Georgia State University
ScholarWorks @ Georgia State University

Applied Linguistics and English as a Second Language Dissertations

Department of Applied Linguistics and English as a Second Language

1-6-2017

Locus of Control in L2 English Listening Assessment

Sarah J. Goodwin
Georgia State University

Recommended Citation

Goodwin, Sarah J., "Locus of Control in L2 English Listening Assessment." Dissertation, Georgia State University, 2017.
https://scholarworks.gsu.edu/alesl_diss/37

LOCUS OF CONTROL IN L2 ENGLISH LISTENING ASSESSMENT

by

SARAH GOODWIN

Under the Direction of Sara Cushing, Ph.D.

ABSTRACT

In second language (L2) listening assessment, various factors have the potential to impact the validity of listening test items (Brindley & Slatyer, 2002; Buck & Tatsuoka, 1998; Freedle & Kostin, 1999; Nissan, DeVincenzi, & Tang, 1996; Read, 2002; Shohamy & Inbar, 1991). One relatively unexplored area to date is who controls the aural input. In traditional standardized listening tests, an administrator-controlled recording is played once or twice. In real-world or classroom listening, however, listeners can sometimes request repetition or clarification. Allowing listeners to control the aural input thus has the potential to add test authenticity but requires careful design of the input and expected response as well as an appropriate computer interface. However, if candidates feel less anxious, allowing control of listening input may enhance examinees' experience and still reflect their listening proficiency. Comparing traditional and self-paced (i.e., examinees having the opportunity to start, stop, and move the audio position) delivery of multiple-choice comprehension items, my research inquiry is whether self-paced listening can be a sufficiently reliable and valid measure of examinees' listening ability.

Data were gathered from 100 prospective and current university ESL students. They were administered computer-based multiple-choice listening tests: 10 identical once-played items, followed by 33 items in three different conditions: 1) administrator-paced input with no audio player visible, 2) self-paced with a short time limit, and 3) self-paced with a longer time limit. Many-facet Rasch (1960/1980) modeling was used to compare the difficulty and discrimination of the items across conditions. Results indicated items on average were similar difficulty overall but discriminated best in self-paced conditions. Furthermore, the vast majority of examinees reported they preferred self-paced listening. The quantitative results were complemented by follow-up stimulated recall interviews with eight participants who took 22 additional test items using screen capture software to explore whether and when they paused and/or repeated the input. Frequency of and reasons for self-pacing did not follow any particular pattern by proficiency level. Examinees tended to play more than once but not two full times through, even without limited time. Implications for listening instruction and classroom assessment, as well as standardized testing, are discussed.

INDEX WORDS: Language assessment, Assessing listening, Second language testing, Rasch analysis, Audio play control, Stimulated recall in L2 listening

LOCUS OF CONTROL IN L2 ENGLISH LISTENING ASSESSMENT

by

SARAH GOODWIN

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2016

LOCUS OF CONTROL IN L2 ENGLISH LISTENING ASSESSMENT

by

SARAH GOODWIN

Committee Chair: Sara Cushing

Committee: YouJin Kim

Stephanie Lindemann

John Murphy

Electronic Version Approved:

Summer 2016

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2016

DEDICATION

To my partner Matt, my mom and dad, and my sister Emily

ACKNOWLEDGMENTS

A dissertation can seem like a solo venture in the end, but really it is a result of the time and efforts of many people I have interacted with throughout my academic and professional career.

The University of Michigan is where I first connected with many wonderful scholars. Thank you to Lisa Del Torto and John Lawler for sparking my interest in linguistics as an undergraduate student. Thanks go to Aaron Ohlrogge for training me to score English language learners' compositions, Sunny Park-Johnson for helping teach me to develop items, and Barbara Dobson, Jeff Johnson, and Amy Yamashiro for hiring me as a language testing research assistant; you all contributed to my initial understanding of English language assessment. Spiros Papageorgiou, I appreciate you teaching me about Rasch analysis and its use in testing listening, as well as being immensely supportive of my graduate studies. Jayanti Banerjee, you reminded me that a dissertation is a marathon, not a sprint; I admire your kind spirit and work ethic. Marianne Bailey, Sarah Briggs, Natalie Nordby Chen, Judy Dyer, Chris Feak, Julia Salehzadeh, and many others from the ELI and CaMLA: thank you for cheering me on in my pursuits.

Thank you to the graduate admissions committee in GSU's Department of Applied Linguistics and ESL who accepted me as a new M.A. student in 2010. I want to thank Sara Cushing, my advisor and dissertation chair, for your encouragement, patience, and support. My other three dissertation committee members also devoted much of their time and care to this project. YouJin Kim, you gave me extensive and critical yet supportive feedback on a Second Language Acquisition course project that was the seedling of this dissertation. Stephanie Lindemann and John Murphy, your research on second language listening, as well as your teachings in applied linguistics, have impacted my instruction and assessment pursuits.

Thank you to many other AL faculty members: Kris Acheson-Clair, for offering insightful remarks that shed more light on my understanding of qualitative research; Diane Belcher, for giving me excellent feedback on my writing; Viviana Cortes, Scott Crossley, and Eric Friginal, for keeping me sane and offering great professional advice and friendship; and Ute Römer, for being a cheerful and compassionate colleague in Ann Arbor and Atlanta. I also would be remiss if I didn't thank the AL, Intensive English Program, and ESL Credit Program faculty and staff members, including but not limited to John Bunting, Susan Coleman, Janie Hardman, Sarah Kegley, Doreen Kincaid, Kim Kleiber, Margareta Larsson, Amanda Starrick, and Diana Wrenn, for their humor, wisdom, and down-to-earth reminders. Debra Snell, thank you for two years' lodging in your wonderful home with a beautiful eastern exposure at which to work.

Thanks go to Robin Cathey, Dave Chiesa, Nia Kapitanova, Minkyung Kim, Jessica Lian, Rurik Tywoniw, and Kátia Monteiro Vanderbilt for helping me with the data. For feeding me with thoughts and food, offering me transportation, inspiring me, and bouncing ideas off of, I also thank all of the GSU AL grad students I've interacted with. Joe Lee, WeiWei Yang, Nur Yiğitoğlu, Meg Montee, Audrey Roberson, Jack Hardy, Cassie Leymarie, Merideth Hoagland, Kris Kyle, Nicole Pettitt, Cindy Berger, Stephen Skalicky, Justin Cubilo, Ju A Hwang, Janet Beth Randall, and many others: I wish you all the best in your careers and lives. Caroline Machado and Gharbeela Sami: a special shout-out for the great gastronomic times. Pam Pearson, thank you for dissertating with me at Aurora and San Francisco Coffee and for all the laughs.

I give thanks to John Field, Larry Vandergrift, and Elvis Wagner, whose writings on second language listening influenced my work tremendously.

Last, but definitely not least, I want to thank deeply and give appreciation to all of the language learners who participated in this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The present project concerns second language (L2) listening assessment in a standardized test situation, focusing on putting the control (playing, pausing, and audio position) of the input in the hands of examinees. With this research, my hope is to contribute to the current body of knowledge in adult L2 listening assessment, along with offering suggestions and future directions for the teaching and learning of L2 listening. In the design of listening assessments, I intend to, as Merrill Swain calls it, *bias for the best* (Swain, 1983; as cited in Fox, 2004), engaging test takers so that they perform at their highest level possible, in order to be able to show their listening ability as best they can.

In most standardized listening tests, learners are only able to listen once to input text, controlled by a test administrator. In such listening constructs, or underlying listening abilities the test purports to tap, automatic processing is expected, so it is assumed that one play of the audio input is sufficient. Once-heard listening is often standard in situations such as listening to a lecture, in which people are expected to be able to comprehend without requesting clarification, questioning, or asking for repetition (Flowerdew, 1994; Sawaki & Nissan, 2009).

In some cases, however, in real-life listening as well as in English for academic purposes (EAP) instruction, there is often the chance to listen more than once, particularly in interactive listening. Students have some control over listening because they can sometimes request that a speaker repeat or paraphrase. If students feel that they want to listen a second time or more, they sometimes have the opportunity to do so in the real world. If they feel confident enough that they understand what they have heard, one hearing may be enough. Listening once or multiple times impacts listening assessment design, because test developers must balance the design of academic listening tests with the content they cover. Test listening should be aligned as much as possible with types of listening in the target language use (TLU) domain. For predictive validity of an assessment, researchers determine to what extent the results of a particular assessment can predict examinees' future performance in the TLU domain (Bachman & Palmer, 2010). If predictive validity can be established for an item set type, this, along with other types of validity, may merit that set's inclusion in a listening test. For instance, an examinee's performance on an academic listening test might be said to be predictive of her listening ability in college courses.

Why put the control of listening input in the hands of the listener? For L2 learning, the transitory nature of listening may create stress and anxiety for learners, who must process input quickly (Arnold, 2000; Graham, 2006). Thus, by being offered a choice in how to engage with listening input, examinees' cognitive load as well as their anxiety may be diminished. Because listening is a skill in which a learner holds less of the locus of control than in any of the other three language skill areas of reading, speaking, or writing (Norton Peirce, Swain, & Hart, 1993), test takers may feel more agency in the listening process if they are able to have control over the stream of sound. Some learners may take control of the audio play, while others may feel that they do not need pauses or that listening once is sufficient.

However, allowing user control with pausing and/or multiple plays may create issues for exam design. First, such a listening test is only feasible when delivered via computing devices, so test centers must have the appropriate technology resources. Additionally, there are practical considerations such as reasonable time permitted for the test sections. Perhaps most importantly, the reliability and validity of the test should not be diminished in a self-paced setting. These features of test design and administration are critical to the creation of listening proficiency exams. With these considerations in mind, empirical research is needed that examines what happens when a self-paced setup is part of a listening test. This dissertation is an initial step to providing this experimental data.

The manuscript is structured as follows. Chapter 2 provides background information about work in five prior research areas: three theoretical (features of spoken language, technology in listening assessment, and who holds control in learning) and two methodological (many-facet Rasch measurement and stimulated recall interview analysis). Chapter 3 introduces the research questions and methodology. In Chapter 4, I present the results, and in Chapter 5, I provide implications of the findings and future directions for research.

2 LITERATURE REVIEW

The section reviewing previous work in the areas of interest is divided into five main components. First, I will discuss features relating to the nature of aural (versus written) language; secondly, I focus on factors affecting listening comprehension tests when technology is considered. The third subsection describes research examining what happens when learners are able to exercise more control in educational situations. The fourth and fifth areas review existing methodology in listening assessment: Rasch measurement in language assessment investigations, and stimulated recall interviews in L2 listening studies. My hope with this dissertation is to connect these areas, as I was particularly interested in what happens when examinees are given the option to self-pace the aural input in a mock standardized test setting, and I wished to address this not only through quantitative (Rasch measurement) but also qualitative (stimulated recall interview) methods. Within each portion, I discuss connections with assessment based on past investigations.

2.1 The nature of listening

2.1.1 Features of spoken language

This section focuses on speech, particularly its similarities with and differences from writing. When considering the TLU domain for listening in academic settings, it may be useful to compare listening and reading constructs, as they pertain to how researchers have conceptualized the acquisition of spoken language and written language. Especially in EAP situations, listening and reading are closely connected because they both not only involve language processing but also are intertwined in students' real-world experiences. Both listening and reading require understanding of the lexical, morphosyntactic, pragmatic, and discoursal features of the input. They both utilize people's experience with previous texts, or what has been read or said already, as well as their world knowledge or background knowledge awareness (Goh, 2000). In addition, academic lectures and discussions are linked to course readings. Instructors can use lecture and discussion opportunities to give context for, provide details about, and critique the content of various reading and listening materials (Murphy, 1996). Reading and listening do share many similarities.

However, speech differs from writing in particular ways. Spoken texts are fast, continuous, and variable (Cutler, 2012); speech may also contain false starts, repetitions, or hesitations. Even in slowly-spoken utterances, there is little pausing at word boundaries unless the speaker is doing so for clarity or emphasis. Spoken language involves blended or reduced forms such as "Whaddayuwannaeat?" (Brown, 2011, p. 4) for "what do you want to eat?", though written language does occasionally include contractions or the use of nonstandard spelling to show interlocutors' accent or variety of speech. A successful listener must be able to comprehend such blendings or reductions as well as perhaps a word's citation form, its version when carefully pronounced in isolation. Moreover, knowing a word's printed form is not enough; listeners must also be able to segment incoming utterances into recognizable words and handle the fact that there may not be just one phonological representation of a word. To name just a few potential factors that can affect the variation in spoken forms, the diverse phonological forms may be due to a word's context in the utterance, the speed at which it is spoken, and/or the variety of speech of the speaker (Buck, 2001; Cutler, 2012; Cauldwell, 2014).

Unless read aloud from a written transcript or memorized from a print-based source, speech is generally not as structurally complex or as filled with long utterances as is writing (Biber et al., 1999). Even the consideration of a concept like the *sentence* is a notion particular to writing and not necessarily to speech (Chafe, 1985). Unlike in reading, there are no sentence boundaries or paragraph markings in speech, but, in English, skilled speakers give explicit markers to guide listeners through what they say. Academic lecture speech in particular contains microstructuring (filled or unfilled pauses or markers such as *ah*, *er*, *um*, *and*, *but*, *now*, *okay*, *so*) and, in video or face-to-face delivery, bodily or facial movements that add to the meaning of spoken utterances (Flowerdew & Miller, 1997). There are also macromarkers (Chaudron & Richards, 1986), when a speaker gives signposts to show what has been said or what is going to be said. Flowerdew and Miller give examples of this from a lecturer they observed, who used macromarkers such as "Okay, let's get started" and "now here/we'll put up our last slide/and come to the conclusions" (1997, p. 38). Such cues are intended to help listeners follow the discourse, but beginning L2 listeners have a great deal to contend with in the modality of listening.

A crucial element with respect to listening is that the speaker, not the listener, controls the rate of input (Goldman, Hogaboam, Bell, & Perfetti, 1980). The listener may suspend the *decoding* process temporarily in her mind but generally cannot directly pause the stream of speech itself. A spoken text is experienced in real time and thus fleeting. Unless a listener has assistive technology, she cannot review a text multiple times or skim-and-scan as she might when reading. I discuss number of plays in a later section, but there may be opportunities to hear listening input more than once in the real world. Test developers must thus consider whether the spoken language variables included on an exam are representative of the target real-world domain of language use in listening. I turn next to how we successfully process speech.

2.1.2 Processing of spoken language

It is important to consider what is happening in the mind when spoken language is being perceived. Prior to the early 1970s, the assumption was that listening in an L2 was the same as in one's native language, and that spoken and written language were processed similarly (Flowerdew & Miller, 2010); thus, instructional approaches did not focus on processing considerations specific to L2 listening. Later, in the 1980s and beyond, communicative methods in L2 listening instruction exerted a great deal of influence, but this meant that teachers' attention was directed toward making students aware of the *product* of listening, or comprehension approaches. As the 1990s and 2000s arrived, there was a rise in research and instructional methods that began to give more specific attention to the *process* of listening, which is a requisite for comprehension. This section focuses on what is needed for successful comprehension in English as one's additional language.

