) model offers several more. It may be fruitful to classify all gestural movements in a video stimulus and investigate attention to them via eye-tracking, although it would likely require much more data, as gestures rather infrequently receive intentional foveal orientation.

7.5.3. Increased Granularity of Task Types

Although both of the present studies only classified task types into "explicit" and "implicit," similar to the work of Wagner (2002, 2006), of Shohamy and Inbar (1991), and of Nissan et al. (1996), these can be broken down into more specific task types as demonstrated by Freedle & Kostin (1999; Kostin, 2004) and by Field (2013). Video viewing and/or item responses may be affected by these in different ways;

investigating them at a higher degree of granularity than presented here may yield useful results.

7.5.4. Use of Multi-Parameter and Multi-Dimensional IRT

With regards to Study II, although MFRM demonstrates well the contrast in item difficulty between the audio and video formats, a limitation inherent to Rasch models is the fact that they estimate a single parameter (ability/difficulty). A two-parameter logistic (2PL) model could also estimate discrimination for comparison between the formats, and a three-parameter (3PL) model could add guessing. It is possible that, for example, video items discriminate more poorly than audio, or video items have a higher guessing level (lower asymptote). However, in order to use either of these methods, an anchoring section between the audio format items and the video format items would be necessary to equate estimates. This would necessitate both a longer test, and a larger sample than that of Study II, but would certainly shed even more light on questions pertaining to item-format and task-format interaction.

Another possibility for analysis would be multi-dimensional IRT, which could be used to explore the possibility that listening comprehension and visual comprehension could load on separate traits. Once again, however, this would require a more complex design and a larger sample.

7.6. Conclusion

This project has sought to shed light on several related questions left unanswered by the literature on video-mediated listening tests. It has employed both quantitative and qualitative eye-tracking methodology to investigate examinee viewing behavior and its interaction with two item task types: explicit and implicit. It then applied those findings to the development of a new video-mediated listening test. The new test was administered to a sample of 279 Japanese university students, alternating between

audio-only and video-mediated formats. The impact of format on overall item difficulty, and, more critically, on the difficulty of the two task types was investigated quantitatively with MFRM, and contrasts were interpreted with the aid of qualitative analysis of the items and videos used. Implications, both theoretical and practical, were discussed, and future research directions were proposed.

References

Adapter. (2014). (Version 2.1.4) [Macintosh]. McKinney, TX, USA: Macroplant LLC.

Alderson, J. C. (1990). Testing reading comprehension skills (Part One). *Reading in a Foreign Language*, *6*(2), 425–438.

Alderson, J. C. (1993). Judgements in language testing. In C. A. Chapelle & D. Douglas (Eds.), *A New Decade of Language Testing Research* (pp. 46–57). Alexandria, VA: TESOL.

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, *26*(4), 621–631. https://doi.org/10.1177/0265532209346371

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556. https://doi.org/10.1177/0265532213489568

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Alma mater index: Global executives 2013. (2013, September 5). Retrieved July 29, 2016, from https://www.timeshighereducation.com/news/alma-mater-index-global-executives-2013/2007032.article

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The "blank screen paradigm." *Cognition*, *93*(2), B79–B87. https://doi.org/10.1016/j.cognition.2004.02.005

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Ambady, N., LaPlante, D., & Johnson, E. (2001). Thin-slice judgments as a measure of interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 89–101). Mahwah, NJ: Lawrence Erlbaum Associates.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, *13*(2), 125–150. https://doi.org/10.1177/026553229601300201

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). Test usefulness: Qualities of language tests. In *Language testing in practice: Designing and developing useful language tests* (pp. 17–42). Oxford: Oxford University Press.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Baltova, I. (1994). The impact of video on the comprehension skills of core French students. *Canadian Modern Language Review*, *50*(3), 507–31.

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

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.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303–310. https://doi.org/10.1177/014662168300700306

Bejar, I. I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (Working Paper No. RM-00-07) (p. 60). Princeton, NJ: Educational Testing Service.

Berne, J. E. (1993). The role of text type, assessment task, and target language experience in L2 listening comprehension assessment. Presented at the Annual Meetings of the American Association for Applied Linguistics and the American Asociation of Teachers of Spanish and Portuguese, Cancun, Mexico. Retrieved from http://eric.ed.gov/?id=ED358737

Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, *78*(2), 316–329. https://doi.org/10.2307/345428

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension* (Technical Report No. TTO 81434 E.3.1). College Park, MD: University of Maryland, Center for Advanced Study of Language. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA550176

Bojko, A. (2005). Eye tracking in user experience testing: How to make the most of it. In *Proceedings of the UPA 2005 Conference*.

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). New York: Psychology Press.

Boland, J. E. (2005). Visual arguments. *Cognition*, *95*(3), 237–274. https://doi.org/10.1016/j.cognition.2004.01.008

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Borsboom, D. (2009). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, *50*(1), 110–114. https://doi.org/10.1111/jedm.12006

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

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

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*(1), 171–191. https://doi.org/10.1017/S0267190500003536

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369–394. https://doi.org/10.1191/0265532202lt236oa

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study* (ARAGs Research Reports Online No. AR/2015/001). London: The British Council. Retrieved from 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

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

Burgoon, J. K. (1994). Nonverbal signals. In M. L. Knapp & G. R. Miller (Eds.), *Handbook of interpersonal communication* (2nd ed., pp. 229–285). Thousand Oaks, CA: SAGE Publications, Inc.

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

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, *70*(2), 205–218. https://doi.org/10.1037/0022-3514.70.2.205

Chang, A. C.-S., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, *40*(2), 375–397. https://doi.org/10.2307/40264527

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Florence: Taylor and Francis.

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

Chovil, N. (1991b). Social determinants of facial displays. *Journal of Nonverbal Behavior*, *15*(3), 141–154. https://doi.org/10.1007/BF01672216

Chung, U. K. (1994). *The effect of audio, a single picture, multiple pictures or video on second-language listening comprehension* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hoboken: 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*, *29*(1), 1–14. https://doi.org/10.1016/S0346-251X(00)00057-9

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Costanzo, M., & Archer, D. (1989). Interpreting the expressive behavior of others: The Interpersonal Perception Task. *Journal of Nonverbal Behavior*, *13*(4), 225–245. https://doi.org/10.1007/BF00990295

Creswell, J. W., & Plano Clark, V. L. (2010). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

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

Daigaku hensachi ichiran/rankingu 2017 [University hensachi list/ranking 2017]. (2016). Retrieved July 29, 2016, from http://hensachimap.com/%e5%a4%a7%e5%ad%a6%e5%81%8f%e5%b7%ae%e5%80%a4%e3%83%a9%e3%83%b3%e3%82%ad%e3%83%b3%e3%82%b02017

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice* (2nd ed.). London: Springer.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt: Peter Lang.

Educational Testing Service. (2016). Administration and scoring. Retrieved July 28, 2016, from https://www.ets.org/toefl_itp/administration_scoring/

Educational Testing Service. (2016). How to use the TOEFL ITP Assessment Series. Retrieved July 28, 2016, from https://www.ets.org/toefl_itp/use/

Educational Testing Service. (2016). Test content. Retrieved July 28, 2016, from https://www.ets.org/toefl_itp/content/

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

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

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it* (Vol. 1, pp. 119–128). British Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=1531311

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

Eibl-Eibesfeldt, I. (1973). The expressive behavior of the deaf-and-blind born. In M. von Cranach & I. Vine (Eds.), *Social communication and movement*. New York: Academic Press.