In the area of L2 English listening comprehension, L1 English listening processing is often used as a starting point for formulation of processing models. Assuming no severe hearing impairments, and a normal neurological system, people acquire their first language via oral input. Rost (2011) divided the elements of listening into linguistic processing, semantic processing, and pragmatic processing, for both first-language (L1) and L2 listening acquisition. Here, his use of "linguistic" processing refers to sound perception, word recognition, and syntactic parsing. Although I would argue that semantics and pragmatics are also language related, Rost (2011) theorized about them as separate branches in his model.

First, Rost (2011) notes, in linguistic processing, especially for adult L2 learners, competence in the L2 phonological system involves the ability to appropriately use lexical and metrical segmentation strategies. Lexical segmentation allows listeners to recognize words in the stream of speech. Metrical segmentation in a language like English that is trochaically timed, or composed of stressed-unstressed syllable patterns, is also important in aural perception (Rost, 2011). Cutler and Carter (1987), based on a corpus analysis, found that 85.6% of content words in running speech are monosyllabic or contain syllable-initial stress, which may help listeners cue into word boundaries. Syntactic and lexicogrammatical development for adult L2 learners are also areas in which a learner must learn to detect new forms in spoken input, often also having to deal with transfer or interference from the L1. Secondly, as far as semantic processing is concerned, Rost states that L2 listening may involve some cultural or individual differences in how we interpret our experiences in the world. It may be the case that some people do not realize that their reasoning and inferencing processes may not be the same in their L2 as they are in their L1. Rost's third consideration, pragmatic processing, involves identification of topic shifts, listenership cues and interactional markers (such as backchanneling), and other understandings of elements related to discourse style. All of these L2 processing elements require substantial input in the target language and a certain amount of motivation on the part of the listener.

Along with Rost's (2011) considerations of linguistic, semantic, and pragmatic processing in L2 listening, there are researchers such as Field (2004) who make connections to top-down versus bottom-up processing. Field (2004) noted that these terms are used in reading and listening research in order to distinguish information gained from contextual sources (top-down) from information gained from perceptual sources (bottom-up). He added:

> Greatly preferable are the terms *lower-level processing* (for decoding what is in the speech stream) and *higher-level processing* (for the building of meaning). However, they have to be used with some circumspection in discussions of L2 listening since they easily become confused with references to lower and higher levels of learner (indicating degree of competence in the target language). (Field, 2004, p. 364, emphasis in the original)

In addition, Field clarified that for his definition of top-down, when he mentioned contextual sources, those refer not just to real-world knowledge but also to "co-text" (Brown & Yule, 1983, p. 46), or information gained from earlier spoken or written content in the context of the utterance. Moreover, top-down and bottom-up processing are not alternatives but rather are in an interdependent, complex relationship, helping a listener form an interpretation of what she has heard (Field, 2004; Tsui & Fullilove, 1998).

The degree to which a listener may use a bottom-up approach more or less often than a top-down approach is dependent upon the purpose for listening. For example, listening for a specific detail in an utterance, such as for a particular number being recited to you, may involve more bottom-up processing. The idea of top-down versus bottom-up is connected to notions of controlled versus automatic processing. When an L2 listener has limited language knowledge, she is not able to automatically process everything that she hears. The person likely engages in controlled processing to be able to focus consciously on what she cannot process automatically. Ideally, eventually controlled processing becomes automatic; however, when conscious attention must be devoted to top-down or bottom-up processing or both, comprehension may suffer. This is because of memory, which plays an important role in listening processing (Vandergrift & Goh, 2012).

Memory, specifically working memory, has a role in the segmenting of meaningful units from the stream of speech. Working memory has a limited capacity before information disappears and new information has to be processed. The retained information is held in a phonological loop until the listener can segment it into words or meaningful chunks (Baddeley, 1986). As listeners' language proficiency increases, their ability to retain and process larger units also grows. Certain components of language become automatic, and listening becomes less of an effort (DeKeyser, 2001). The more familiar a unit of sound is to listeners, the more quickly they can draw on long-term memory to supply the previously acquired linguistic knowledge (Vandergrift & Goh, 2012).

A three-stage element of working memory specific to listening was described by Anderson (1995; 2009). Based on L1 processing theories, he differentiated listening comprehension into three interrelated phases: perceptual processing, parsing, and utilization. One can draw on these components in an integrated way, not necessarily in a sequence each time. Perceptual processing involves the decoding of the acoustic message by segmenting phonemes from the stream of speech. Parsing is the creation of a mental representation of how words are combined to form meaning, while utilization is the relating of that representation to existing knowledge or drawing inferences to complete the interpretation. Vandergrift and Goh (2012) took the concepts of perceptual processing, parsing, and utilization and extended them, showing how they are related to top-down and bottom-up processing; in top-down, utilization informs parsing, while in bottom-up, perception directly contributes to parsing. In other words, top-down processing results in information gained from contextual sources while bottom-up uses information gained from perceptual sources (Field, 2008).

One can imagine that, especially if listening to a long utterance, or needing to listen for various purposes (e.g., local and global ideas), both top-down and bottom-up processing would need to be occurring nearly simultaneously. Of special note with regard to L2 listening, particularly adults learning a new language, is the cognitive demand on the listener. Beginner L2 listeners often try to first perceptually process, extracting information word-by-word, and if that does not work, they may give up (Brown, 2011). Goh (2000) compared verbal reports of two groups of learners: listeners in a higher-ability group (TOEFL Paper-Based Test [pBT] scores of approximately 550 to 600) and a lower-ability group (TOEFL pBT 440-500). Examinees at higher- and lower-proficiency levels both had difficulties recognizing words they knew or quickly forgetting what was heard. However, there were some differences between the two groups: lower-proficiency learners reported they did not hear the next part of a text because they were thinking about what they had just heard; higher-proficiency learners said they understood words but not their intended message. Thus, lower-level learners' difficulties generally stemmed from the perceptual processing and parsing phases, with unsuccessful word recognition, while higher-level learners had more success processing and parsing but not utilizing what they had heard, comprehending the words but not their meaning in a particular context. These findings suggest that language learners may have comprehension troubles relating to some or all three of the phases described by Anderson (1995; 2009): perceptual processing, parsing, or utilization.

Some L2 processing models do not always appear to draw a clear distinction between processes related to decoding (sound and word recognition) and processes relating to meaning building (the listener bringing in outside knowledge to enrich understanding of what has been decoded). Field (2008) noted that the first process of the two, decoding, may require a slight shift from decoding in the L1, requiring the L2 learner to decompose the stream of sound in new ways to result in morpheme or meaningful phrase recognition. However, the second of the two, meaning building, involves processes that are similar in the L1. He also explained that inexperienced L2 listeners may not only be uncertain about their decoding accuracy but also may require much more focus of their mental resources on decoding, reducing the mental resources that can be directed at meaning building. For example, Field gives the sample sentence "I've lived in Italy for ten years" (2008, p. 87). If a learner hears this but does not realize that the speaker still lives in Italy, it may be the case that she has missed the /v/ signaling the present perfect verb phrase, which includes a reduced form of "have" that may be quite difficult to perceive in connected speech. Moreover, that phoneme is the key difference between the present perfect version of the sentence and the simple past form "I lived in Italy for ten years." This reiterates that, especially for beginners, listeners are using a great deal of mental resources in decoding the L2 speech stream, and as with Goh's (2000) higher-level listeners, ideally these processes are gradually made more automatic, leading to successful perception, parsing, and language utilization.

Another important consideration is that many L2 listening models focus mainly on language as a somewhat standalone process, while avoiding evidence from second language acquisition (SLA) suggesting that language in use is impacted by other attributes, such as social factors (Gardner & Lambert, 1959, 1972; Pavlenko, 2002). These qualities include not only word recognition itself but also one's identity and one's sociolinguistic control in language interactions. To reflect these features, Rost's later (2014) model of L2 listening ability is divided into an affective domain, a cognitive domain, and an interpersonal domain. The affective domain involves abilities such as a language learner demonstrating resilience (recovering from a loss of face in an interaction) or taking initiative. The cognitive elements are the spoken language processing that takes place in the mind of the listener. Here, Rost stressed that it is not necessary to listen like an L1 listener, but that a listener may have to retrain her auditory perceptual system to account for an L2 phonological system. The interpersonal characteristics of L2 listening include understanding appropriateness and types of engagement in a variety of contexts.

Rost's (2014) framework for L2 listening is in line with the observations of Field (2008), who noted that there has been a recent shift in the descriptions of L2 skills, bringing them more in line with definitions of the processes underlying real-world performance. This is evidenced in models such as Weir's (2005) cognitive validity framework, used in much of the recent language assessment literature. Field (2013) explained that, for a language test to possess cognitive validity, although it may not necessarily be possible for the conditions of the test administration to simulate an actual listening event, the test should elicit those mental processes that are representative of the processes in a target-language-use listening context. The model has been adapted by Cambridge English for their suite of listening exams, and it involves goal setting, decoding acoustic/visual input, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, creating a text-level representation, and monitoring comprehension (Weir, 2005, p. 45). A note should be made here, though, that Field (2008) warned that such a model does not necessarily favor native-speaker language *forms* but rather an expert listener's *processes* underlying listening performance. This underscores the concept that a language learner, as she becomes more able, should ideally draw on existing components of listening competence in her L1 in order to be able to cope with the circumstances of an L2.

(Field likened this to learning to drive a left-hand-drive car after having learned on a right-hand-drive one.)

To summarize processing models relating to L2 listening: Rost's (2011) division of components was linguistic, semantic, and pragmatic processing; Vandergrift and Goh (2012) described how Anderson's (1995; 2009) phases of perceptual processing, parsing, and utilization are connected to top-down and bottom-up processing; and Rost's (2014) model involved affective, cognitive, and interpersonal domains of L2 listening ability. In addition, Weir (2005, p. 45) had suggested listening elements of goal setting, decoding acoustic/visual input, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, creating a text-level representation, and monitoring comprehension.

Although Rost's 2011 and Vandergrift and Goh's 2012 (drawing on Anderson's 1995, 2009) listening comprehension models tend to focus on moving from perception to meaning making, Weir's 2005 and Rost's 2014 models suggest that we contribute not only our own core linguistic knowledge but also elements of our strategic competence and social identity when we listen. Strategic competence concerns all skills and components, linguistic and nonlinguistic, that L2 users utilize for successful understanding (Canale & Swain, 1980). A unified model that incorporates all of these characteristics may be necessary for understanding what is occurring in L2 listening, especially with processing considerations for beginner English language learners not acquiring their additional language in childhood. We can see from these frameworks that listening involves not only strictly linguistic processing but also other cognitive and metacognitive elements. Listening, thus, rather than being conceptualized as a passive skill, is immensely complex and requires active use on the part of a language learner. Even more complexity arises, moreover, when listening is done in high-stakes testing situations.

2.1.3 The L2 listening construct for assessment

Research in language assessment that has focused on the measurement of listening proficiency has often concentrated on academic listening, listening taking place in the educational TLU domain. The nature of the spoken language that takes place in schools, particularly in the upper levels of the U.S. K-12 contexts as well as at the university level, means that this must be taken into consideration when defining the listening construct. The TLU domain for academic listening would include settings such as formal lectures and instructor-fronted talk, seminar discussions, laboratory sections of courses, office-hour interactions or advising sessions, guest speakers, conference paper or poster presentations, and even campus tours. These academic listening scenarios occur in a variety of contexts ranging along the orality continuum and varying in their degree of interactivity with other interlocutors. Test developers must also consider that the construct may not include strictly listening but also other modalities. For example, besides listening, a poster presentation requires integrated skills in that the presenter must read, write, and speak, and its viewer often must read as well as listen. Because the TLU domain for university scholars contains one or more of these spoken language types, a listening assessment of English, especially one designed for academic purposes, should cover this domain to the extent necessary in order for there to be construct validity. Construct validity relates to the underlying knowledge or skills purported to be assessed in an exam.

In order for assessment to be valid and to operationalize a construct for academic listening, we must be able to understand what distinguishes EAP listening from general listening. Taylor and Geranpayeh, describing the academic listening construct, observed that "[a]cademic study makes ... heavy cognitive demands and is often characterised by context-reduced communication in which logic and inference play a key role" (2011, p. 93). Moreover, the idea of the academic lecture being a "non-collaborative monologue" (Lynch, 2011, p. 85) is generally not the situation in U.S. university class sessions; at least, this is perhaps more true of graduate-level than of undergraduate courses. Particularly at the graduate level, if classes tend to be more monologic than dialogic, spoken discourse is still participatory and interactive. The increase in the accessibility of technology has also changed the nature of listening (King, 1994), because a speaker can interact with slide show presentations or other audiovisual materials such as online videos, and listeners must process that input in tandem with the instructor's or other students' spoken language. Students also interact in one-on-one or small-group settings such as office hours, study groups, or tutoring sessions. Another crucial detail for L2 communication is that, whether in whole-class or smaller academic listening situations, there may be intercultural communication difficulties to contend with, not simply at the linguistic level but also at an academic-culture level. Lynch (2011, p. 83) gives the example of a speaker's raised eyebrow, with the speaker attempting to show that his speech was shifting from literal to humorous, as a facial gesture cue that may not be interpreted universally by listeners from all cultures, or even seen in a large audience.

Other characteristics of academic listening include speakers' vocabulary and discourse signaling cues (Vandergrift, 2007). There are academic and field-specific words and phrases present that may not necessarily be used in communication for general purposes; students may not have been exposed to such vocabulary prior to encountering it in their discipline. Discourse markers that a speaker uses such as "Let me now turn" or "I want to look at", as well as prosodic cues (using intonation to indicate relationships among different ideas presented), ideally provide signposts for a listener to be guided through instances of spoken academic language. Listeners may struggle if they lack background knowledge for understanding text content or if they are not familiar with texts' type or structure (Vandergrift, 2007). Moreover, instructors in a class "contextualize, elaborate, and critique the content of course readings through lecture and discussion" (Murphy, 1996, p. 107), so listeners also need to be familiar with those functions of academic language. According to Taylor and Geranpayeh, "Academic listening tests will ideally be both linguistically challenging and cognitively demanding" (2011, p. 96). Goh and Aryadoust (2016) note that "[m]uch of academic listening involves learning through listening to lectures" (p. 1), and for EAP language programs, it involves learning and attending to not only content but also language. This can be immensely challenging for beginner listeners, as successful L2 listening involves not only language mastery but also background or cultural knowledge, few internal or environmental distractions, and understanding of speakers' varied accent or speech rate (Miller, 2009), among other factors.

Any linguistic or cognitive characteristics of an exam, as well as traits of an individual engaging with the test, are details that can potentially impact test performance and are thus essential for researchers and test developers to keep in mind. When computing devices are introduced that may allow listeners to experience the input more than once, this, as well as the factors already introduced, may have an impact on L2 learners' understanding of spoken language.

2.2 Technology and listening

In L2 listening, there are a number of test method variables that may affect the measurement of examinees' language proficiency. Test method characteristics have been categorized by Bachman (1990), Bachman and Palmer (2010), and Douglas (2000) into factors including the physical and temporal features of the test, the input, the expected response, and the interactions between the input and the expected response. Expected response, as used here, refers to the linguistic and nonlinguistic actions a test taker likely must perform in an assessment (Bachman & Palmer, 2010). The input and the expected response format may both be affected by the use of computer technology. If the audio input is designed to be administered via computer, test developers must decide how many times listeners hear audio as well as whether play, pause, and play position controls are available for examinees to control. This control then becomes part of the way examinees respond to test items. In the case of a multiple-choice listening test, candidates must select their chosen answer from multiple options, and they also likely have to navigate among different screens or sections of the test. When technology or other elements of a test interface are introduced, care must be taken to ensure that computer familiarity does not unfairly benefit or impede successful listening.

With computer audio controls in a listening test situation, there is the chance for some construct-irrelevant variance to be added to the listening construct, depending on how that construct is defined. By construct-irrelevant variance, I mean nonlinguistic knowledge such as background knowledge, feelings, intentions, or past experiences (Banerjee & Papageorgiou, 2016; Elliott & Wilson, 2013). These nonlinguistic factors should ideally not inhibit examinees' display of their listening abilities. Computer technology involves examinees being able to control their strategic competence, as any resulting test score reflects not only language ability but also the strategic use of technology (Chapelle & Douglas, 2006). Thus, with a test involving computer controls, exam developers must be very cautious about how to justify the inferences derived from such a test. Relevant to the present investigation with regard to the audio input and expected response are the number of plays and ability to pause the stream of sound.

2.2.1 Number of plays

In a high-stakes listening test, learners have to devote a great deal of cognitive resources to process the audio input. With repetition, learners' cognitive load may be decreased, freeing up resources for attention or working memory, and their anxiety can be reduced (Vandergrift & Goh, 2012). In a self-study or classroom setting, learners can become more accustomed to the content, vocabulary, and structure of spoken input. However, there are effects of repeated listens to a text that can have an impact on listening test performance. For example, Berne (1995) and Jensen and Vinther (2003) found that repeated listens had a significant effect on final test performance, as did Chang and Read (2006), though Chang and Read's participants benefited more from being provided general background information about tested topics. Buck (2001) also reported that playing the text a second time generally makes tests easier. Brindley and Slatyer's (2002) twice-heard text was slightly easier than a once-heard baseline text; however, that baseline set was on a different topic. Much of the work in this area indicates that twice-played audio is generally easier, though the audio play has generally been controlled by a test administrator.

Roussel (2011) is an example of a study in which the number of plays was controlled by participants. There were three different proficiency levels represented in her study, and the learners' initial levels were operationalized by their performance on a listening-twice test. Roussel's participants, after an initial test, had been classified as B1 and B2 (intermediate) level on the Common European Framework of Reference for Languages (the CEFR levels range from A1 beginner to C2 advanced; Council of Europe, 2001). She found that, for French-speaking learners of German, self-regulated listening led to better scores than administration-imposed listening once or twice, but she uncovered no significant differences between listening once versus twice.

Ruhm et al. (2016) gave Austrian eighth-graders at the A2/B1 level a listening test. They hypothesized that not only frequency of input presentation but also item difficulty and audio input length may impact test performance. Using a logistic regression model, they investigated sources of variance in the data. Listening twice generally made items easier. Double listening helped more on short, decontextualized items that were more difficult items; on easier items, however, double listening made less of a difference in reducing the difficulty of items. Playing the recording twice had a larger effect on longer (>60 second) input than on shorter input. However, contextual information and the level of vocabulary interacted with item length and hence difficulty; short items had less contextual information than longer ones, and persistently difficult items had more advanced vocabulary. This illustrates that the effect of once- and twice-played audio may be more complex than initially thought, and parallel presentation of vocabulary in listening sets can sometimes be difficult to control for.

For standardized academic English proficiency tests, listening sections are generally played once or twice. Geranpayeh and Taylor (2008) summed up the stances from both positions aptly: in listening just once, the argument could be made that second hearings are sometimes impossible in the TLU domain of real-world listening. There is also the expectation of automaticity of processing, in which case listening once should be expected and sufficient. In listening twice, there is recognition of the state of listening in the non-testing context, in which listeners often have the chance to ask their interlocutor for repetition or clarification. There, however, the artificiality of the testing context is recognized. When there is repetition in the real world, the speaker rarely makes exactly the same utterance: prosodic elements may differ, even if the words are identical (Buck, 2001). Since the 1970s, with the availability of playback equipment that is affordable and reliable, once-played and twice-played audio have been delivered as recorded media rather than read aloud live. With that input type, the control has always lain with the test administrator.

2.2.2 Pausing and computer controls

In real-world listening to recordings, pausing and repetition is often available. Even native speakers or very proficient listeners take advantage of replay or pause options in audiobooks, streaming videos, or online podcasts. Occasionally there are play-pause or "go back 15 seconds" buttons available to the listener. East and King (2012) warned, however, that pausing and repetition may actually create a trend away from authenticity, as the speech stream becomes different from nonrecorded speech if it is sometimes paused or repeated. Vandergrift (2007) explained that, for listening input for L2 students, "authentic contexts, form and speech rate should not be sacrificed in the interest of simplifying L2 listening for the language learner" (p. 200). If we consider podcasts, audiobooks, or other replayable media, adding technological possibilities such as replaying or pausing is a way of "'authenticizing' practices that were once considered inauthentic" (Robin, 2007, p. 109). Because of timing and practicality considerations, however, with technologically-advanced testing capabilities, the inclusion of innovative techniques in the presentation of audio is something a test developer must consider carefully.

In Robin's (2007) discussion of computer-assisted language learning (CALL) trends and the consideration of repeated audio delivery, he contended that "listening has become a semi-recursive activity ... inching its way closer to reading, which is fully recursive" (p. 110). Additionally, Alderson, as early as 1990, supported the appropriate integration of computers in the assessment process. He highlighted the "element of learner choice" (Alderson, 1990, p. 24);

although he was referring to examinees being provided with computerized help dialogues or other types of feedback, this feature could also be applied to test delivery. The rapid spread of technology that allows language learners to hear repeated aural input thus creates opportunities not only for students themselves but also for instructors and assessment professionals.

Within CALL studies, researchers have investigated the impacts of innovative audio materials on listening comprehension. Studies such as those by McBride (2011) and Roussel (2011) have given learners control over components of the input. (I discuss control further in an upcoming section.) For much of the body of work in L2 listening in CALL, researchers' interests have lain mainly in language learning strategies (O'Bryan & Hegelheimer, 2007; Roussel, 2011; Smidt & Hegelheimer, 2004) and/or captioning, subtitles, annotations, topic or section organizers, and help dialogues (Grgurović & Hegelheimer, 2007; Hegelheimer & Tower, 2004; Hulstijn, 2003; Smidt & Hegelheimer, 2004). These areas reflect the importance of careful sequencing of test components and planned computer setup to assist language learners. The particular interest in strategy training has revealed that it is important to not only help learners become accustomed to the cognitive attributes necessary for focusing on listening comprehension, but also familiarizing them with, and considering the cognitive load of, engaging with a computer interface. Smidt and Hegelheimer (2004) highlighted making learner-controlled materials available but cautioned that learners should be taught how to operate them for their benefit, particularly regarding how to be able to identify meaningful information. Although language learning materials have become more innovative in many settings, with language instructors and SLA researchers designing more learner-controlled tasks, large-scale language proficiency testers have been more reticent to embrace a learner-controlled component of listening tasks, due to the variability it introduces.

In the listening literature, especially before the year 2000, the term *pausing* has not generally referred to a test taker being able to take control herself. Pausing has tended to refer to administrator-inserted pauses to allow test takers more time to process language, or necessary pauses built into a recording, as in the stop of audio play between items or sets so that examinees ideally have sufficient time to read options and respond to multiple-choice items. For example, Blau (1990) and Zhao (1997) included administrator-controlled pauses as part of their research designs. Zhao (1997) digitized each input sentence as a separate unit so that learners in the study could control speech rate per sentence, investigating how the rate of speech, not the pausing itself, impacted listening items. Blau (1990) found that pausing generally improved comprehension, but there was a proficiency threshold "to be able to take advantage of the extra processing time provided by the pauses" (p. 752). For higher-proficiency learners, pausing was not necessary. However, comprehension of input with pauses was, in general, higher than that of the original version.

With the availability of more accessible and user-friendly audio recording playback methods afforded by computing devices, listeners have sometimes been able to take control of the aural input. McBride (2011) gave examinees the ability to pause during the second listening of a set of dialogues, but they could not rewind or fast-forward, so students experienced some interrupted words when they paused and then played the recording from its stopped point. The test results suggested that people able to pause reported having a more positive learning experience than those who were not permitted play-pause control. Roussel (2011) gave French L1 listeners of L2 German control over audio play and found four different profiles of self-regulation, or "the capacity of the listener to exercise physical control over the listening input by using the mouse" (p. 100): learners who listened 1) once without pauses, then another time with pauses; 2) first with pauses then again without; 3) once or several times through without pauses; and 4) once with pausing throughout. The first strategy was generally the most effective for learners, with one first hearing globally then later hearings split into meaningful chunks to identify specific details. Learners in the second category were neither high proficiency nor low proficiency listeners. Proficiency effects were thus found in listening strategy use, as also uncovered by Blau (1990) and Chiang and Dunkel (1992). For the third type of profile, Roussel suggested that self-regulation may have put too much cognitive load on listeners, as it was difficult for them to know where to pause. For the fourth group, these listeners were the least successful, not able to successfully segment or perceive enough text, and their pausing was frequent and disorganized. The learner in Cross's (2014) podcast investigation adopted the strategy of full listening before listening in segments (Roussel's first listener profile), reporting that her eventual goal was to be able to understand as much as possible from an initial listen. Read and Barcena (2016) gave students listening input via a mobile phone application; higher-proficiency learners were better able to self-regulate. Students listened at least once and would typically pause before beginning the recording on their phones, then would often listen more than one time. Hence, it can be seen from these studies that learners are contending with a number of variables, particularly relating to play-pause control as well as their own listening proficiency level, in order to comprehend what they have heard.

I review these features about technology because, in a listening assessment setting, characteristics of the input, the expected response, or their interaction may have an impact on examinees' ability to accurately show their listening proficiency. The most important concept a test designer ought to remember, having considered these varied features, is what effect these characteristics have on engaging the listening knowledge, skills, and abilities desired to be assessed (Buck, 2001). I next turn to a feature that has been relatively little explored in the standardized language assessment literature: whether the examinee has control over the listening audio recording.

2.3 Locus of control

In this section, I focus on who has the ability to control the input of a given situation. I use the term "locus of control" in describing this characteristic of the audio input. This has connections with being able to help listeners *bias for the best* (Swain, 1983; Fox, 2004), or show their proficiency for a given purpose as best they can, in testing situations. If listening examinees feel more able or less anxious in a testing situation in which they have the locus of control, test developers can create test method conditions that allow for the best examinee performance.

Norton Peirce et al. (1993) used locus of control to discuss results in a language learning study they conducted with French immersion students. They explained that "the locus of control is said to reside with the participant (or participants) who exercise dominant control over the rate of flow of information in a communicative event" (pp. 36-37). The term can thus be relevant to not only co-constructed spoken language such as a face-to-face conversation, but also to recorded spoken input such as a radio program. Norton Peirce et al. (1993) explain that the locus of control is with the language learner in literate activities, but not as much in oral activities; moreover, "the locus of control is more favorable to the learner in oral production than in oral reception" (p. 37). In other words, in listening, which is a receptive oral skill, the learner by nature has less control than in any of the other three modalities of speaking, reading, and writing.

An L2 listening study that did investigate control was that of Zhao (1997), in which each of the study participants comprehended input better when they had control over the speech rate; moreover, they self-reported that slower speeds helped them listen. Additionally, Roussel (2011) noted that listening self-regulation may be beneficial to learners at certain proficiency levels but not others; her results indicated that, for lower-proficiency listeners, self-pacing may have overloaded their working memory. She argued that if they had not had to control the listening input, they might have been better able to focus on comprehension processes. In the language learning studies mentioned in this subsection, however, listeners interacted in self-study sessions or language classrooms and not in a standardized language testing scenario.

The concept of locus of control has applications to language assessment because, if candidates feel they have more of a decision in listening to the input, their anxiety may be reduced. Tsui (1996) conceptualizes anxiety as possibly being related to fear of negative evaluation or fear of failure. Horwitz (1986) was one of the first scholars to operationalize a foreign language learning anxiety scale for classroom research, which allowed researchers to standardize how they were conceptualizing anxiety. Later researchers such as Aida (1994) separated anxiety into its different components, identifying test anxiety in a factor analysis as loading on a different factor than general foreign language anxiety. Elkhafaifi (2005) also distinguished anxiety into its subfactors, separating learning anxiety from listening anxiety, in the context of U.S. university students learning Arabic. Extending to listening assessment, In'nami (2006) and Chang and Read (2008) investigated connections between anxiety and listening proficiency. Some of Chang and Read's participants reported that only being able to listen once may have been a source of their anxiety. In'nami, using factor analyses and structural equation modeling, found that test anxiety factors (such as emotion, general test worry, and test-irrelevant thinking) did not impact listening test performance; his results supported Aida's (1994) finding that test anxiety was conceptualized as a different type of anxiety than general foreign language learning fear or discomfort. Thus, anxiety, as it relates to applied linguistics and language acquisition, is clearly a multifaceted phenomenon. However, few, if any, language learning studies have explored the reduction of anxiety specifically in a case of the ability to have control in a situation, much less in a language assessment context.

Locus of control may relate to principles of learner autonomy, or "the philosophy that students should have a large amount to say about what, how, and how fast they learn" (Bailey, 1999, p. 41). Broadly, for successful learning, it has been argued that "adults demonstrably learn more, and more effectively, when they are consulted about dimensions such as the pace, sequence, mode of instruction and even the content of what they are studying" (Candy, 1988, p. 75). Choice in the locus of control may also have a connection with language learner and examinee motivation. Being able to make one's own selections while learning may enrich intrinsic motivation (Patall, Cooper, & Robinson, 2008). Sage, Bonacorsi, Izzo, and Quirk (2015), in an investigation of self-paced versus computer-controlled advancing of slide-show panels, reported that participants did not prefer the computer-set-paced format; the format that users found the easiest to learn from was one during which they could freely pause. The respondents noted that "they could stop during moments of confusion ... were in control ... could select their own focus ... and could guide their own learning" (Sage et al., 2015, p. 185).

Turning back to language learning, Vandergrift (2005) remarked that there appear to be relationships among motivation and L2 listening proficiency. He, however, used the term "locus of control" differently from the way in which Norton Peirce et al. (1993) employed it. Control here, as Pintrich (1999) and Vandergrift (2005) utilized it, is conceived as the ability to self-regulate one's metacognition (thinking about one's own cognition). Although this concept may be connected to who has control over task features, it is more concerned with learners' or examinees' intrinsic motivation.

It should be clarified here, though, that although one portion of the control may be given to test takers, the rest of a standardized listening test is still largely under test administrator control. Lawless and Brown (1997) refer to this as external locus of control, while components that learners themselves manipulate would be under an internal locus of control. The audio chosen, the items included, the overall results and score reporting, and the purposes for which the results are used are all externally controlled by the institution that developed the test. If every part of a listening exam were highly variable, the test would not truly be standardized for all examinees, and it would be very difficult for assessment professionals to be able to interpret test results. With learning environments that may be too open ended, learners with little prior knowledge of a task may experience difficulties (McNamara & Shapiro, 2005; Roscoe, Allen, Weston, Crossley, & McNamara, 2014; Scheiter & Gerjets, 2007). Nevertheless, if examinees are able to exercise some control over the incoming stream of speech, they may find the high-stakes listening testing scenario overall less alarming. Being provided a choice may have a positive effect on perceived competence and observed performance (Patall, Cooper, & Robinson, 2008).