Ekman, P. (1984). Expression and the nature of emotion. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 319–343). London: L. Erlbaum Associates.

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

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

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129. https://doi.org/10.1037/h0030377

Ekman, P., & Friesen, W. V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Los Altos, CA: Malor Books.

Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System (FACS): A technique for the measurement of facial action*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York: Pergamon Press Inc.

Elkhafaifi, H. (2005). The effect of prelistening activities on listening comprehension in Arabic learners. *Foreign Language Annals*, *38*(4), 505–513. https://doi.org/10.1111/j.1944-9720.2005.tb02517.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J: Psychology Press.

Engelhard, G. (2009). Using item response theory and model–data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*(4), 585–602. https://doi.org/10.1177/0013164408323240

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Routledge.

Engelhard, G., Jr. (2007). Differential rater functioning. *Rasch Measurement Transactions*, *21*(3), 1124.

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Engelhard, G., Kobrin, J. L., & Wind, S. A. (2014). Exploring differential subgroup functioning on SAT writing items: What happens when English is not a test taker's best language? *International Journal of Testing*, *14*(4), 339–359. https://doi.org/10.1080/15305058.2014.931281

Exam English Ltd. (2014). TOEIC® Listening part 3: Conversations. Retrieved May 15, 2016, from http://www.examenglish.com/

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

Feldman, R. S., & Tyler, J. M. (2006). Factoring in age: Nonverbal communication across the life span. In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 181–199). Thousand Oaks, CA: SAGE Publications, Inc.

Fernández-Dols, J. M., & Carroll, J. M. (1997). Is the meaning perceived in facial expression independent of its context? In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 275–320). Cambridge: Cambridge University Press.

Feyten, C. M. (1991). The power of listening ability: An overlooked dimension in language acquisition. *The Modern Language Journal*, *75*(2), 173–180. https://doi.org/10.1111/j.1540-4781.1991.tb05348.x

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

Filipi, A. (2012). Do questions written in the target language make foreign language listening comprehension tests more difficult? *Language Testing*, *29*(4), 511–532. https://doi.org/10.1177/0265532212441329

Final Cut Pro. (2015). (Version 10.1.4) [Macintosh]. Cupertino, USA: Apple Inc.

Fisher, W. P. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*(3), 238.

Flowerdew, J., & Miller, L. (2010). Listening in a second language. In A. D. Wolvin (Ed.), *Listening and Human Communication in the 21st Century* (pp. 158–177). John Wiley & Sons.

Floyd, K. (2006). An evolutionary approach to understanding nonverbal communication. In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 139–157). Thousand Oaks, CA: SAGE Publications, Inc.

Folger, J. P., & Woodall, W. G. (1982). Nonverbal cues as linguistic context. In M. Burgoon (Ed.), *Communication Yearbook 6* (pp. 63–91). Beverly Hills, CA: SAGE Publications, Inc.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*(1), 2–32. https://doi.org/10.1177/026553229901600102

Fridlund, A. J., & Russell, J. A. (2006). The functions of facial expressions: What's in a face? In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 299–319). Thousand Oaks, CA: SAGE Publications, Inc.

Fulcher, J. S. (1942). "Voluntary" facial expression in blind and seeing children. *Archives of Psychology (Columbia University)*, *38*(272).

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

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*(2), 133–167. https://doi.org/10.1191/0265532202lt225oa

Goldberg, J. H., & Whichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 493–516). Amsterdam: Elsevier Science.

Gorin, J. S. (2006a). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Gorin, J. S. (2006b). Using alternative data sources to inform item difficulty modeling. Presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, *10*(1), 57.

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, *3*, 182–200.

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

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

Gruba, P. (2006). Playing the videotext: A media literacy perspective on video-mediated L2 listening. *Language, Learning & Technology*, *10*(2), 77–92.

Guion, R. M. (1977). Content validity—The source of my discontent. *Applied Psychological Measurement*, *1*(1), 1–10. https://doi.org/10.1177/014662167700100103

Hall, J. A. (2006). Women's and men's nonverbal communication: Similarities, differences, stereotypes, and origins. In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 201–218). Thousand Oaks, CA: SAGE Publications, Inc.

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

Hernandez, S. S. (2004). *The effects of video and captioned text and the influence of verbal and spatial abilities on second language listening comprehension in a multimedia learning environment* (Unpublished doctoral dissertation). New York University, New York. Retrieved from http://search.proquest.com/pqdt/docview/305166044/abstract/13FE4D5FFBD2C1FDDC/

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures. *Gesture*, *3*(2), 127–154.

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: Oxford University Press.

Imhof, M. (2010). What's going on in the mind of a listener? The cognitive psychology of listening. In A. D. Wolvin (Ed.), *Listening and Human Communication in the 21st Century* (pp. 97–126). John Wiley & Sons.

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

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The Mind's Eye* (pp. 573–605). Amsterdam: North-Holland. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780444510204500311

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115–122. https://doi.org/10.1111/jedm.12007

Kassner, M., & Patera, W. (2012). *Pupil: Constructing the space of visual attention* (Unpublished master's thesis). Massachusetts Institute of Technology, Cambridge, MA. Retrieved from http://hdl.handle.net/1721.1/72626

Kassner, M., Patera, W., & Bulling, A. (2014). *Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction*. Retrieved from arXiv.org. (1405.0006)

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.

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

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Knapp, M. L. (2006). An historical overview of nonverbal research. In V. L. Manusov & M. L. Patterson (Eds.), *The SAGE handbook of nonverbal communication* (pp. 3–19). Thousand Oaks, CA: SAGE Publications, Inc.

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

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (Research Report No. RR-04-11). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2004/hsif

Koyama, D., Sun, A., & Ockey, G. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology*, *20*(1), 148–165.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1992). Treatment effects: Fixed-effects chi-square of homogeneity. *Rasch Measurement Transactions*, *6*(2), 218–219.

Linacre, J. M. (2002a). Facets, factors, elements and levels. *Rasch Measurement Transactions*, *16*(2), 880.

Linacre, J. M. (2002b). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Mean-square and standardized chi-square fit statistics. *Rasch Measurement Transactions*, *17*(1), 918.

Linacre, J. M. (2013). Disconnected subsets, Guttman patterns and data connectivity. *Rasch Measurement Transactions*, *27*(2), 1415–1417.

Linacre, J. M. (2014). FACETS (Version 3.71.4). Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm

Linacre, J. M. (n.d.). Bias interaction DIF DPF DRF estimation. Winsteps.com. Retrieved from http://winsteps.com/facetman/biasestimation.htm

Logic Pro X. (2015). (Version 10.1.1) [Macintosh]. Cupertino, USA: Apple Inc.

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics*, *17*(1). Retrieved from http://escholarship.org/uc/item/0c080191.pdf

Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics*, *18*(1), 3–19. https://doi.org/10.1017/S0267190500003457

MacWilliam, I. (1986). Video and language comprehension. *ELT Journal*, *40*(2), 131–135. https://doi.org/10.1093/elt/40.2.131

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emotion*, *14*(3), 195–214.

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

Matsumoto, D., Consolacion, T., Yamada, H., Suzuki, R., Franklin, B., Paul, S., … Uchida, H. (2002). American-Japanese cultural differences in judgements of emotional expressions of different intensities. *Cognition & Emotion*, *16*(6), 721–747.