The next two sections focus on studies which have used Rasch measurement (a quantitative method that can be applied to language test item analysis) and on investigations that have used stimulated recalls (interviews with stimuli that help a researcher collect examinees' thoughts about an event), both of which have been receiving more attention recently in the testing of L2 listening.

2.4 Rasch analysis of listening comprehension items

In language testing, standardized test administrators must be able to, in some way, operationalize comparisons among items being easier or more difficult and what the items tell us about test takers being more or less able. A score on a test gives us data about how a given examinee performed on a given set of items. This means that if this examinee takes two tests of different levels of difficulty, she may appear to have high ability based on the first test, but average ability based on the second test (Bachman, 2004). Thus, psychometricians record individual item statistics with respect to relevant samples of examinees so that we can better understand what inferences we are making about language learners' ability level. Item response theory (IRT) has been used in various standardized tests for such a purpose. IRT explicitly acknowledges that test performance is based on not only test takers' ability on an underlying trait (such as L2 listening proficiency) but also on item traits that reflect examinees' ability. For classical item analysis, statistics are calculated, such as the difficulty (sometimes called facility) of an item and its discrimination index or item-total score correlation. Test developers examine whether an item is satisfactorily discriminating, or whether it sufficiently groups examinees into different ability levels. Information about test questions is useful in order to justify whether to keep items on a test, eliminate them, or revise and retry them. Supplementing with difficulty and discrimination statistics is ideally done after assessment professionals have ensured that the items fit test specifications based on their content (Bachman, 2004).

Item responses "are a function of both person achievement and item difficulty" (Engelhard, 2009, p. 591); however, person ability and item difficulty are not the only factors that may contribute to item responses and thus test scores. Test performance can be conceptualized as a function of various facets, or other variable aspects of a test beyond person ability and item facility that may have an effect on scores. For example, a variable such as number of plays in a listening test could be considered in a measurement model for a listening test. Rasch (1980) measurement has been used to examine the effects of various examinee or test characteristics (e.g., age, gender, topic of test section, time or day test was taken, prior examinee language learning or testing experiences) on test scores. Many-facet Rasch modeling is a type of IRT for examining not only item-level and examinee performance but also other factors that may have an impact on scores, which can all be analyzed on a common, interval-based scale. In Rasch measurement, two assumptions need to be made: the data must fit the model, and the test must measure a single, unidimensional construct (Bond & Fox, 2007; Eckes, 2008; McNamara, 1996).

In language assessment, Rasch modeling has been used primarily in performance assessment to examine the effects of various facets, such as rater traits, rating scales, or examinee characteristics, on scores that raters assign (Bachman, 2004; Eckes, 2009; Lim, 2011; Weigle, 1998). These facets are all variables that may have an impact on scores and thus merit investigation, particularly in high-stakes settings. Language testers have analyzed such data using FACETS multi-faceted Rasch modeling software (or WINSTEPS software for dichotomous items), developed by Linacre (2014). This tool provides ability estimates for variables (test questions, persons, or other facets) as well as fit statistics. Ability estimates from the software output offer information about how much variation there is along a facet, such as differences in examinee ability or how far raters are from an average. Fit statistics show how well the facets of interest fit the statistical model produced. For instance, Weigle (1998) and Lim (2011) are two studies which used FACETS fit statistics to examine rater behavior on writing assessments; satisfactory model fit indicated rater consistency, which is desirable for a standardized test in which raters are expected to behave similarly.

The Rasch measurement model has also been used with regard to dichotomously-scored items, such as MC standardized test questions. It has been employed in listening assessment investigations such as those conducted by Aryadoust and Goh (2014), Batty (2015), Goh and Aryadoust (2010), and Papageorgiou, Stevens, and Goodwin (2012). Goh and Aryadoust (2010) used Rasch analysis to examine fit of items to a statistical model, indicating that the test items were measuring a unidimensional construct (Wright & Linacre, 1994). Also of interest in Rasch analysis are person reliability and item reliability indices, which Aryadoust and Goh (2014) used in their investigation of Michigan English Test (MET) listening comprehension items. The person reliability index shows how precise the measurement of test taker ability is, or "the sensitivity of the test to distinguish among high- and low-ability test takers" (Aryadoust & Goh, 2014, p. 12). The item reliability index serves as a measure of how sufficient the sample size of test takers is (Aryadoust & Goh, 2014). Item reliability is also used to interpret how precisely the items are measuring an underlying latent variable (Beglar, 2010; Wright & Linacre, 1994): listening comprehension, for example. A value closest to 1 for reliability indices is ideal.

In Rasch analysis, a researcher can also obtain information about how item performance may vary based on different testing conditions. For instance, Papageorgiou et al. (2012), in examining the effect of monologic versus dialogic input on listening comprehension, administered the same topics to examinees; some listened to audio with one speaker and others heard audio with two speakers, though the same information was presented in both sets. They then gave identical listening comprehension items after dialogic or monologic audio and compared the items using Rasch measurement. Of three different topics, they found that the dialogue was significantly easier than the monologue for only one topic of the three. Their Rasch analysis was supplemented by a content analysis of the listening item set language and what items were assessing; the topic that led to easier items in the dialogic condition may not have had equal information density, redundancy, or discourse markers as in its monologic condition. Batty

(2015) used Rasch measurement to explore how video or audio input impacted listening comprehension; he was able to compare item-level analysis from the statistical model in tandem with a content analysis. He concluded that there were no meaningful differences between audio or video delivery of the listening test.

Although quantitative approaches such as many-facet Rasch modeling can help us compare aspects of the test input and thus make inferences about how examinees are able to show their listening proficiency, qualitative measures are also valuable for tapping learners' perceptions. One qualitative approach in particular, stimulated recall, is discussed in the next section.

2.5 Stimulated recall in listening comprehension

Stimulated recalls (SRs) are one type of introspective method, specifically a retrospective means, to gather data about participants' thought processes in performing a task (Ericsson & Simon, 1996; Gass & Mackey, 2000). As *retrospective* indicates, this is done by asking participants about what was going through their mind *after* they undertake a task, prompted by some sort of stimulus. With introspection, there is an assumption that it is important to be able to reflect on mental processes because, in L2 research, it is ideal to have data sources other than language production data (Gass & Mackey, 2000).

In SR, an aural and/or visual stimulus of a participant's task performance is presented to her and utilized to stimulate recall of the mental processes that were in use during the earlier completed event. It is valuable for L2 researchers because it may help identify, from a participant's stream of consciousness, the type(s) of knowledge drawn on when processing linguistic information. It can also help us reveal more about the mental structures or representations that are used to be able to organize such information. With production data, L2 researchers sometimes make inferences about the reasoning behind learners' spoken or written behaviors, but understanding the source of this behavior cannot be done only by consulting learners' production. There may exist more than one explanation for why learners produced something that can only be examined by attempting to investigate learners' process data (Gass & Mackey, 2000).

It should also be mentioned how SRs differ from other introspective methods. SRs are retrospective, in that the information meant to be elicited is not done at that moment but rather immediately afterward. This differs from, for example, a think-aloud protocol, in which learners are asked to verbalize *as* they complete a task (Vandergrift, 2010). For SRs, to target the cognitive processes of interest, a strong stimulus is needed to elicit information that will help the participant think about what had been going on in her mind at the time of the earlier event. Gass and Mackey (2000) explain that ensuring a strong stimulus may involve use of more than one data source, such as watching a video and reading a transcript, particularly if the recall is more delayed. Methods of immediate retrospection such as SR differ from simultaneous introspection or delayed retrospection with regard to the temporal distance between the action and the verbalization. Immediate retrospection may be particularly appropriate for studies in which participants are attending to strategies for task completion, as Faerch and Kasper (1987) noted. Attempting to elicit a participant's thoughts during a task may not be a possibility because disturbing the person too much may cause interference in accomplishing the language activity (Sasaki, 2014).

The use of introspective methodology in applied linguistics has been influenced by related fields such as psychology and theoretical linguistics. Prior to the 1950s, due to the influence of behaviorism, introspective methods were perceived as unreliable. Behaviorism's assumptions involved the "objective" external observation and analysis of human behavior. From the middle of the twentieth century onward, researchers began to become more interested in humans' cognitive processes, and research paradigms shifted somewhat, with introspection being viewed as an important resource. With behaviorism falling out of favor, valid and reliable data elicitation and analysis methods for introspection gradually became more acceptable in many circles. Since the 1950s, and especially rising in use in the late 1980s and through the 1990s to the present day, researchers have investigated process data through introspection (Gass & Mackey, 2000; Sasaki, 2014).

Much of the past introspection research has addressed a broad range of research questions. The SR methodology in particular has been utilized to learn about oral interaction (either comprehension or production), classroom-based research, reading, writing, or vocabulary learning (Gass & Mackey, 2000), and naturally some of these areas overlap. Within applied linguistics, SLA researchers have frequently employed verbal reports (Sasaki, 2014); Rebuschat (2013) observed that much of the verbal report research within SLA has given attention to form-meaning connections, investigating how participants verbalize rules or patterns relating to certain linguistic structures of interest. Verbal reports are also becoming more widely used in language assessment (Sasaki, 2014). Buck (1990; 1991), Suvorov (2013), Wagner (2008), and Wu (1998) are examples of studies in assessing listening that have employed retrospection, investigating the processes leading to successful comprehension. These as well as other qualitative methods are needed for researchers to gain a fuller understanding of the adult L2 listening assessment construct.

2.6 Summary of prior research

The review of existing research focused on the nature of oral (versus written) language, the effect of technology, the factor of locus of control, Rasch measurement in assessment investigations, and stimulated recall research. Quite recently, listening has been receiving more attention within applied linguistics, second language studies, and language assessment. Listening assessment researchers have employed psychometric methods such as Rasch modeling in order to investigate how test results may be impacted by the effects of examinees' first language (Harding, 2012), monologic versus dialogic input (Papageorgiou et al., 2012), examinees' age (Banerjee & Papageorgiou, 2016), or speaker accent in the input audio (Ockey, Papageorgiou, & French, 2016). Introspective and retrospective methods have been utilized in investigations of listening examinees to examine their reactions to audio and video stimuli (Buck, 1990; 1991; Révész & Brunfaut, 2013; Wagner, 2008; Wu, 1998). Recall that many factors of the input or expected response, as well as test takers' experiences that they bring to the test, have the potential to impact listening comprehension items and thus examinee scores, as well as possibly the uses of those scores.

Within listening studies in standardized assessment contexts, item performance has been investigated with regard to number of plays and/or speech rate, but not, to the best of my knowledge, regarding who has control over the audio play. CALL studies have focused their attention on language learners' strategies or overall results rather than on item-level analysis. Although it is important to consider test takers' results, I argue that the results are a function of not only examinees' level but also item characteristics. Because some people may choose to play the audio once while others may self-pace, this variable feature thus has the potential for introducing variance into observations of examinee performance on individual items. However, it is unknown what effect self-pacing may have on items and hence examinee listening performance in a mock standardized test. Seeking to fill this gap, I now turn to the present study.

3 METHODOLOGY

In order to discover more about whether a self-paced test is a valid and reliable measure of L2 listening proficiency, I employed both quantitative and qualitative methods of data collection and analysis. The quantitative data include (a) test item responses, (b) examinees' post-test survey Likert-scale item responses, and (c) computer-collected interactivity data such as number of clicks and time spent on each page of the test interface. The qualitative data include (a) examinees' post-test survey open-ended item responses and (b) retrospective stimulated-recall (b1) video capture of the listening test and (b2) follow-up interview comments. By *valid*, I refer to construct validity (is the test measuring L2 listening proficiency? do items discriminate between low- and high-performing examinees?), as well as face validity (do examinees perceive the test to be valid?). By *reliable*, I refer to internal consistency (as measured by Kuder-Richardson-21 internal consistency figures). I formulated the following research questions and hypotheses for the study.

3.1 Research questions and hypotheses

Research Question (RQ) 1: Is a self-paced (i.e., examinee control of playing, pausing, and audio position) play condition as valid and reliable as one administrator-controlled play in a listening exam?

RQ 2a: Given the opportunity, do examinees take control of (i.e., pause and/or repeat) the listening audio play, and does this vary by examinee listening proficiency level?

RQ 2b: If listeners do take control of the audio play, when and why do they do so, and does this vary by examinee listening proficiency level?

I wanted to explore whether examinees take control of the listening audio play, and, whether they do or do not, how to use test items to make inferences about listening proficiency. In a mock standardized testing setting, I was curious whether this setup would help adult English learners best display their listening abilities. My hypothesis for RQ 1 was that, to bias for best, items in a self-paced listening assessment condition would permit listeners to show their ability better than on items presented just once, and examinees would prefer self-paced items and feel that they were a good measure of their listening proficiency. Regarding RQ 2a, for beginner English listeners, even after repeated listens in a self-controlled condition, I imagined that their listening comprehension item performance would still be weak. For advanced listeners, I did not think opportunity to use play control would result in significant differences; in fact, I did not expect many of them to engage in pausing or repetition of the audio input. For intermediate-proficiency listeners, I expected that they would take play control, but that they might have more success with items following self-paced rather than once-played audio. For RQ 2b, when listeners take control of the audio play, I thought they might do so when they wanted to confirm their hunches about a selected choice and/or when they wanted to hear a specific detail again. I expected more of this behavior in beginner and intermediate than in advanced listeners.

3.2 Data collection overview

I collected data in two parts: a main test and a supplementary test. Table 1 shows a breakdown of the approximate chronological order of events with who the examinees and interviewees were, the items they responded to, and the length of the components.

Table 1 Testing events

Session and Participants	*Event*	*Approx. Duration*
98 students took main test (97 Intensive English Program students in group sessions, one grad student individual session)	took main (43-item) test and post-test survey	1 hour 15 minutes
2 other graduate students took main test and participated in SR in individual sessions	informed consent, participant information collection, and setup	5 minutes
	took main (43-item) test and post-test survey	45 minutes
	participated in SR interview	45 minutes
8 (of the 97) IEP students took supplementary items and participated in SR in individual sessions	informed consent, participant information collection, and setup	5 minutes
	took supplementary (22-item) test and second post-test survey	30 minutes
	participated in SR interview	45 minutes

The main test (43 items) included 100 prospective and current students at a U.S. university; two of the students participated in stimulated recall (SR) about these items. The supplementary test (22 items) was taken by eight of the students who also took the main test; these students all participated in follow-up interviews. My original goal had been to administer a minimum of 100 tests with both matriculated and non-matriculated students for the main test, and I also had wanted to recruit minimally ten SR interview participants. Because I was not able to capture the actions occurring on participants' screens for the main test administered in classes, the supplementary test was used for the eight follow-up participants so that I could gather their reactions to items new to them in the interview sessions. I also decided to invite two graduate students who had not yet seen the main test to not only take that listening test but also participate in SR. This means that some of the SR data are reactions to items from the main test, while other SR data concern supplementary test items.

3.3 Participants

Of the 100 examinees for the main test, 97 were Intensive English Program (IEP; pre-university) students who took the test during regular class time in computer classrooms. The IEP consists of five levels; participants came from composition classes at levels 3, 4, and 5, and an oral fluency class at levels 1 and 2 (levels 1 and 2 had been combined into one class due to low enrollment). The other three participants were ESL Credit Program (matriculated) graduate students who took the test in separate sessions each on their own. Two of the three graduate students each took the test with video screen capture, followed by a stimulated-recall interview.

Because all IEP students take a composition class that meets in a computer laboratory and had been attending classes for at least one month, they knew how to log into and out of campus computers and navigate a web browser to access their course content. The graduate students were also highly proficient with computers. I thus considered everyone's computer skills to be sufficient for computer-based language learning activities. There were 101 tests completed, but one student took the test in his oral fluency as well as his composition class, so I invalidated his second result, which left a total of 100 unique examinees.

The 100 examinees' average age was 25 (SD = 6.64, median age 23), and they reported they had begun learning English around age 17 on average (SD = 7.53, median age 17). They came from eight language backgrounds: Arabic, Chinese, French, Korean, Portuguese, Spanish, Ukrainian, and Vietnamese. There were 43 students who reported they had been in English-speaking countries less than six months, 33 students six months to a year, and 24 students longer than one year.

3.4 Instruments and procedure

3.4.1 Listening test

There were 43 items used in the main test and 22 items in the supplementary test; the 22 supplementary items were not part of the main statistical analysis because not enough examinees took them. The data analyzed for this research study were provided by Cambridge Michigan Language Assessments (CaMLA) in Ann Arbor, Michigan (CaMLA, 2014). The items come from the Michigan English Test (MET). The MET is a test of receptive skills (listening and reading) given in English as a foreign language (EFL) contexts. It aims at A2 (upper beginner) to C1 (lower advanced) levels of language proficiency on the Common European Framework (Council of Europe, 2001). The MET listening and reading sections are used for academic and professional purposes to make inferences about examinees' receptive skill use in English. The listening section in an operational exam consists of 60 multiple-choice items with four options each: one correct answer and three distractors. Listeners hear short audio recordings ranging from approximately 10 seconds to 1 minute 45 seconds, and in a typical exam setting, these are delivered via recorded media played once by the test administrator to all examinees simultaneously. Recordings are of trained voice actors who read the items as naturally as possible, with contracted and reduced word forms and hesitation markers, in a recording studio without background noise. There are three types of conversations or monologues: (a) short conversations between a female speaker and a male speaker, followed by one comprehension item; (b) longer conversations between two speakers with three to four comprehension items; and (c) monologues with four comprehension items. After each input plays, test takers hear a narrator read the text of the comprehension question(s) and can also read the questions printed in the test booklet, but they must read the options themselves. A sample dialogue and monologue can be seen in Appendix A.

I received permission from CaMLA to use retired listening test audio and questions. Some are publicly available on the web (CaMLA, 2014), but it is unlikely that examinees had seen them before this study. The sets had undergone experienced item writer review and include a variety of academic, professional, and personal/daily life situations. Various subskills (listening subconstructs) were assessed; item types included questions tapping the main idea of each set, important details, inference and implicature, understanding lexicogrammatical elements (vocabulary and/or phraseology) in context, and/or speaker's purpose. In an operational form, item question stems (but not the options) are delivered aurally, so examinees must read the four options themselves for each item but can listen to and read the item question.

One special listening item type that attempts to tap the understanding of vocabulary in context is the replay-context item. The listener hears the conversation or monologue, and later, when the items are presented, the candidate hears, "Listen to a part of the talk/conversation again, then answer the question", "what does the speaker mean when (s)he says: ...", or "Why does the speaker say: ...", then part of the audio input plays again. Candidates then must read the replay-context item options and indicate what the speaker meant when a certain utterance was said. The text of the replayed portion is not printed in the test booklet, though the question is.

Normally the test is paper-and-pencil, but so that I could offer an option for examinees to control the main audio for certain sets, I administered the items using Qualtrics web-based survey and test software (Qualtrics Research Suite, 2016). All students used classroom computers running at least 3-gigahertz processors with 8 gigabytes of memory. Students each had their own pair of earphones. I asked everyone to use the Google Chrome browser; this ensured that all students had the same Shockwave Flash browser plugin so that listening audio loaded and appeared identically. Question text was not delivered aurally except for replay-context items; examinees thus had to read all options for themselves as well as the text for most questions. Participants tested that their headsets and computer audio play functioned properly before beginning the scored items. Mobile devices and other resources were not permitted during testing. Qualtrics provides extensive output for item responses; in addition to item results, I also collected information about how often examinees clicked per screen, as well as time spent on each page.

Examinees took MET items in administrator-controlled and self-paced audio play conditions. The administrator-controlled sets were played once with the audio play control bar not accessible to test takers. For the self-paced sets, candidates could play, pause, and/or move the audio position. Examinees were warned on the instructions page that the audio was set to autoplay as soon as they navigated to a new page of items; audio began as soon as they clicked the ">>" (Next) button. They were also told that replay-context items, which had their own player, would not autoplay, so they needed to click those to hear them. Four of the nine sets each had one of this item type. When there was a replay-context item, examinees were able to play it as many times as they wanted, within the time limit of the set. The Next button was hidden for the time length of one main audio play so that candidates could not proceed forward until they had ideally listened once. The Next button was set to appear at the bottom of the screen after the length of the recording plus one second; for example, for a 59-second dialogue, the Next button was clickable after 60 seconds. Participants could preview the questions and options while the audio played, and notetaking was permitted but not required. Making a selection for every item was not mandatory; examinees could leave questions unanswered before proceeding to a new page. Unanswered items were categorized as incorrect responses. Examinees could not return to previous pages of the interface.

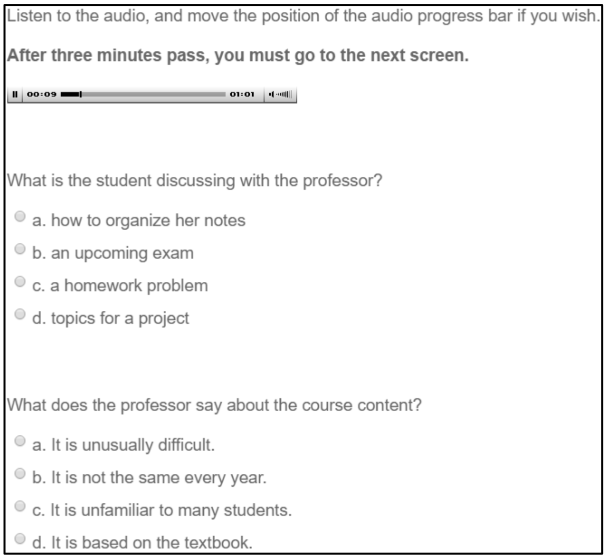
For the main (43-item) test, all participants took the same 10 items administrator-controlled with no start or stop option visible, played once, hereafter called anchor items, to have a comparison set of items in the analysis with less examinee variability. These were followed by 33 items with audio in three conditions of 11 items each: administrator-controlled audio with no start/stop permitted and played once ("1x" hereafter), self-paced with the ability to pause or replay with a three-minute time limit (self-paced short, or "SPS"), and self-paced with no time limit (self-paced long, or "SPL"). In the 1x condition, the audio control bar was not visible, but for other conditions it was available at the top of each page. For SPS sets, a three-minute timer counted down at the top of the page; the timer could only be hidden if examinees scrolled down the page until it was out of view.

The 33 items were part of nine sets, and each set was presented on its own page in Qualtrics with either three or four items. In each condition, there were two conversations with four and three items, followed by one monologue with four items. This meant that listeners heard 9 total sets and saw 3 sets (11 items) per condition, and each examinee was assigned to one of three versions ("Forms"), A, B, and C, of the test so that not everyone heard the same audio in the same play condition. For example, if test taker 1 heard the student-professor conversation set in the 1x condition, test taker 2 heard that same set in the SPS condition. Item codes for the 33 items in sets are labeled Q23 to Q51 and Q57 to Q60, sequential numbers except for 52 to 56, and anchor items are labeled A01 to A10. Table 2 shows the arrangement of listening items; read down one column to view the item order for that form.

Table 2 Listening test layout

Condition	*Form A*	*Form B*	*Form C*
1x anchor items listening once, no play control	A01-10	A01-10	A01-10
1x sets listening once, no play control	Q23-26, 31-33, 44-47	Q27-30, 34-36, 48-51	Q37-40, 41-43, 57-60
SPS sets listening self-paced with 3-minute time limit	Q27-30, 34-36, 48-51	Q37-40, 41-43, 57-60	Q23-26, 31-33, 44-47
SPL sets listening self-paced with no time limit	Q37-40, 41-43, 57-60	Q23-26, 31-33, 44-47	Q27-30, 34-36, 48-51

The distribution of Forms A, B, and C was kept as equal as possible: I used a Qualtrics master link, via the Survey Flow feature, to allocate similar numbers of respondents to each test version. This resulted in 32 examinees taking Form A, 35 Form B, and 33 Form C. The audio recordings I used ranged from 54 seconds to 1 minute and 42 seconds in duration. Candidates took 44 minutes on average, but they had 1 hour and 15 minutes of class time to complete the whole test. The figure shows a screenshot of the Qualtrics computer interface.



Listen to the audio, and move the position of the audio progress bar if you wish.

After three minutes pass, you must go to the next screen.

What is the student discussing with the professor?

○ a. how to organize her notes
○ b. an upcoming exam
○ c. a homework problem
○ d. topics for a project

What does the professor say about the course content?

○ a. It is unusually difficult.
○ b. It is not the same every year.
○ c. It is unfamiliar to many students.
○ d. It is based on the textbook.

Figure 1 Screenshot of Qualtrics computer interface

The audio control bar can be seen in the screenshot. This slider bar was visible in the SPS and SPL sets, and on items (Q30, Q33, Q40, and Q51) with a replay-context snippet. The participant left-clicks the box at the far left to play or pause, and she can left-click-drag the black tick mark that moves from left to right along the slider to control the audio position.

3.4.2 Post-test survey

Immediately following the test items, examinees completed a post-test survey shown on the last page of the Qualtrics interface. I asked respondents to indicate from 1 (strongly disagree) to 6 (strongly agree) on Likert-scale items as well as respond to open-ended questions. I inquired of candidates whether they preferred being able to have control over the audio play, as well as questions about their perception of the difficulty of and their familiarity with the topics. The survey items (response types follow in parentheses) were:

"Did you prefer being able to have control over the audio?" (Yes/No)

"Why did you choose Yes or No above?" (text box)

"The topics presented in the listening sets were easy." (disagree 1 2 3 4 5 6 agree)

"I was familiar with the topics covered." (1 2 3 4 5 6)

"I did better on the test items when I was able to control the audio." (1 2 3 4 5 6)

"This is a good test of my listening ability." (1 2 3 4 5 6)

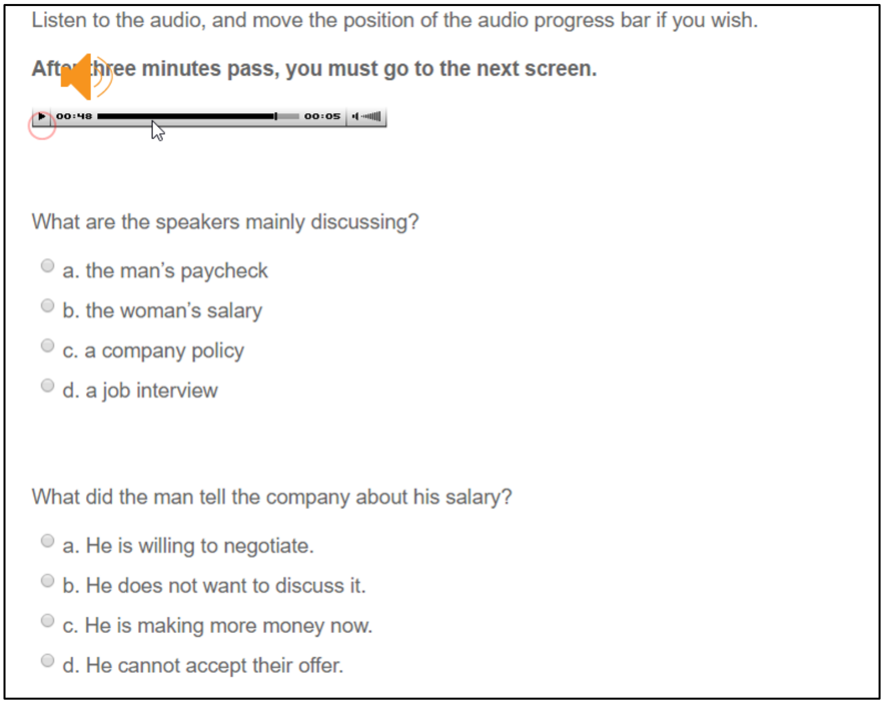
"Place any comments or questions here." (optional text box)

These post-test survey items were incorporated because I feel it is important to gather information about examinees' perceptions of what makes a valid test, particularly because face validity can be difficult to operationalize. By face validity, I refer to the appearance of validity: stakeholders' perceptions that the test is a good test of one's abilities. To gain information about listeners' perspectives on the test, as Révész and Brunfaut (2013) did, I included this perception questionnaire data.

3.4.3 Stimulated recall (SR) interviews

Recruitment for a follow-up stimulated recall and interview session was open to all examinees from the main test as well as to ESL Credit Program students. For the main test, I provided a space at the end of the exam for IEP students to offer their contact information. Also, two students heard about the testing opportunity through their ESL Credit Program colleagues. Each participant scheduled a one-on-one interview, all conducted by me. Participants were paid $10.00 cash for their time. All participants signed informed consent documentation as required by the university's Institutional Review Board. When I started an interview, I did not explicitly tell listeners that I was specifically examining stimulus play behaviors. Rather, I explained that the purpose of the project was to investigate how people listen in English. The SR and interview process had been piloted with ten other listeners one year prior to the collection of this data in order to check that all software and the interview procedures functioned.

The participants took listening test items in one-on-one sessions with a computer running ActivePresenter, freely available screen-capture software. While people used the Qualtrics test interface to listen and respond to comprehension items, ActivePresenter was simultaneously running to gather a recording of mouse trajectories and clicks. Whenever they clicked, a mouse-click sound effect played. Roussel (2011) tracked listening test pausing and mouse movements in a similar way using the software Camstudio. The image here, captured using ActivePresenter, is a still from the video of one participant. The listener had just clicked the pause button, turning it to the play icon shown, and was proceeding to click on the audio progress bar about one-fourth of the way into the recording:



Listen to the audio, and move the position of the audio progress bar if you wish.

After three minutes pass, you must go to the next screen.

What are the speakers mainly discussing?

○ a. the man's paycheck
○ b. the woman's salary
○ c. a company policy
○ d. a job interview

What did the man tell the company about his salary?

○ a. He is willing to negotiate.
○ b. He does not want to discuss it.
○ c. He is making more money now.
○ d. He cannot accept their offer.

Figure 2 ActivePresenter screenshot

SR participants then watched the video of their test behavior, which was played back using ActivePresenter's video preview feature. When they had clicked, a circle appeared on the screen, as can be seen in the sample screenshot; it was a small growing circle that originated from the clicked point. I audio-recorded participants' comments and asked them to pause the video whenever they wanted to comment on what they had been thinking about during the test. If the examinee did not pause the video herself, I paused the video at times I wanted to elicit more information and encouraged her to tell me what had been going through her mind. Interviews ranged from 16 to 58 minutes. The interview protocol is provided in Appendix B. After each interview, I exported the ActivePresenter video file of the screen capture with participants' behavior, and I saved the audio file from the interview session, both stored securely to consult later. Interviews were transcribed verbatim with repetitions and false starts included and hesitations and pauses indicated.