Matsumoto, D., Kazri, F., & Kooken, K. (1999). American-Japanese cultural differences in judgments of expression intensity and subjective experience. *Cognition & Emotion*, *13*(2), 201–218.

Matsumoto, D., Takeuchi, S., Andayani, S., Kouznetsova, N., & Krupp, D. (1998). The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian Journal of Social Psychology*, *1*(2), 147–165. https://doi.org/10.1111/1467-839X.00010

Matsumoto, D., Yoo, S. H., Hirayama, S., & Petrova, G. (2005). Development and validation of a measure of display rule knowledge: The display rule assessment inventory. *Emotion*, *5*(1), 23–40. https://doi.org/10.1037/1528-3542.5.1.23

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, *32*(1), 1. https://doi.org/10.1207/s15326985ep3201_1

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31. https://doi.org/10.1207/s15434311laq0301_3

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

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). USA: American Council on Education.

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

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 615–622. https://doi.org/10.1037/0278-7393.18.3.615

Morris, D. (1977). *Manwatching: A field guide to human behavior*. New York: H.N. Abrams.

Morris, D. (1994). *Bodytalk: A world guide to gestures*. London: Cape.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement*, *4*, 386–422.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (Research Report No. RR-95-37). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/1996/hxsu

Noller, P. (1985). Video primacy—A further look. *Journal of Nonverbal Behavior*, *9*(1), 28–47. https://doi.org/10.1007/BF00987557

NVivo qualitative data analysis software. (n.d.). (Version 10.0.638.0) [Windows]. Doncaster, Australia: 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

Parkinson, B. (2005). Do facial movements express emotions or communicate motives? *Personality & Social Psychology Review*, *9*(4), 278–311. https://doi.org/10.1207/s15327957pspr0904_1

Parry, T. S., & Meredith, R. A. (1984). Videotape vs. audiotape for listening comprehension tests: An experiment. *OMLTA Journal*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED254107

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*(3), 247–269. https://doi.org/10.1111/j.1745-3984.2008.00063.x

Penfield, R. D. (2010a). DDFS (Version 1.0). Miami, FL. Retrieved from http://www.education.miami.edu/Facultysites/Penfield/DDFS.zip

Penfield, R. D. (2010b). DDFS: Differential distractor functioning software. *Applied Psychological Measurement*, *34*(8), 646–647. https://doi.org/10.1177/0146621610375690

Pernice, K., & Nielsen, J. (2009). How to conduct eyetracking studies. Nielsen Norman Group. Retrieved from http://www.nngroup.com/reports/how-to-conduct-eyetracking-studies/

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174. https://doi.org/10.1037/0096-3445.109.2.160

Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal*, *14*(1), 34–44.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

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

Ricci Bitti, P. E., & Poggi, I. (1991). Symbolic nonverbal behavior: Talking through gestures. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 433–457). Paris: Cambridge University Press.

Rilcy, R. (2008). *Achieve TOEIC Bridge™: Test-preparation guide*. London: Cengage Learning.

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

Rosenberg, E. L. (1997). Introduction. In P. Ekman & E. L. Rosenberg (Eds.), *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (pp. 3–17). New York: Oxford University Press.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: Johns Hopkins University Press.

Rost, M. (1990). *Listening in language learning*. New York: Longman.

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). St. Ives, UK: Pearson Education.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, *78*(2), 199–221. https://doi.org/10.2307/329010

Rubin, J. (1995). The contribution of video to the development of competence in listening. In D. J. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 151–165). Singapore: Dominie Press, Inc.

Russell, J. A. (1995). Facial expressions of emotion: What lies beyond minimal universality? *Psychological Bulletin*, *118*(3), 379–391. https://doi.org/10.1037/0033-2909.118.3.379

Russell, J. A., & Fernández-Dols, J. M. (1997). What does a facial expression mean? In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 3–30). Cambridge: Cambridge University Press.

Russell, J. A., Fernández-Dols, J. M., & Chovil, N. (Eds.). (1997). Facing others: A social communicative perspective on facial displays. In *The psychology of facial expression* (pp. 321–333). Cambridge: Cambridge University Press.

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

Scherer, K. R. (1980). The functions of nonverbal signs in conversation. In R. N. St. Clair & H. Giles (Eds.), *The Social and psychological contexts of language* (pp. 225–244). Hillsdale, N.J.: L. Erlbaum Associates.

Schroeders, U., Wilhelm, O., & Bucholtz, N. (2010). Reading, listening, and viewing comprehension in English as a foreign language: One or more constructs? *Intelligence*, *38*(6), 562–573. https://doi.org/10.1016/j.intell.2010.09.003

Shin, D. (1998). Using videotaped lectures for testing academic listening proficiency. *International Journal of Listening*, *12*, 57–80.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*(1), 23–40. https://doi.org/10.1177/026553229100800103

Short business conversations / TOEIC® listening. (2016). Retrieved May 15, 2016, from http://www.english-test.net/toeic/listening/short_business_conversations.html#list

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*(1), 66–78.

Stevenson, D. K. (1985a). Authenticity, validity and a tea party. *Language Testing*, *2*(1), 41–47. https://doi.org/10.1177/026553228500200105

Stevenson, D. K. (1985b). Pop validity and performance testing. In Y. P. Lee & U. of H. Kong (Eds.), *New directions in language testing: Papers presented at the International Symposium on Language Testing, Hong Kong* (pp. 111–118). Exeter, UK: Pergamon Press.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

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). Ames, IA: Iowa State University.

Suvorov, R. (2013). *Interacting with visuals in L2 listening tests: An eye-tracking study* (Doctoral thesis). Iowa State University, Ames, IA.

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

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. USA: Pearson Education.

Tanenhaus, M. K. (2007). Spoken language comprehension: Insights from eye movements. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 309–326). Oxford: Oxford University Press.

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL® ITP scores onto the Common European Framework of Reference* (Research Memorandum No. RM–11-33) (p. 31). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RM-11-33.pdf

Taylor, L. (2013). Introduction. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 1–35). Cambridge: Cambridge University Press.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, *10*(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Thompson, J. (1941). Development of facial expression of emotion in blind and seeing children. *Archives of Psychology (Columbia University)*, *37*(264).

Toulmin, S. E. (2003). *The uses of argument*. Cambridge: Cambridge University Press.

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 & Applied Linguistics, Teachers College, Columbia University*, *2*(1). Retrieved from http://journals.tc-library.org/index.php/tesol/article/viewArticle/7

Wagner, E. (2006). *Utilizing the visual channel: An investigation of the use of video texts on tests of second language listening ability* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York.

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

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

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

Weidenmann, B. (1989). When good pictures fail: An information-processing approach to the effect of illustrations. In H. Mandl & J. R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 157–170). Tokyo: Elsevier Science.

Wolvin, A. D. (2010). Listening engagement: Intersecting theoretical perspectives. In A. D. Wolvin (Ed.), *Listening and Human Communication in the 21st Century* (pp. 7–30). John Wiley & Sons.

Wolvin, A. D., & Coakley, C. G. (1996). *Listening* (5th ed.). Madison: McGraw-Hill.

World university rankings. (2015, September 30). Retrieved July 29, 2016, from https://www.timeshighereducation.com/world-university-rankings/2016/world-ranking

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, *9*(4), 472.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions*, *16*(3), 888.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*(1), 107–122. https://doi.org/10.1016/j.system.2007.12.003

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

Appendix A.

Study I Procedural Checklist and Base Interview Script

Note: “Local” and “global” items correspond to “explicit” and “implicit,” respectively.

Experimental Procedure

Before Session

1. Add student to Moodle course
2. Boot Ubuntu
3. Test eye tracker cameras
4. Test audio capture
5. Test keyboards and mice
6. Place fixed chair
7. Close curtain

Session Script

	Respondent	Researcher	Assistant
Headset	Sit in fixed chair; put on headset	1. Wipe headset with alcohol sheet 2. Assist respondent with putting on the headset. [Adjust with nosepiece risers if necessary.] 3. Adjust monitor height to eye level. *Please sit comfortably, but please keep your back against the chair back so that you stay sitting up straight.* *十分にリラックスしてもらいたいのですが、背筋だけ、伸ばした状態を保つようお願いします。*	
Practice (#1)	Read screen	*Now let's get started. First you will see a written introduction to some of the characters of the video. After reading that, please click the "Next" button. You will then see a question. After reading that, click "Next" and watch the video one time. After the video, please click "Next" again and choose the answer that you think is best.* *じゃあ、始めましょう。まずは、ビデオに登場する何人かのキャラクターについての説明文が出てきます。この文章を読み終わったら、「次へ」のボタンを押してください。すると、質問が出てきます。それが読み終わったら、「次へ」のボタンを押して、ビデオを一回だけ見てください。ビデオが終わったら、「次へ」をもう一度押し、最も適していると思う答えを選んでください。*	
	Do quiz #1	Move behind divider; watch monitoring screen.	
Calibration	Sit in eye-tracking chair. Follow calibration target.	1. Operate pupil calibration from behind the divider. 1. Check calibration metrics. 2. Redo if necessary. If you must redo, do not say you need to redo it. Simply say, "here it comes again." じゃあ、もう一回します。もう一回対象物を追ってください。	*You will see a target appear on the screen. Please follow it only with your eyes.* *スクリーンに対象物が見えると思います。眼だけで、その対象物を追ってください。*
Test (#2)		*Okay, it looks like you're all ready! Let's do the main test now. You'll see 6 more videos, each with a question, and then we'll talk about them.* *はい、準備ができたようですね！本番のテストに移りましょう。これから6つのビデオを見てもらい、一つのビデオにつき一つずつ質問が出てくるので答えてもらいます。それからお話しましょう。*	
		Return to the monitoring area.	
	Take test.	Watch the monitor. Make note of interesting behavior to talk about later.	

	Respondent	Researcher	Assistant
Interview	Relax.	Return to respondent. Remove headset and let them get comfortable. Queue up the video of their eye movements.	
		Begin screen recording with Kazam. State: • Respondent name • Date / time	
	Answer questions.	Conduct stimulated recall interview (see questions on following page).	[Assist with language if necessary]
Wrap-up	Take money.	*Thank you so much for your time. This was really interesting, and you've been a big help. As promised, here is your [payment].* *貴重(きちょう)なお時間をどうもありがとうございました。おかげで、とても面白いデータがとれました。お約束通り、代金をお支払します。* Pay.	[Assist with language if necessary]
	Farewell pleasantries.		

Interview Questions

The interview process will be fairly ad hoc and informal. The following are suggestions.

Warm-up

How did it feel using the eye-tracker?

アイ・トラッカーを使ってみてどんな感じがしましたか？

- Just to warm them up, get them relaxed.
- Reassure them that they did fine; we're more interested in *them* than in their answers.

General questions

Starting/restarting question

Please try to recreate your thought processes here. What are you looking at and why?

これを見たときの思考のプロセスを振り返ってみてください。

何を見ていました？それはなぜですか？

When there is a note or something interesting in the video

That's interesting! Why did you look there at that time?

おもしろいですね！なぜこのとき、そこを見たのですか？

How does looking there help you answer the question?

そこを見ることが、質問を答えるうえで、どんなヒントになったと思いますか。

Was there anything in particular you were looking for to help answer the question?

質問を答えるために何かを特別に探していましたか？

Video-specific questions

Video 1: Larry went shopping (Local)

Just to remind you, the question here was "What did the man buy?"

念のために伝えますが、今回の質問は「男性は何を買いましたか。」でした。

Video 2: Ted and Mary aren't calling (Local)

Just to remind you, the question here was "What does the man think their friends should do?"

念のために伝えますが、今回の質問は「男性は友達が何をすればいいと思っていますか。」でした。

Video 3: Cheryl interrupting football (Global)

Just to remind you, the question here was "Why is the woman upset at the man?"

念のために伝えますが、今回の質問は「どうして女性は男性に腹が立っていますか。」でした。

Video 4: Making dinner plans (Global)

Just to remind you, the question here was "Why doesn't the man like the woman's suggestion?"

念のために伝えますが、今回の質問は「どうして男性は女性の提案が好きではないですか。」でした。

Video 5: Annoying salesgirl (Local)

Just to remind you, the question here was "What is the man shopping for?"

念のために伝えますが、今回の質問は「男性は何を探していますか。」でした。

Video 6: Cheryl gets a call (Global)

Just to remind you, the question here was "What does the man who is driving think?"

念のために伝えますが、今回の質問は「運転している男性は何を考えていますか。」でした。

Wrap-up

Do you feel that you could have answered the items as well without the video?

ビデオがなくても、質問を同じようにうまく答えることが出来たと思いますか？

Appendix B.

Study II Main Instrument Test Specification

Note: “Local” and “global” items correspond to “explicit” and “implicit,” respectively.

Test Specifications for Video-Mediated Listening Test

Aaron Olaf Batty
July 11, 2015

Purpose of Test

The purpose of the present test is to assess English language listening comprehension, with a focus on the genre of informal conversation. The broader purpose of the test is to investigate the impact of the presence of video in the listening passages upon item response. As such, all items will have both an audio-only version and an audio-video version.

Construct

The present test is intended to measure informal communicative listening proficiency in English. This can be broadly defined as incorporating the following:

- *Grammatical knowledge* encompassing phonology, lexis, and syntax.
- *Pragmatic knowledge* required to comprehend the gist of utterances based on an understanding of the illocutionary force of speech acts.
- *Sociolinguistic knowledge* of the meanings of common features of informal conversation such as idiom, figures of speech, and register.

Examinees

The examinees will be Japanese university students studying English. Japanese will be their first language, and all will have studied English over a minimum of six years of compulsory education. Target proficiency level of the examinees is TOEFL IBT 400 – 525.

Instrument

Content

Informal conversational scenes written to include not only the spoken lines, but also the nonverbal cues performers are to display. The scenes will conform to the following:

- Approximately 30 seconds long
- One male, one female
- Minimal use of setting and props
- Natural informal language use

Scenes will be performed by paid local semi-professional actors. The audio for the video and audio-only versions will be identical, but items administered in the audio-only format will feature a still from the video to control for the lost of contextual information for schema setting.

Item Task Types

Two broad item task types will be prepared: “global” item tasks and “local” item tasks, as described by Shohamy and Inbar (1991). These item types are operationally defined as below:

- *Local items* are those whose answer appears explicitly in the conversation, requiring comprehension of the “local,” linguistic content.