3.5 Data analysis

I was first interested in examinees' responses to the listening items themselves, so I converted the multiple-choice item results to values indicating correct (1) or incorrect (0) responses. I then calculated students' percentage correct values on the test in order to be able to provide them feedback. (I warned students that, because this test is designed to be difficult, it was not expected that any of them would score 100%; an average on a test like this was closer to 66%. I sent their scores via email with this explanation, and I also sent scores to their IEP instructors for diagnostic feedback.) I then used Microsoft Excel to calculate facility (difficulty) and point biserial indices for items across conditions to determine which were the most difficult or easiest, and which best discriminated among different learner proficiency levels. Also using Excel, I compared reliability among different conditions.

Since not everyone took every condition with each topic and item, but I wished to compare the effect of the facet of input control on identical items, I employed many-facet Rasch measurement (MFRM). This type of item response theory, a probabilistic model, permitted me to be able to plot items and examinees along a like scale for comparison. The use of an interval scale means that an equal distance between any two data points represents an equal difference in person ability or item difficulty (Bond & Fox, 2007), offering an advantage over traditional test statistics. For example, raw scores of 40% and 80% cannot be compared on an interval scale for examinees or items; the former score does not indicate that items were "twice as difficult" than for someone with the latter score. Thus, item difficulty, person ability, and input condition difficulty could all be compared on one scale.

MFRM uses a logit, or log-odds, interval scale, to examine both person ability and item model fit. A logit value for person ability is a calculation of the natural logarithmic value of the odds of item success. In other words, with Rasch modeling, I wanted to answer, "For an examinee at a particular ability level on a particular item, what is the probability of success?" or "For an examinee at a particular ability level on a particular item *in a particular condition*, what is the probability of success?"

The data were set up as a specification file to be run in the software FACETS (Linacre, 2014). FACETS outputs the interval scale in the form of a variable map, or Wright map, allowing for the direct comparison of test facets of interest (Eckes, 2009). This meant I could compare the effect of play condition (listening-once, self-paced shorter time limit, or self-paced longer time limit) on items. I ran two models: a two-facet model with examinees and items as facets, as well as a three-facet model with examinees, items, and condition as facets.

For a dichotomous item in the two-facet model, the model can be expressed as:

$\ln (P_{ni} / (1 - P_{ni})) = B_n - D_i$

where

P_{ni} = probability of examinee n with ability B_n succeeding on item i with difficulty level D_i (Wright & Mok, 2004).

This model, which involves just examinees and item difficulty, does not take into account the facet of test condition. To allow investigation of bias or interactions among facets, I also ran a three-facet model. For a dichotomous item in the three-facet version, the model can be expressed as:

$\ln (P_{nij} / (1 - P_{nij})) = B_n - D_i - C_j$

where

P_{nij} = probability of examinee n with ability B_n succeeding on item i with difficulty level D_i in condition j with difficulty level C_j.

The two-facet model with examinees and items was a model conceptualized as involving examinee performance on 109 different items: 10 anchor items and 99 variable items (33 items * 3 conditions); I could thus compare item logit values, since a separate difficulty level was calculated per item. For instance, Q23 was separated into its Q23-listening-once version, Q23-self-paced-short, and Q23 self-paced-long depending on what input it had been presented with. With the two-facet model, there was less connectivity in the data, which is necessary for this item analysis; in other words, examinees did not all respond to every item in every condition. All examinees took the 10 anchor items, but then only 33 of the 99 other items each. The anchor items, which everyone took, allowed me to make comparisons among probability of success on the other 99 items.

In the three-facet model, the 1x *condition* facet was set at zero logits, to be able to compare other conditions in relation to that point. The three-facet model, with 43 different items (10 anchor items and 33 items in variable sets), allowed me to view condition as a separate facet and whether there were any bias or interaction effects of condition on items. Thus, for example, all three versions of item Q23 (1x, SPS, and SPL) were grouped together as the same item in the three-facet model, but as three different items in the two-facet model.

The two-facet model was required so that I could compare the individual logit (difficulty measure) values of different versions of the same item. However, I used the three-facet model for determining item fit and reliability statistics, because examinees really took 43 items, not 109, each, and because condition could then be modeled as a facet potentially affecting test scores. I also used Excel and the R statistical programming package (R Core Team, 2016) to compare item and examinee scores, and I conducted a content analysis of items to link scores with stimulus and test question traits.

I calculated means and standard deviations of responses to Likert-scale post-test survey items, and I read and sorted open-ended survey comments by examinees who preferred control separately from examinees who did not prefer control. I listened to interviews and simultaneously read transcripts, identifying key quotations, behaviors, and themes. I watched SR videos and made observation notes about when candidates used play controls. I then checked these video observations, revisiting interview audio and transcripts in an iterative fashion. Based on the SR videos, I examined where and when candidates paused or replayed. I coded interviews to determine why candidates said they stopped, replayed, or chose not to replay, and I enlisted a second coder for the ten transcripts and checked their codes against my notes. I used all of the above information, integrating SR participants' video behavior with their interview remarks, to determine whether the possibility to have control over the audio play had an effect on item performance and thus examinees' ability to display their listening proficiency.

The next chapter describes the results of the study.

4 RESULTS AND DISCUSSION

Here I describe the results by data source, followed by a discussion organized by research question.

4.1 Listening test results

4.1.1 Classical item statistics

Classical item statistics are presented here for test takers and items. The following figure presents a histogram, created in RStudio (an interface for the statistical programming language R), of the 100 examinees' raw scores on the listening test:

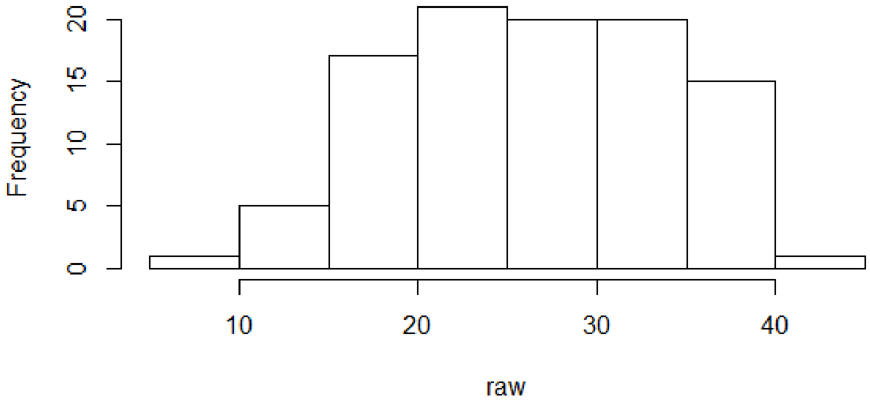


Figure 3 Histogram of raw listening scores, out of 43

Examinee scores ranged from 9 to 41 items correct out of 43. The mean score was 26.99 (SD 7.67) and the median 27.50. As can be seen from the figure, the data are slightly negatively skewed (skewness value of -0.18), with more examinees clustering graphically to the right of the average point; the left "tail" of the figure is longer than the right tail. Examinees fell into a pattern somewhat resembling a normal distribution (which would have a skewness value of 0). The skewness value of -0.18 indicates a slight lack of symmetry in the distribution, with a few more examinees on the right side of the average, suggesting that the test was rather easy for many test takers. Kurtosis quantifies the peakedness of the data. The kurtosis value is -0.94 (where 0 is the kurtosis of a normal distribution), indicating that the data are leptokurtic, or having a slightly higher peak with more clustering around the average than generally found in a normal distribution.

I also checked whether examinees who took Form A, B, or C were fairly comparable in terms of test performance. Form A examinees scored 26.00, Form B 26.60, and Form C 28.36, so Form C test takers did score slightly higher in terms of raw score. An ANOVA, however, indicated that these were not statistical differences in raw scores (*F* = .837; *df* = 2; *p* = .436). On the ten anchor items everyone took, Forms A, B, and C also did not statistically differ (F = .096; df = 2; p = .909).

I grouped all item conditions (1x, SPS, and SPL) together for the calculation of item facility (percentage correct) and discrimination figures. In other words, when I discuss the difficulty of an item, that facility value is a result of examinees' responses to all of its play control versions. A point biserial calculation was used for the item discrimination figure:

$r_{\mathrm{pbi}} = ((X_{\mathrm{p}} - X_{\mathrm{q}}) / s_{\mathrm{x}}) * (\sqrt{}(pq))$

where

r_{pbi} = point biserial correlation

X_{p} = mean total test score for students who answered the item correctly

X_{q} = mean total test score for students who answered the item incorrectly

s_{x} = standard deviation of all test scores

p = proportion of students who answered the item correctly

q = proportion of students who answered the item incorrectly (Hatch & Lazaraton, 1991).

Item facility and discrimination values are shown in Appendix C. Item facility values ranged from 15.84 to 91.09 percent of candidates responding to an item correctly, with a mean item facility value of 62.58 (SD = 21.01).

Overall, the most difficult question was item Q58, the second item of a four-item set with a monologue about an academic guest lecture:

> Listen to a professor speaking to her philosophy class.
> We've got just a few minutes before class ends, and I want to let you know about a public lecture that's scheduled for tonight in Dodge Hall. The lecture is on "Climate change and ethics: What do we know and what should we do?" We'll be turning our attention to ethics in a few weeks, so I'm hoping if you attend this lecture - it is optional - that it'll whet your appetite for the subject. Dr. Steven Willis, a professor from Central University, will be the speaker. He's a theoretical meteorologist, an expert on computer simulations of the atmosphere, and, and this is where he really diverges from most other

> climatologists, he's published articles in major philosophical journals. It says here in the flyer that Dr. Willis will summarize key scientific findings, and he'll examine policy options and their ethical implications, that is, what we should do about them. I believe he'll be discussing the question of intergenerational fairness, fairness between generations. It's an interesting idea. We're used to thinking about ethics in terms of the here and now, and dealing with questions about the relations between living beings. But one of the questions Dr. Willis is going to ask is whether those of us who are living today have ethical obligations to those who will inherit the earth hundreds of years from now. Most philosophers believe, and I concur, that we do. And frankly, based on the work he's done, I'd be surprised if Dr. Willis thought otherwise.
>
> Q57 What is the speaker's main purpose?
> a. to help students understand climate change
> b. to prepare students for tomorrow's class
> c. to encourage students to attend a lecture [intended key]
> d. to explain an article the students have read
>
> Q58 Why does the speaker think the event will interest the students?
> a. Dr. Willis will meet with them individually.
> b. They will write a paper on climate change.
> c. Dr. Willis is a famous philosopher.
> d. They will study ethics later. [intended key]
>
> Q59 What does the speaker say about future generations of human beings?
> a. People today have responsibilities toward them. [intended key]
> b. They will have the same problems as people today.
> c. Their lives may be more difficult than people's today.
> d. They will find a solution to climate change.
>
> Q60 What can be inferred about the speaker and Dr. Willis?
> a. They know each other personally.
> b. They hold some similar opinions. [intended key]
> c. They have degrees in philosophy.
> d. They have written about climate change. (CaMLA, 2014)

This monologue was full of low-frequency vocabulary and long utterances, especially more than other audio; compare to the dialogue and monologue in Appendix A. The items in the set were all quite difficult (facility values Q57 = 45.54; Q58 = 15.84; Q59 = 22.77; Q60 = 32.67). The easiest item was Q57, asking examinees to identify the main purpose of the monologue, but still rather difficult compared to the mean item facility value of 62.58. Items Q58 and Q59 required examinees to understand details, but because the text contains many propositions, test takers had to pay special attention to certain points the speaker made, in order to key the items successfully. The item Q58 ("Why does the speaker think the event will interest the students?" / "d. They will study ethics later.") required examinees to understand the utterance *"We'll be turning our attention to ethics in a few weeks, so I'm hoping if you attend this lecture - it is optional - that it'll whet your appetite for the subject."* The item Q59 ("What does the speaker say about future generations of human beings?" / "a. People today have responsibilities toward them.") hinged on listeners being able to understand *"But one of the questions Dr. Willis is going to ask is whether those of us who are living today have ethical obligations to those who will inherit the earth hundreds of years from now."* The item Q60 ("What can be inferred about the speaker and Dr. Willis?" / "b. They hold some similar opinions.") required successful inferencing based on the text and the context in which it was likely uttered, as well as possibly eliminating incorrect options from the item text.

Item point biserial figures were used to calculate how well the items were distinguishing test takers from one another. The closer to 1 the value, the better the item separates high-performing test takers from lower-performing examinees. Hatch and Lazaraton (1991) recommend a .20 to .40 discrimination value to indicate a good item; based on my testing experience, I used a cutoff of .30 or above to identify satisfactorily discriminating items. Main test items' point biserial values ranged from .19 to .62. Nine items had values lower than .30, indicating that they did not discriminate as well as the other 34 items. Four of the nine less-discriminating items had acceptable discrimination values between .25 and .30, so 38 of the 43 items discriminated well or acceptably. The items mentioned above in the set with Q57 to Q60, although quite difficult for the overall group, generally discriminated quite well (discrimination indices Q57 = 29.81; Q58 = 35.12; Q59 = 40.92; Q60 = 46.18), indicating that the most advanced listeners were correctly answering them.

I also examined classical item facility and discrimination values among different item conditions, which can be seen in full in Appendix C. Table 3 shows item conditions' facility and discrimination means and standard deviations:

Table 3 Descriptive statistics for item facility and discrimination

	1x		*SPS*		*SPL*	
	Facility	*Point Biserial*	*Facility*	*Point Biserial*	*Facility*	*Point Biserial*
mean	0.59	0.38	0.62	0.41	0.67	0.44
SD	0.20	0.19	0.23	0.15	0.20	0.16

Self-paced conditions, on average, did result in slightly easier items, with 59% of examinees correctly responding in the 1x condition, 62% for SPS, and 67% for SPL. However, these differences were not statistical, according to an ANOVA (F = 2.32, p = .13). The SPL version was the play condition that best discriminated best among test takers on average, with a .44 point biserial mean value, compared to .41 in SPS and .38 in 1x. These differences were also not statistical (F = 1.59, p = .21). Comparing across individual items, all of which can be seen in Appendix C, the SPL condition discriminated best for 14 of the 33 items. For 10 items, the 1x condition discriminated best, and for the other 9 items, the SPS condition best discriminated. Of the SPL condition items, only five had point biserial values of below .30; however, in the SPS condition 9 items did not discriminate well, and in 1x there were 11 items that did not discriminate well.

Some items were quite difficult overall in any condition, as with Q60. Only 17% of examinees responded correctly to it in the SPS condition, but the correct respondents were not always the most able test takers, as the point biserial value was .15. Item Q60 in the 1x condition had a facility value of .45 with a point biserial value of .53, while in the SPL condition it had a .38 facility value and a .72 point biserial value. It discriminated quite well in its 1x and SPL conditions but not in the SPS condition; it was a difficult item requiring inferencing beyond the text. It may also have been that examinees' time expired, as Q60 was the last item in its set, and this was a demanding input and set of items to respond to in only three minutes.