- *Global items* are those whose answer does not appear explicitly in the conversation, but which can be surmised by the conversation. Such items require comprehension of the "global," extralinguistic content (e.g., context, social cues, vocal tone, etc.).

It is important to note that no items will be written in such a manner that they could not be answered without the visual content.

Items

The final (i.e., post-pilot) test will be comprised of twenty (20) items of the following characteristics:

- Two (2) sections of ten (10) items each, to be delivered in either audio-only, or audio-video format
- Five (5) local items in each section (10 total)
- Five (5) global items in each section (10 total)
- Four-option multiple choice item format
- One (1) example item

Item stems will be displayed prior to watching the video or listening to the audio. They will not be accessible during viewing/listening. The stem will then appear along with the answer options after viewing.

Administration

Delivery Method

The test will be delivered online via a web browser using the Moodle quiz module or the LimeSurvey questionnaire server software package. It will require the examinee to use headphones.

Procedure

The test procedure will be as follows:

1. Informed consent
2. Instructions
3. Audio test
4. Example item
5. Test content

Total test time is expected to be approximately forty (40) minutes.

Scoring

All items will be weighted equally. Many-facet Rasch modeling will be used for format, item, and person difficulty/ability estimation.

Appendix C.

Scripts and Final Items for Study II

1. LICORICE

FADE IN:

INT. OFFICE

MAN is seated at a table with a bag of licorice. He pops a piece in his mouth and chews. WOMAN enters and walks toward a donut box.

MAN

Hey! Whatcha up to?

WOMAN

(WOMAN stops; rubbing forehead and face)

Hey. Oh, I'm just tired. Need some sugar.

MAN

(extending arm, shaking the bag)

Licorice? It's sweet.

WOMAN

(scrunched face)

Ugh. I don't know how you can eat that.

MAN

(frowning)

Huh?

(palms upturned)

Why?

WOMAN

Uh, don't you think it tastes like medicine?

MAN

(glances into the bag, scoffs, shrugs)

No?

WOMAN

Well, I do.

(moving to donut box)

No, a *donut* is just what the doctor ordered.

(opens box; it's empty)

Hey! Who ate all the donuts?

MAN

(beat; raising eyebrows, extending bag again)

Licorice?

WOMAN

(stares incredulously, then rolls eyes, sighs in resignation, shakes head quickly)

Oh, what the hell . . . *God.*

FADE OUT

[Implicit]

1) What will the woman do next?（女性は次に何をするでしょうか。）

a) eat some licorice（リコリスを食べる。）

b) go to the doctor（病院に行く。）

c) buy more donuts（ドーナツをもっと買う。）

d) go home to rest（家に帰って休む。）

2. COFFEE OR TEA

FADE IN:

INT. SCHOOL HALLWAY

MAN is standing in hallway looking at iPad. WOMAN exits a classroom, backing out. BOTH carry bookbags.

MAN

(looking up)

You ready to go?

WOMAN

Yeah, but, well...

(quizzical)

Isn't it a little early for lunch?

MAN

(looking at her watch)

Ten o'clock? Yeah maybe. Do you have any other ideas?

WOMAN

(eyebrows raised)

Coffee? I know a good place.

MAN

I drink tea.

WOMAN

I'm sure they'll have that there.

MAN

You're sure? How can you be sure? Have you ordered it?

WOMAN

(big eyes)

—Well no, but—

MAN

(pointing)

Ahah! See! You don't know! What if we get there and they don't have it? What will I drink?

WOMAN

(calming)

That isn't going to happen! They will have tea! I promise!

MAN

If you say so.

(hands out in frustrated resignation)

BOTH start to walk down the hall, off-camera

WOMAN

(big eyes)

I don't know why I agreed to have lunch with you.

FADE OUT

[Explicit]

2) What is the man worried about?（男性は何を心配していますか。）

a) he will not be able to drink tea（紅茶が飲めないこと。）

b) he will not be able to drink coffee（コーヒーが飲めないこと。）

c) it is too late for lunch（昼食には遅すぎること。）

d) he does not have enough time for lunch（昼食を食べるには十分な時間がないこと。）

3. NOTES

FADE IN:

INT. CLASSROOM

WOMAN sits at a desk with a notebook and pen. A book bag rests on the floor next to her.

MAN enters, hurriedly, carrying a bookbag, and sits next to WOMAN. WOMAN does not look up, conspicuously not noticing MAN.

MAN

(leaning in, conspiratorially, quietly)

Hey.

(WOMAN does not respond; louder)

Hey.

(WOMAN does not respond; louder, more insistent)

Hey!

WOMAN

(not looking up)

No.

MAN

(leaning back, faux shock)

"No," what?

WOMAN

(still not looking up)

No, you cannot borrow my notes.
You should come to class and take your own notes.

MAN

(holding one finger up)

No, wait, wait, there's a reason
I couldn't make it to class!

WOMAN

(looking up)

There's always a *reason.* You drank too much the night before. You didn't set your alarm. There was a... a... pizza-eating contest. I dunno. There is always a reason.

(pointing)

Just not a very good one.

(turns attention back to notes)

MAN

Sooo... what would be a good one?

WOMAN

(looking back up)

Oh, for example... "My train was delayed."

(looking back down)

MAN

(beat)

Is it too late to say "My train was delayed?"

WOMAN

(laughs quietly, shaking head, slides notebook over to MAN)

Here. Just give them back.

MAN

(beaming)

Thanks!

FADE OUT

[Explicit]

3) Why couldn't the man attend class?（男性はなぜ授業に出席しなかったのですか。）

a) It is not stated.（言及無し。）

b) He drank too much the night before.（前の晩、飲み過ぎたから。）

c) He forgot to set his alarm.（目覚まし時計をかけ忘れたから。）

d) His train was delayed.（乗っていた電車が遅れたから。）

4. CONCERT

FADE IN

INT. FOYER

MAN is sitting; WOMAN enters with a flyer in hand.

WOMAN

(showing MAN the flyer)

Hey look, there's a concert tomorrow. Why don't we go?

MAN

(frowning)

Tomorrow?

(looking at poster)

But we have to work tomorrow. What would we tell the boss?

WOMAN

Why don't we just call in sick? Everybody does that.

MAN

(looking uncomfortable)

Yeah, maybe, but *I* don't do that. And besides, I am a terrible liar.

WOMAN

Oh, come on! It's only a *little* lie.

MAN

(shaking head)

Won't it be strange if we're both sick on the same day?

WOMAN

(flopping hands around as she thinks)

Okay, how about this:

(pointing)

You call in sick, and I say that
I have a family emergency.

(lean back, sure of herself)

MAN

(beat; sternly)

***That* is really dishonest...**

(conspiratorially, raising eyebrows)

but it just might work...

FADE OUT

[Implicit]

4) What will the man most likely do after this?（この後、男性は一番何をしそうですか。）

a) call in sick（病気で仕事を休む。）

b) say he has a family emergency（家族の緊急の用事があると言う。）

c) refuse to go to the concert（コンサートに行くことを断る。）

d) go to work as normal（いつも通り仕事に行く。）

5. DOUBLE BOOKING

FADE IN

INT. LECTURE HALL FOYER

MAN is seated. WOMAN enters and plops down next to him, looking troubled.

MAN

Something wrong?

WOMAN

Oh, yesterday I made an appointment at a spa for Saturday. I really need to relax.