I also examined items by subskill tested. Of the 33 items of interest, 13 assessed the comprehension of a main idea or the key message of the input, 12 required test takers to understand supporting details, and 8 assessed understanding of information that required extension beyond the text, such as implicature, inferencing, prediction, speaker's purpose, or vocabulary in context. Items that best discriminated in one particular play condition did not share one particular subskill. For example, the 14 items that discriminated best in the SPL condition were all three types: four main idea items, six detail items, and four extension items. Therefore, one particular subskill of item was not overall more difficult than another; additionally, play condition did not seem to discriminate better among examinees for certain item types (e.g., extension items did not all discriminate best in the SPL condition).

From the classical item analysis, it appears that self-paced items with no time limit were slightly easier than self-paced timed, which were easier than listening-once conditions, but not remarkably so. Based on point biserial figures, items best discriminated in the self-paced condition with no time limit as compared to items in listening-once and self-paced timed conditions. Moreover, the SPL condition had the fewest items with a point biserial figure of less than .3: the 1x condition had 11 such items, SPS had 9, and SPL had only 5.

Reliability, calculated using Kuder-Richardson-21 (KR-21; Hatch & Lazaraton, 1991), served as a measure of items' internal consistency. The reliability value for the 1x condition was

.84, SPS .89, and SPL .87, indicating that the item versions in different play conditions were similarly internally consistent.

4.1.2 Rasch analysis

I next turned to an analysis that took into account item and play condition difficulty as well as person ability.

4.1.2.1 Variable map of examinees, conditions, and items

Here I describe the analytic scale and results used to compare examinees, items, and conditions. The variable map is presented here and described below.

```
+-------------------------------------------------------+
|Measr|+Examinees|-Condition      |-Items               |
|-----+----------+----------------+---------------------|
|   4 +  (more   +  (harder)      + (harder)            |
|     |   able)  |                |                     |
|     |          |                |                     |
|     | *        |                |                     |
|     |          |                |                     |
|     |          |                |                     |
|     |          |                |                     |
|   3 + *        +                +                     |
|     |          |                | Q58                 |
|     | **       |                |                     |
|     |          |                |                     |
|     | ***      |                |                     |
|     | *        |                | Q59                 |
|     | ******   |                |                     |
|   2 + *        +                + A01                 |
|     | *        |                | A06                 |
|     | **       |                | Q39    Q60          |
|     | ***      |                | Q42                 |
|     | *******  |                |                     |
|     | ***      |                |                     |
|     | *        |                | Q43                 |
|   1 + ****     +                + Q28    Q57    A08   |
|     | *****    |                | Q30    A02          |
|     | *****    |                | Q48                 |
|     | ****     |                |                     |
|     | *****    |                | Q29                 |
|     | *****    |                | Q26    Q45          |
|     | *****    |                | Q41                 |
*   0 * ******   * onceplayed     * Q31    Q47          *
|     | **       | SPS3minlimit   | Q35    Q50          |
|     | ****     |                | Q23    Q24    Q44   |
|     | ****     | SPLnotimelimit | Q27    Q38          |
|     | ****     |                | Q34    Q37    A07   |
|     | *****    |                | Q40                 |
|     | ****     |                | Q51                 |
|  -1 + *        +                + Q32    A09          |
|     | **       |                |                     |
|     |          |                | Q36    A10          |
|     | *        |                | Q33    Q46          |
|     | *        |                | A04    A05          |
|     |          |                | Q25                 |
|     | *        |                |                     |
|  -2 +          +                + Q49                 |
|     |          |                | A03                 |
|     |          |                |                     |
|     |          |                |                     |
|     |          |                |                     |
|     |          |                |                     |
|     |  (less   |                |                     |
|  -3 +   able)  +  (easier)      + (easier)            |
|-----+----------+----------------+---------------------|
|Measr| * = 1    |-Condition      |-Items               |
+-------------------------------------------------------+
```

Figure 4 All-facet vertical ruler

The first column of the variable map ("Measr", or Measure) shows the equal-interval log odds scale ranging from -3 (bottom of figure) to +4 (top of figure). In the other columns, examinee ability, item difficulty estimates, and condition difficulty estimates are plotted along the logit scale. The more positive a value on the plot, the more able the examinee or the more difficult the item. For an examinee and item at the same logit value, the person has a 50% probability of scoring that item correctly; if an item is 1.1 logits less difficult than a person is able (person logit value minus 1.1 logits = item logit value), the probability of success rises to 75% (Linacre, 2014). Test takers positioned higher on the variable map are at a more advanced proficiency level and thus have a higher probability of responding to items correctly; examinees lower on the variable map have a lower probability of item success. For this variable map, logit values are shown relative to the once-played condition logit value, which was anchored at 0 logits. Examinee and item logit values can be seen in more detail, with fit statistics and other data from the FACETS output, in Appendix D.

I must give a warning here about estimates for test taker ability and item difficulty. Likely because there were so few examinees per group (32 took Form A, 35 B, and 33 C), the model standard error (SE) is quite high: the average model standard error was .26 for items. It is ideal if this figure is as low as possible, generally lower than .10, because the higher the SE, the less accurate the measurement is. With many more examinees, the SE would probably be lower, so the results here need to be interpreted with caution.

The mean logit value for examinees was 0.60. Examinee logit values ranged from -1.89 (less proficient listener) to 3.55 (more proficient), spanning 5.44 logits. The separation index was 2.61, which meant examinees were divided into between two and three statistically distinct groups, with 0.87 reliability. The fixed chi-square value was 677.0 (df = 99, p < .01), indicating that people are significantly different in terms of ability measure, or listening proficiency. Form A, B, and C examinees did have slightly different average logit values, reflecting the trend from the raw scores: Form A examinees' average logit value was 0.46, Form B 0.56, and Form C 0.79. It would have been more ideal if these values were equal, but there do not appear to be meaningfully different (more than half a logit) values among examinees who saw different forms of the test.

I also was interested in item difficulty from the FACETS output. The mean logit value for items was -0.22. Because this is slightly lower than the examinee average logit value (of 0.60), the examinees were generally quite able with respect to this test. Ideally, for examinees and items to be well matched, their mean logit value should be roughly the same. If the spreads of items and examinees are drastically different, the test may have been far too easy or difficult. However, because the mean logit values of 0.60 for examinees and -0.22 for items are not more than 1 apart, in this statistical model, the examinees and items still appear to be mostly well matched.

Item measures ranged from -2.16 (easier item) to 2.86 (more difficult), spanning 5.12 logits. The separation index was 4.58, which meant items were divided into between four and five statistically distinct groups, with 0.95 reliability. The fixed chi-square value was 843.3 (df = 42, p < .01), indicating that items are significantly different in difficulty. Because there are roughly four to five item levels, the test is sufficiently distinguishing test takers into different proficiency levels overall. The higher the separation index for a test such as this, the better; a low separation index indicates that items are not differentiating examinees into enough ability levels.

For item model fit, values closest to 1 are ideal. Item fit ranged from infit mean square values of 0.79 to 1.23; one item (Q24) was slightly overfitting (InfitMnSq = 0.79) and another item (Q38) was misfitting (or underfitting; InfitMnSq = 1.23). I used a rather conservative measure of categorizing items with an infit mean square value of less than 0.80 or greater than 1.20 as not fitting the statistical model; this is more suitable to dichotomous items (Aryadoust, Goh, & Kim, 2011; Wright & Linacre, 1994). Item fit was thus satisfactory for 41 of the 43 items, suggesting that those items are tapping a similar listening construct.

4.1.2.2 Interactions among items and conditions

I next compared logit values among conditions. Table 4 shows the mean logit values, standard error, and confidence intervals by play type:

Table 4 Logit value comparison among conditions

	1x	*SPS*	*SPL*
mean	0.00	-0.18	-0.48
SE	.06	.08	.08
CI$_{.95}$	(-0.12, +0.12)	(-0.34, -0.02)	(-0.64, -0.32)

The 1x item mean logit value was anchored at 0 logits (SE .06), and SPS items' logit value was -0.18 (SE .08) and SPL -0.48 (SE .08), indicating that SPL items were slightly easier than SPS items, and SPS items were easier than 1x items. By estimating confidence intervals derived from the model SE, I calculated plus-or-minus 1.96 times the SE from the condition logit value (Bachman, 2004; Ockey et al., 2016). For example, for SPS items, because the model SE was .08, the reported logit value of -0.18 may really be anywhere from -0.34 to -0.02, or -0.18 +/-0.16. This means that, for the .95 probability level, the confidence intervals for the logit values of the three conditions are CI$_{.95}$(-0.12, +0.12) for 1x, CI$_{.95}$(-0.34, -0.02) for SPS, and CI$_{.95}$(-0.64, -0.32) for SPL. The claim that SPL items were easier than SPS which were easier than 1x also must be interpreted carefully because the confidence intervals overlap. (Studies such as Ockey et al. [2016] and Papageorgiou et al. [2012] consider logit differences of more than .5

logit to indicate meaningful differences in items.) Some values are more than .5 logits apart when confidence intervals are examined, so there are differences between conditions, but generally the intervals of logit values are quite close, so these differences may not be meaningful. However, again, this statistical model has some error due to the small sample size, so the claims must be treated carefully.

To check for differential item functioning (DIF) of items by condition, I ran a bias analysis in FACETS to analyze unexpected responses, or significant differences in observed versus expected responses to items, considering the interactions of those items with the facet of condition. Table 5 gives more detail about the bias size:

Table 5 Items with significant bias by condition

Item	*Bias Size*	*Probability (p < .05)*	*Bias and Condition*
Q59	.84	.0452	examinees performed **better** than expected in **1x** condition
Q47	.91	.0494	examinees performed **better** than expected in **1x** condition
Q31	-.95	.0266	examinees performed **worse** than expected in **1x** condition
Q60	-1.03	.0467	examinees performed **worse** than expected in **SPS** condition
Q33	-1.16	.0103	examinees performed **worse** than expected in **SPS** condition

Based on the statistical model, the bias sizes among the same items in different conditions range from an absolute value of .84 to 1.16. Examinees performed significantly worse than expected on items Q31 in the 1x condition and Q33 and Q60 in the SPS condition. They performed significantly better than expected on Q47 and Q59 both in the 1x condition. Two of these items had appeared on Form A, one on Form B, and two on Form C. This indicates that item biases were not experienced only by test takers of one form but may be a result of item presentation or other item features.

Q59 and Q60 were difficult comprehension questions, as the classical item analysis indicated. They were two of the items appearing after the set about a professor discussing an upcoming guest speaker event, a set containing a proposition-dense monologue with less-frequent vocabulary. On Q59 ("What does the speaker say about future generations of human beings?" / "a. People today have responsibilities toward them."), examinees performed better than expected when listening only once; this is unusual. Test takers may have taken advantage of extra plays or time in the self-paced conditions and possibly second-guessed their answers or initial hunches. On Q60 ("What can be inferred about the speaker and Dr. Willis?" / "b. They hold some similar opinions."), which was most difficult in the self-paced timed condition, examinees performed worse than the model expected for the SPS play condition. As mentioned previously, test takers perhaps had run out of time and/or had difficulty making the inference required to key the item successfully, as the correct answer was not mentioned directly in the announcement about the guest lecturer.

The data for the examinee, item, model fit, and bias/interaction analyses were all based on the three-facet model (examinees, items, condition). Because all versions of each item were included in its analysis (e.g., Q23's 1x, SPS, and SPL versions were not calculated separately), I next compared logit values for individual items using the two-facet model (examinees and items) to examine more closely test taker and item relationships.

4.1.2.3 Item comparison across conditions

I examined the two-facet model output in order to be able to identify the items with the most variation among conditions. Recall that a higher logit value means a more difficult item; the lower the value, the easier the item. The two items with the most variation were two that were flagged in the bias analysis: items Q31 and Q33. To say that a particular item is more difficult than another, however, their logit values need to be more than 3 SEs apart (Linacre, 2014). Table 6 gives more detail about the logit values and standard errors of biased items, with the unexpected values from the FACETS output in boldface:

Table 6 Comparison of logit values for biased items

	1x		*SPS*		*SPL*	
Item	*Logit*	*SE*	*Logit*	*SE*	*Logit*	*SE*
Q59	**1.57**	.40	3.01	.54	2.59	.50
Q47	**-0.69**	.45	0.66	.39	-0.32	.41
Q31	**1.21**	.41	-0.69	.46	-0.85	.44
Q60	1.11	.39	**2.73**	.51	1.38	.42
Q33	-2.51	.75	**-0.31**	.43	-2.18	.63

Q31 in its 1x version was a logit value of 1.21, over 1.90 logits (and thus more than 3 SEs) more difficult than its self-paced versions. This item follows the pattern I had expected, with 1x being more difficult than SPS, and SPS a slightly higher logit value (more difficult item) than SPL. However, Q33-SPS showed a somewhat unexpected pattern: its SPS version was more difficult than the 1x or SPL versions: over 1.87 logits more difficult (also above 3 SEs). This suggests that the time limit may have contributed to the item's difficulty. Aside from Q31, none of these items follow the expected pattern of SPL being easier than SPS and SPS being easier than 1x. Q60 being easier in its SPS version is particularly surprising. The logit values and standard errors of different versions of all items can be seen in Appendix D. Subskills tested included understanding the main idea, supporting details, and inference and implicature. Q31 was a main-idea item, Q59 tested a specific detail, and Q33, Q47, and Q60 required examinees to infer or understand vocabulary in context, so the items did not belong to only one category of subskill assessed.

Q31 and Q33 were part of a set, the next-day delivery dialogue that can be seen in Appendix A. Examinees performed worse than the Rasch model expected on Q31-1x as well as worse than expected on Q33-SPS. Q31, the item with a difference of more than three times the standard error, may have been substantially easier in the self-paced conditions (SPL logit value -0.85; SPS logit value -0.69) than in the 1x condition (logit value 1.21) because the information that keys the first item ("What does the woman want?" = "to have a package delivered the next day") is presented in the woman's and man's first turns:

> F: Hello, I'm wondering if it's too late to arrange for a next-day delivery to Los Angeles?
>
> M: It's too late to schedule a pickup. Our drivers are already out making their final rounds. But if you can bring your package to one of our offices, we can still guarantee next-day delivery. (CaMLA, 2014)

The language "next-day delivery" in the first two turns may be much more salient for examinees who had an opportunity to replay the audio from the beginning. In the self-paced conditions, Q31 was much easier. The item discriminated best in the once-played condition, but examinees also performed worse than the Rasch two-facet model expected on its 1x version. The speakers spend a turn each discussing at which street intersections the shipping company is located so that the woman can drop off a package, but this is not part of the tested information in the items.

Moreover, as Q33 is a replay-context item ("Listen to a part of the conversation again. Then answer the question: 'Oh, really? Great! I have plenty of time then.' Why does the woman say: 'Oh, really?'" / "d. She is surprised by what the man told her."), it required examinees to play a special audio stem to be able to hear the full item. Test takers had to remember that the woman had said "Oh, really?" and understand the context in which she uttered it, as well as the purpose for why she said it. Q33 had its own player requiring the examinee to listen again to part of the audio input, which may have introduced extra navigation difficulty on SPS sets where the player for the main audio was also accessible. The same phenomenon may have also occurred with replay-context item Q40, which was most difficult in the SPS condition (0.00 as compared to -1.16 in 1x and -1.38 in SPL) but not identified in the bias analysis. The other two replay-context items (Q30 and Q51), however, did not show this pattern; they were easier in SPL than in SPS, and most difficult in 1x. There may be issues beyond the items themselves that are causing trouble for listeners; the difficulty of the items may not be contingent simply on play condition, as the input contained some distracting information, and the extra audio player may have caused confusion for some test takers.