MAN

That's true... So what's the problem?

WOMAN

The problem is that last month I told my friend that I'd go to a fancy restaurant with her that day, and then forgot all about it.

MAN

So just cancel the spa.

WOMAN

It's for tomorrow. I'll have to pay a cancellation fee.

MAN

Well, that may be, but your friend is probably looking forward to your dinner. You can't cancel on her now.

WOMAN

Yeah, I suppose...

(pout)

FADE OUT

[Explicit]

5) What is the man's advice?（男性のアドバイスは何ですか。）

a) The woman should cancel the spa.（スパをキャンセルするべき。）

b) The woman should cancel the dinner.（夕食をキャンセルするべき。）

c) The woman should relax.（リラックスするべき。）

d) The woman should call her friend.（友達に電話をするべき。）

6. OLD FLAME

FADE IN:

INT. CAFÉ

MAN and WOMAN sit with coffee cups.

WOMAN

(leaning forward)

Hey, can I talk to you about something?

MAN

(concerned, leaning forward)

Sure. What's wrong?

WOMAN

You may think this is strange, especially after all this time, but I think I've started to have feelings for my ex-boyfriend.

(look down, a bit ashamed)

I'm kind of shocked at myself!

MAN

(leaning back)

Well, I'm not that surprised, actually. You two were a great couple.

(shaking head)

It wouldn't be that bad if you got back together again.

WOMAN

(looking relieved)

You think so?

(leaning back)

Thanks, I really needed to hear that.

(leaning back in, lowering voice)

But I'm still kind of scared about telling him my real feelings, so

(raising eyebrows)

could we keep this between us for now?

MAN

(slight smile, gesture zipping lip)

My lips are sealed.

FADE OUT

[Explicit]

6) What does the woman want the man to do?（女性は男性に何をしてもらいたいですか。）

a) keep her secret（秘密にして欲しい。）

b) talk to her ex-boyfriend（元彼氏に話して欲しい。）

c) comfort her（慰めて欲しい。）

d) ask her on a date（デートに誘って欲しい。）

7. SMOKING

FADE IN

INT. CAFÉ

MAN and WOMAN are sitting at a table.

(MAN pulls pack of cigarettes from his bag)

WOMAN

Still smoking, huh?

MAN

Yeah, well, I'd like to quit but it's not easy.

WOMAN

Believe me, I know. I used to smoke too, but I was finally able to quit.

MAN

How?

WOMAN

Well, I tried all the normal things: gum, patches... nothing. And then I found the secret.

MAN

What's the secret?

WOMAN

You have to *want* to quit. You have to think, "I do not want to do this anymore." And then you just... *quit.*

MAN

(scoff)

Yeah, there's no hope for me, then.

FADE OUT

[Implicit]

7) What does the man think at the end?（最後に男性はどのように思いましたか。）

a) He cannot quit smoking.（タバコをやめられない。）

b) He will take the woman’s advice.（女性のアドバイスを受け入れる。）

c) He should start by smoking less each day.（毎日少しずつ吸う事を減らし始める。）

d) He wants to know the woman’s secret.（女性が禁煙出来た秘訣を知りたい。）

8. ATTIRE

FADE IN

INT. HALLWAY

MAN and WOMAN meet in hall. Both are wearing conference name badges and carrying papers.

WOMAN

(reading)

Hey, did you see that the invitation for the reception tonight says to dress "casually?" To me, "casually" means jeans and a T-shirt.

(looks at MAN)

What do you think it means here?

MAN

(uneasy)

Yeah, what *does* that mean? Jeans? Shorts? Polo shirt and khakis?

(palms upturned)

What?

WOMAN

Right? So what do we do?

MAN

You know what I'm going to do?

(pointing gesture)

Shirt and a tie.

(palms up)

That way, I can take the tie off if it's too formal.

WOMAN

Well that's fine for you, but what about me?

MAN

Yeah, I guess you're on your own there.

WOMAN

(sarcastically)

Gee, thanks.

FADE OUT

[Implicit]

8) What does the woman think at the end of the conversation?（会話の後、女性はどのように思いましたか。）

a) the man was not helpful（男性の助言は役に立たなかった。）

b) she knows what to wear（何を着るか分かった。）

c) she has to go to the reception alone（一人で宴会に行かなければならない。）

d) she will wear jeans and a t-shirt（ジーンズと T シャツを着るつもりである。）

9. PRESENTATION CONSTERNATION

FADE IN

INT. FOYER

MAN wears conference name badge on a lanyard, holding some papers, looking at a schedule, nervously. WOMAN enters, also wearing a badge.

WOMAN

Nice conference, huh?

MAN

Yeah, I guess. Hey, don't you have to give your presentation today?

(pointing at the schedule)

Yeah, your name's right here.

(looking intensely at WOMAN)

Oh my God, aren't you nervous?

WOMAN

Maybe a little bit, but I'm more excited, really. I love to stand up and tell people what I've done; it's fun! Yours is tomorrow, right?

MAN

Yeah. It's my first time presenting at a conference.

(stressed/nervous voice)

I'm terrible at things like this! How do you deal with it?

WOMAN

Oh, don't worry; once you stand up and start speaking, you'll be fine.

MAN

(shaking head, disbelieving)

I sure hope you're right...

FADE OUT

[Implicit]

9) How does the man feel after talking to the woman?（女性と話した後、男性はどのように感じていますか。）

a) nervous（緊張している。）

b) calm（落ち着いている。）

c) excited（興奮している。）

d) bored（退屈している。）

10. FRIEND CALLING

FADE IN

INT. CAFÉ

MAN and WOMAN sit across from each other, chatting. WOMAN'S phone rings. WOMAN pulls it from her pocket, looks, sighs, cancels the call.

MAN

(pointing casually)

You're not taking the call? Who was it?

WOMAN

(looking conflicted, putting phone away)

No, it's just some old friend from high school. She always finds the worst time to call.

MAN

I don't mind if you take a call for a few minutes.

WOMAN

That's just it; it won't *be* a few minutes. First she'll ask what I'm doing, and then I'll tell her, and then she'll sigh. I'll spend the next ten minutes trying to find out what's wrong with her, and then it'll be some stupid boyfriend problem or something. I swear, she's like a middle school student.

MAN

Maybe you should have talked to her. She sounds like she has a lot of problems.

WOMAN

(sassy)

More like she *is* a lot of problems.

FADE OUT

[Explicit]

10) What does the man think the woman should do?（男性は、女性はどうするべきだと思っていますか。）

a) talk to her friend（友達と話す。）

b) ignore the call（電話を無視する。）

c) call her friend back（友達を心配することをやめる。）

d) stop talking to her entirely（友達との関係をやめる。）

11. LOST ITEM

FADE IN

INT. CLASSROOM (set to only see one table, with wall behind)

Table has a sign saying "Lost and Found." WOMAN sits behind table. Behind her are a pile of backpacks, etc. MAN enters.

MAN

(worried)

Hi, is this the lost and found table?

WOMAN

(bubbly)

Sure is!

MAN

Great. Um, I'm looking for my bag? It's a white briefcase that you can wear like a backpack.

WOMAN

Oh! I know what you mean. Is this it?

(holds up a white backpack)

MAN

Um, no. That's just a white rucksack? Mine's more like a briefcase?

WOMAN

Oh! That one! This is it, right?

(holds up a black briefcase)

MAN

(frustrated, raising voice a little)

No, that's just a regular briefcase, and it's not even white! It's black!

WOMAN

(scolding, pointing)

Hey hey hey! I'm trying to help you here, mister!

MAN

(sigh, resignation)

You're right; I'm sorry. I'll keep looking.

FADE OUT

[Explicit]

11) What is the man looking for?（男性は何を探していますか。）

a) a white briefcase（白のブリーフケース）

b) a white rucksack（白のリュックサック）

c) a black briefcase（黒のブリーフケース）

d) a black rucksack（黒のリュックサック）

12. REGISTRATION

FADE IN

INT. CLASSROOM (set to only see one table, with wall behind)

Table has a sign that says "Registration." There are name badges laid out in a grid on top. MAN stands behind it. WOMAN walks up.

MAN

Welcome to the conference! Have you already registered online?

WOMAN

I tried to, but your webpage wouldn't take my card.

MAN

I'm sorry to hear that! You can register now and pay in cash, though, no problem.

WOMAN

Um, okay, but how much is it?

MAN

Registering onsite is 230 US dollars.

WOMAN

(beat)

That's thirty dollars more than the online rate!

MAN

I'm afraid so.

WOMAN

(raising voice a little)

I shouldn't have to pay the onsite rate just because your web page was broken!

MAN

(nodding)

That's a good point. Let's just do 200 then.

WOMAN

(still exasperated)

Thank you!

FADE OUT

[Explicit]

12) Why didn't the woman pay online?（女性はなぜオンラインで払いませんでしたか。）

a) There was a problem with the website.（ウェブサイトに問題があったから。）

b) She did not have time.（時間がなかったから。）

c) She did not realize it was more expensive onsite.（現地で払う方が高いということを知らなかったから。）

d) Her credit card was maxed.（クレジットカードの限度額が超えていたから。）

13. THE PARTY

FADE IN:

INT. HALLWAY

WOMAN is walking toward camera, carrying a shoulder bag; MAN catches up from behind.

MAN

(calling after WOMAN as approaching)

Hey! Wait up!

(catches up)

Hey, are you going to Yoko's party on Saturday?

WOMAN

I am.

MAN

(hands forward)

What're you bringing?

WOMAN

I was going to bring some homemade cookies. You?

MAN

I'm thinking fruit.

(pointing)

Oh, and beer. She never has any.

WOMAN

I was originally planning on bringing stuff to make cocktails, but I don't have time to go shopping.

MAN

Y'know, that's probably for the best; most of the people going don't even drink.

WOMAN

Oh! I didn't even think of that.

(nodding in emphasis)

Good thing I didn't spend any money on it, then.

FADE OUT

[Explicit]

13) What is the woman bringing to the party?（女性はパーティーに何を持って行きますか。）

a) cookies（クッキー）

b) cocktails（カクテル）

c) beer（ビール）

d) fruit（果物）

14. MYSTERIOUS KEY

FADE IN:

INT. LECTURE HALL FOYER

MAN and WOMAN sit next to each other. MAN is wearing a light jacket (because it'll be sweltering in September). MAN reaches in his pocket and pulls out a small key.

MAN

(turning the key in his hand)

Huh.

(showing the key to WOMAN)

Hey, take a look at this key I just found in my pocket. What do you think it is?

WOMAN

(takes it)

I dunno...

(turning it in her hand)

It kind of looks like the key to a bike lock.

MAN

Yeah, except I don't have a bike.

(takes the key back, looks at it more)

Maybe it's the key to a desk or something... Wonder how long it was in there?

WOMAN

When's the last time you wore this sweatshirt?

MAN

I dunno... a long time ago. Like... five years?

WOMAN

Did you have a bike then?

MAN

No, but my roommate used to lend me his until one day when I...

(look of recognition)

WOMAN

(curious frown)

When you what?

MAN

(sheepishly)

When I lost the key.

WOMAN

Well, it looks like you found it.

(scoff)

FADE OUT

[Implicit]

14) What is the key most likely for?（何の鍵の可能性が一番高いですか。）

a) The roommate's bicycle.（ルームメイトの自転車）

b) The man's bicycle.（男性の自転車）

c) A desk.（机）

d) A motorcycle.（オートバイ）

15. WHOSE NUMBER IS THIS?

FADE IN:

INT. CAFÉ

MAN and WOMAN sit across from each other, coffee cups in front of them; they are lightly holding hands on top of the table. Clearly a couple.

WOMAN

Oh! Hey, do you have the number for the restaurant we're having our anniversary dinner at?

MAN

(lets go of her hand, gets his phone out)

Oh yeah. It's in the call history.

(hands WOMAN the phone)

WOMAN

(swipes to enter the phone and begins looking through the call list)

Thanks; I just want to confirm our reservation for Friday...

(drifts off, face goes serious)

MAN

Something wrong?

WOMAN

Who called you at 1AM?

MAN

Sorry?

WOMAN

(looking up from phone)

There's a call here at 1AM?

MAN

(very uncomfortable; stammering)

Uhhh... I don't... I don't know.

WOMAN

You talked for 2 minutes.

(scrolls down, is shocked, angered)

And here it is again, two days ago! You were on the phone for over an hour when I was on that business trip!

(locks eyes with MAN, demanding, staccato delivery)

Whose number is this?

MAN

(looks uneasy)

FADE OUT

[Imp]

15) Why is the woman upset?（女性はなぜ気分を害していますか。）

a) She thinks the man is having an affair.（男性は浮気していると思っているから。）

b) She thinks the man forgot to call the restaurant.（男性はレストランに電話をかけ忘れたと思っているから。）

c) She thinks the man has spent too much time on the phone.（男性は電話に時間をかけ過ぎていると思っているから。）

d) She thinks the man did not invite her on his trip.（男性は女性を旅行に誘わなかったと思っているから。）

16. CRAFTS FAIR

FADE IN:

INT. HALLWAY

MAN and WOMAN walk together with bags.

MAN

Doing anything interesting this weekend?

WOMAN

Yeah, I'm going to a craft fair.

MAN

(stop walking)

A what?

WOMAN

(stop, turn to MAN)

A crafts fair. You know, people selling handmade jewelry and knickknacks and stuff.

MAN

(nodding)

Yeah, I know what a craft fair *is*.

(shaking head)

I just can't picture *you* at one.

WOMAN

(frowning, shrugging)

Why?

MAN

(scoffs, palms upturned)

You hate stuff like that!

(beat)

C'mon now, why are you really going?

WOMAN

(self-satisfied)

I'm just going with some new friends. They go every week.

MAN

(unconvinced, arms crossed)

And you're just pretending to be interested.

WOMAN

(defensive)

So what if I am? I think it's good to make new friends.

(BOTH continue walking)

FADE OUT.

[Explicit]

16) Why is the woman going to the craft fair?（女性はなぜ手工芸フェアに行くのですか。）

a) To impress some friends.（友達に良い印象を与えるため。）

b) To buy some jewelry.（アクセサリーを買うため。）

c) She goes every week.（毎週行っているため。）

d) To take pictures.（写真を撮るため。）

17. PARTY AFTERCARE

FADE IN:

INT. LECTURE HALL FOYER

MAN and WOMAN sit on a bench.

WOMAN

Hey, do you know if Professor Kimura was okay after the party at my place the other night?

MAN

Totally! He was in really high spirits. We talked all the way back to the station.

WOMAN

That's good. It's just that he hasn't replied to an email I sent, so I was worried that he maybe didn't have a good time or felt sick from my food or something!

MAN

(shaking head)

No no no no no. He was very happy with the food. He seemed to have had a really good time.

WOMAN

Phew, that's a relief. It's just that he's usually really good about writing back right away so I guess I just wondered.

MAN

Nah, you're just thinking too much. I really wouldn't worry about it. He's probably just busy.

FADE OUT

[Explicit]

17) Why is the woman worried about the professor?（女性はなぜ教授のことが気になっていますか。）

a) He has not replied to her email.（教授がまだ女性からメールに返事をしていないこと。）

b) He got sick from her food.（教授が女性の作った食べ物で病気になったこと。）

c) He had a bad time at her party.（教授が女性の開いたパーティーで、楽しくなかったこと。）

d) No one has talked to him since the party.（パーティー以来、誰も教授と話をしていないこと。）

18. THE RIDE BACK

FADE IN:

INT. CLASSROOM, SPEAKER'S LECTERN

WOMAN is attaching a laptop to a projector. MAN approaches.

MAN

Dr. Brandt? I'm sorry to bother you while you're preparing for your talk, but I wanted to see if you needed me to call you a taxi to take you back to your hotel afterwards?

WOMAN

(looking up)

What? Oh, is it far?

MAN

Not really. It's about a 10-minute walk from here.

WOMAN

(dismissively)

Well, oh, okay. That's fine, then.

MAN

I'm sorry?

WOMAN

(short, irritated)

I'll just walk.

MAN

Right.

(turns to leave, then remembers and quickly turns back to WOMAN)

Oh, and will you need a taxi to the airport tomorrow, or are you taking the shuttle bus?—

WOMAN

(stops what she's doing, turns and looks squarely at MAN)

—You know what? Can we maybe talk about this later?

MAN

(hand up, apologetically)

Oh, right. I'm sorry.

FADE OUT

[Implicit]

18) Why does the woman ask to finish the conversation later?（女性はなぜこの話を後でしたいと頼みましたか。）

a) She is busy right now.（今は忙しいから。）

b) She doesn't know the answer to the man's question.（男性の質問に対する答えを知らないから。）

c) She doesn't understand the question.（質問の意味が分からないから。）

d) She doesn't know when she is leaving.（いつ帰るか分からないから。）

19. THE PLAY

FADE IN:

INT. CAFÉ

MAN and WOMAN sit across from each other; BOTH are looking at their smartphones.

MAN

(still looking at phone)

Oh yeah, I forgot to tell you. I made plans to play golf with some people from work tomorrow.

WOMAN

(still looking at phone)

Uh, okay, but don't forget that we have play tickets for tomorrow...

MAN

(looking up from phone)

Oh yeah! The play. What time was that again?

WOMAN

(exasperated, looking up from the phone)

I've told you over and over! It's 8 PM.

MAN

(adding item to phone)

Okay, got it. Mm, I should be able to get back from golf before then. We're starting at 4:30!

WOMAN

(looking, snapping at him)

What? You need to be back *before* 8, not *at* 8! It'll take you at

least an hour to clean up, and we have to be there no later than 7:30. *Plus* it's an hour's drive! There's no way you'll make it back in time!

FADE OUT

[Implicit]

19) Why is the woman upset?（女性は何に不満がありますか。）

a) The man is bad at scheduling.（男性が時間の予定を立てるのが下手だから。）

b) The man is always playing golf.（男性がいつもゴルフをしているから。）

c) The man is ignoring her.（男性は女性を無視しているから。）

d) The man did not invite her to play with him.（男性は女性を演劇に誘わなかったから。）

20. JOB PROSPECTS

FADE IN:

INT. CAFÉ

MAN and WOMAN sit with coffee cups.

WOMAN

Have you given any thought to applying for that position I forwarded to you?

MAN

(wincing, scratching back of head, leaning back)

Yeah, I thought about it, but

(leaning forward again, shaking head)

I don't think I'm going to apply. I looked into it, and the pay isn't even as good as what I'm making now.

WOMAN

That may be true, but that company is growing fast. I think you could be promoted pretty quickly there.

(looking concerned)

In a year or two, you could easily be making more than you are here.

(leaning back)

I think you should reconsider.

MAN

(looking conflicted/unsure)

Yeah...

(changing mind)

I suppose there's no harm in applying just to see how it goes.

(palms upward)

You never know, maybe I won't even get an interview.

FADE OUT

[Explicit]

20) Why hasn't the man applied for the other position?（男性はなぜ他のポジションに申し込みをしていませんか。）

a) It does not pay as much as his current salary.（男性の現在の給料ほどもらえないから。）

b) He does not feel that he is qualified.（男性に申し込む資格がないと思っているから。）

c) The job sounds stressful.（ストレスが多そうな仕事だから。）

d) He is expecting a promotion from his current position.（現在のポジションからの昇進を期待しているから。）

Appendix D.

Filming Checklist and Procedure

Filming Shot Setup

Camera

- ☐ SD card inserted
- ☐ Battery remaining
- ☐ Shutter open
- ☐ Camera mounted
- ☐ Mic levels appropriate
  (adjust if necessary on the preamp on top of the camera—must adjust both sides)

Lapel Mics

- ☐ Mic cables cleaned
- ☐ Mics attached to body packs
- ☐ Body packs switched on
  (Blinking light indicates dead battery)
- ☐ Receiver antennas raised
- ☐ Receiver plugged into AC
- ☐ Outputs connected to preamp
  (make note of label for A)

Mic Preamp

- ☐ Preamp plugged into AC
- ☐ Powered on with switch
- ☐ Inputs from wireless receiver attached
  (make note of label for A)
- ☐ Outputs from preamp attached to digital audio recorder
  (make note of label for L)
- ☐ Mic levels appropriate
  (adjust with Gain and Output as necessary)

Digital Audio Recorder

- ☐ SD card inserted
- ☐ Battery remaining
- ☐ Set to 4CH recording
- ☐ Internal mics flipped inward
- ☐ Mounted on tripod below camera
- ☐ Press Record once to enable monitoring
- ☐ Attach headphones
- ☐ Levels for internal mic appropriate
  (adjust if necessary by pressing the 1/2 button and using the Input Level buttons on the side)
- ☐ Levels for external mics appropriate
  (adjust if necessary by pressing the 3/4 button once for L and twice for R and using the Input Level buttons on the side)

Lights

- ☐ Height/angle appropriate
- ☐ Light levels appropriate
  (adjust with + and – buttons)

Procedure

Preparation

1. Set setting props (tables, etc.)
2. Set camera, lights, audio (beware backlighting)
3. Locate and distribute hand props
4. Bring actors in to fine-tune light and framing
5. Spike the floor (if necessary)
6. Level check (press Record once; see checklist)
7. Rehearse 1-2 times

Recording

1. Start audio recorder by pressing "Record" (it should already be flashing for monitoring)
2. Start video camera
3. "In five..."
4. Start stopwatch
5. Run scene
6. Stop stopwatch
7. Stop video camera
8. Stop audio recorder
9. Re-take as necessary
10. Watch scene for check
11. Repeat for other scenes in the location

Finishing

1. 10 seconds of room tone
2. Collapse as little of equipment as possible to move to the next location