Of the 33 items presented in different conditions, eight of them (Q 26, 30, 31, 32, 36, 41, 45, and 51) followed the pattern I had anticipated, with their 1x version being the most difficult, followed by SPS, followed by SPL. According to the output from the three-facet model, Q31 was meaningfully different in its 1x version, with examinees answering it correctly more often in the self-paced versions. Aside from Q31, the difference was not three times the standard error, so there may not be a meaningful difference among item play conditions for most items. The other 25 of the 33 items showed mixed patterns, or even the opposite pattern, with regard to play condition potentially impacting item difficulty. Again, though, because the model standard error is rather high, any differences among items should be considered cautiously.

4.1.3 Qualtrics timing and click data

Qualtrics records the total time a survey user spends per screen as well as the number of clicks per screen. The timing data let me examine how long candidates spent on each page. I also was able to discern whether examinees used the whole three minutes of time allotted to SPS sets.

However, Qualtrics does not record *where* examinees click on the screen. Because I was not able to capture everyone's screen behavior using ActivePresenter, I thought I might have been able to use Qualtrics' click data as a proxy for interactivity, presuming that more clicks on the screen might have meant more instances of control on the SPS and SPL sets. I did not tell examinees that their clicks would be recorded, so some people may have clicked aimlessly on the test pages, and I did not inform them that their time-on-task would be calculated or that they ought to complete the tasks as quickly as possible, as I wanted them to carefully absorb the audio input and items. I also checked whether there was any connection between time on task and listening proficiency as operationalized by raw score or logit value.

Because some sets had three items and some four, I adjusted the click data by dividing the screen-click figure by the total number of items. Examinees clicked on average 2.24 times per test item. They clicked 1.79 times per item for 1x sets, 2.20 for SPS, and 2.72 for SPL, though naturally these figures will be higher for SPS and SPL because candidates could click on an audio player as well as on items. More clicking was also necessary for replay-context items, which had their own separate player, but not every set had a replay-context item.

Listeners spent an average of 2 minutes and 43 seconds on each page of the test interface. They spent 2:35 per 1x page, 2:31 per SPS page, and 3:02 per SPL page. The slightly shorter time on SPS pages compared to 1x sets could be due to the three-minute time countdown on SPS sets pressuring candidates to budget their time, as there was not a three-minute timer on the 1x condition pages. Because there was also no time limit on SPL pages, it fits that examinees spent slightly longer on those sets. Listening proficiency level, as measured by logit value, was not correlated with either time spent on page (r = -0.02) or number of clicks (r = 0.05).

4.2 Post-test survey results

Examinees responded to post-test survey items about the difficulty of and their familiarity with test items, as well as their reaction to the ability to have control over the audio play.

For the item "Did you prefer being able to have control over the audio?", 80 responded Yes and 20 No, or 4 of every 5 candidates responding Yes. However, as this item consisted of a binary choice, examinees had to choose a preference; they could not indicate shades of meaning or be indifferent, unless they typed their caveats in the open-ended comment boxes.

In the box for "Why did you choose Yes or No above?", candidates' responses sometimes concerned play controls. They commented that they used controls when their attention or concentration drifted, when they felt they didn't hear something well, and/or when they wanted to confirm or double-check a selected answer. Those who did not prefer having control remarked that it took away from their focus or concentration, or that they wanted to simulate real-life once-heard situations. The responses from candidates stating why they responded Yes or No are provided in Appendix E.

Recurring themes in examinees' open-ended comments included checking what they may have missed, confirming their understanding, and improving their listening ability. Five of the 80 examinees who preferred control commented that it made them feel more comfortable while listening. Several also remarked that they could better focus on reading the questions and options when they had control. For the "No" respondents, one of the 20 examinees who did not prefer control commented that it would reflect her "true hearing ability" by not having to listen again; others also seemed to have a similar sentiment about not desiring to take audio control because they felt it would not show their real listening proficiency.

For the following items, all "1" strongly disagree to "6" strongly agree, descriptive statistics are listed in Table 7:

Table 7 Post-test survey responses

Survey Item	*Mean*	*SD*	*Median*	*Correl. w/ Logit*
PTS1: The topics presented in the listening sets were easy.	4.04	1.16	4	0.21
PTS2: I did better on the test items when I was able to control the audio.	4.37	1.52	5	0.04
PTS3: I was familiar with the topics covered.	4.06	1.22	4	0.07
PTS4: This is a good test of my listening ability.	5.01	1.31	5	0.15

Examinees generally responded that they agreed that the test was a good assessment of their listening skills, though these items concerned the test in general, not any one specific play condition. They also somewhat agreed that topics were easy and familiar, and they agreed they felt they did better with the ability to have control on the item sets. I was particularly interested whether there was any connection between item PTS2 "I did better on the test items when I was able to control the audio" and examinees' listening ability. PTS2 had no (r = 0.04) correlation with listening proficiency as measured by logit value. For that survey item, the average was 4.37 (SD = 1.52), minimum 1, maximum 6.

I also checked whether there were any differences in responses to PTS2 depending on whether candidates took Form A, B, or C of the test. Because the data were not normally distributed, I ran a non-parametric test of differences in means. A Kruskal-Wallis non-parametric ANOVA (H = .542; df = 2; p = .763) indicated that there were no significant differences in this response among the three groups. This suggests that no one form, or arrangement of items in a certain condition, biased examinees in making their choice for the survey item about performing better with control.

Additionally, the responses to PTS2 only had a low (*r* = 0.14) correlation with IEP or graduate student level. This suggests that the perception of better performance during the option to pause or replay did not have a strong statistical relationship with student proficiency as operationalized by IEP or ESL level. Other post-test survey items, similarly, had little connection to IEP or ESL level (PTS1 *r* = 0.21; PTS3 *r* = 0.04; PTS4 *r* = 0.17).

4.3 SR interview findings

As noted above, 100 examinees took the main 43-item test, and for two of them I also collected SR interview data on those items. Eight IEP examinees took the supplementary 22-item test and participated in interviews immediately after having responded to the supplementary sets. The eight IEP students took the supplementary test one to three weeks after they had taken the main test in their classes without an interview specifically about the main test items.

Because I was interested in not only whether candidates exercised audio control but also why they said they did so, I focused on their introspective remarks, using verbal report methodology. For all ten interviewees, I had planned to collect screen-capture video, but for two of the IEP participants, ActivePresenter had caused the mouse to flicker so much on the screen that sometimes people couldn't see where their cursor was while taking the listening sets. When the flickering became too severe, examinees told me and I stopped the screen capture software, which let them see their mouse navigation more clearly. With those two people, I conducted audio-recorded interviews without SR.

The other eight interviewees (two graduate students and six IEP students) watched the ActivePresenter video screen capture, played back to them as the stimulus for SR as I audio recorded. When I would play the video of participants' mouse clicks and movements, they often had a fascinated or amused initial reaction to the mouse moving on the screen seemingly by itself. One participant even "cheered on" the mouse trail at different points of the interview with exclamations such as "Come on!", hoping that his past self on the video screen had selected the correct answers.

Interviewees' comments generally focused on selecting what they had felt were the right answers on the test. When I asked questions such as "what were you thinking about at that time?", examinees generally did not discuss strategies or describe how they felt about listening; they tended to focus on the content they had heard. I did not want to steer people toward the discussion of any cognitive or metacognitive strategies that I knew about, since I wanted to hear their frank opinions. Thus, I generally did not prompt anyone with questions other than "what had you been thinking?" or "what was on your mind?", and I sometimes asked for clarification of remarks people had made. Still, I was surprised that more people's comments did not center on their listening behaviors or processes. Some people discussed why they chose a certain option or another, drawing on information from the stimuli; these are perfectly logical responses to a question such as "what were you thinking about during this part?", although such replies did not focus on metacognitive behavior about listening but rather seemed to be concerned with seeking the correct answer in each instance.

Transcription conventions for quoted excerpts from interviews include the following: A comma indicates a brief pause of one to two seconds with non-phrase-final intonation, and three dots indicate a pause of more than three seconds. A period indicates phrase-final falling intonation even if not at a clear phrase boundary. An underscore indicates a word that was completed but the utterance was cut off, generally followed by a rephrasing or restating, while a hyphen immediately followed by a space indicates a cut-off syllable. An all-capitalized word indicates emphasis with a raised pitch contour.

Based on the SR videos and interview data, I referenced my notes about when examinees clicked on items and whether they took control, linking those to what participants said about self-pacing. I describe first the findings from the two graduate students, followed by data from the eight IEP students. All participant names used throughout the manuscript are pseudonyms.

4.3.1 Graduate student interviews

Table 8 shows the two graduate student interviewees who participated in stimulated recall on the 43 main test items. Included are their score on the main test by logit value and percentage, their gender and first language, and time spent on average on the three sets seen in each condition.

Table 8 Graduate student interviewee information

Name	*Logit value*	*Score on main test*	*Gender*	*L1*	*Average time spent on 1x sets*	*Average time spent on SPS sets*	*Average time spent on SPL sets*
Xinyi	1.74	81.40%	M	Chinese	01:31	02:41	02:46
Roshan	2.17	86.05%	M	Persian	02:21	02:45	03:39

Xinyi and Roshan each took the ten anchor items, followed by three 1x sets, three SPS sets, and three SPL sets. They scored 35 and 37, respectively, on the 43 main items, with logit values of 1.74 and 2.17. The top two of the 100 main test listeners' logit values were 2.97 and 3.55; Xinyi and Roshan were among some of the most able listeners with respect to these test items, scoring in the top one-third of examinees. Based on the main listening test, and also due to the fact that they had gained graduate admittance to an English-medium university, they can be characterized as advanced-proficiency-level listeners. They both spent increasingly more time on average from the 1x sets to SPS to SPL.

Both candidates took the 3 SPS sets and 3 SPL sets on Form B of the main test. The SPS sets were a dialogue about a laptop warranty (Q37-40), a dialogue about a retiring coworker

(Q41-43), and the climate change lecture monologue (Q57-60; the very dense speech event mentioned previously in the quantitative results section). For the self-paced audio with no time limit, the SPL sets were a dialogue about viewing an apartment (Q23-26), the next-day delivery dialogue (Q31-33), and a monologue about a research study on teens using the internet to discover music (Q44-47). The sets with items Q31-33 and Q44-47 can be found in Appendix A. Table 9 provides details about how often the two examinees used audio controls on self-paced sets:

Table 9 Graduate student control use

Name & Instances of Control	*Instances of control during SPS sets*			*Instances of control during SPL sets*		
	Pausing & Unpausing	*Stopping*	*Replaying*	*Pausing & Unpausing*	*Stopping*	*Replaying*
Xinyi **14**	N/A	1 (in 1st set)	5 (1 in 1st set, 2 in 2nd set, 2 in 3rd set)	1 (in 3rd set)	1 (in 2nd set)	6 (3 in 1st set, 2 in 2nd set, 1 in 3rd set)
Roshan **12**	N/A	1 (in 2nd set)	2 (in 2nd set)	4 (1 in 2nd set, 3 in 3rd set)	N/A	5 (1 in 1st set, 2 in 2nd set, 2 in 3rd set)

Both Roshan and Xinyi engaged in pausing and unpausing as well as replaying of the audio. Replay does not necessarily mean another full play of the audio from beginning to end; generally replay was partial plays, sometimes as short as a few seconds. *Unpausing* refers to resuming play from the place the audio was temporarily stopped, but without playing to rehear part of the audio; otherwise, I categorized the behavior as *replaying*. I also categorized *stopping* on its own separately from *pausing and unpausing*; sometimes candidates would discontinue play without resuming, and usually that was to respond to items or proceed to the next screen.

Roshan used play control on only one of the SPS sets but on all of the SPL sets, engaging in pausing and replaying controls a total of 12 times on the six sets. Xinyi used the controls 14 times throughout all six sets, replaying a portion of each set and also using the pause/play button on one of the SPS sets and two of the SPL sets. For both, play control was used approximately 2 times per set. Xinyi and Roshan's detailed behaviors on the sets, along with SR interview comments where relevant, are presented in Appendix F.

Xinyi preferred having control on the main test. He reported that he liked being able to pause or replay in order to not have to read the items at the same time as listening to the audio input. He also wanted to check that his mind was on track, as well as to verify that he hadn't missed something. Xinyi took control to be able to engage in process of elimination of incorrect multiple-choice options, as well as to determine how a speaker felt in a certain situation. Remarking that having the play control allows him to feel more comfortable, he generally engaged in one-and-a-half to two plays of each main audio.

Roshan also did tend to prefer control. He said that he controlled the input in order to have the ability to check or confirm what he had heard, especially to confirm his notes or in case his attention had drifted. He felt, however, that he shouldn't use controls too much, explaining that because in real life one doesn't always have unlimited time to revisit what has been heard, and on high-stakes listening proficiency tests there is almost always a time limit. For one self-paced set, Roshan had a unique reason for replaying: because he seemed to feel that the topic was interesting and just wanted to hear the information again. He did this on the set with a man discussing the results of the research study about teens using the internet to find out about music, a self-paced set with no time limit.

On the last set Roshan and Xinyi heard, with items 44 to 47, they each used play controls on more than one occasion, sometimes for similar and other times for different reasons. Both wanted to be sure what they heard was correct. Roshan took play controls in part because he felt the information presented was intriguing. However, there may be issues with the wording of items, not necessarily relating to the opportunity to take play control. Both Xinyi and Roshan expressed concern with how one of the items, Q44, was phrased:

> Q44 What was the research study about?
> a. how often teenagers use a computer
> b. where teenagers buy their CDs
> c. what kind of music teenagers like
> d. how teenagers use the Internet [intended key] (CaMLA, 2014)

For Xinyi, this item confused him to the point where he felt there was no correct answer. Although the audio concerns how teenagers use the internet to explore bands or music, option D does not directly address music. He thought option D was too general but eventually selected it because the other options were inaccurate enough. To reduce his confusion about the item, he took control of the audio play. Roshan also commented on this same item, replaying this set not only because he felt it was interesting but to ensure that what he had heard was correct in order to key item 44:

> Roshan: Because this question is ambiguous for me. [Sarah: ah uh-huh] how often use_ you know_ uh "how teenagers use the internet". I know that. I remember the most part is about the music, what kind of the music you know. where teenagers buy this (it's not) C-Ds_ the_ i only hear that boys only download music and they write it.
>
> Sarah: Uh-huh, so D that option didn't have music in it so it made you confused?
>
> Roshan: Yes, and how often? no. I'm sure that these three are not correct answer but (important xx is the correct answer). Most participants would. Is it the correct answer?
>
> (Roshan and Sarah, interview comments)

For this item, then, there appeared to be a problem not directly relating to play control but rather to how the intended correct answer was worded. This main idea item was of average difficulty for the 100 examinees on the main test, with an overall -0.31 logit value (0.00 was the average item difficulty, with positive values indicating more difficult items and negative values easier). It tended to be most difficult overall in the 1x play condition (0.54), then easier in SPL (-0.67) and even easier with SPS (-1.16). However, in this case, the item difficulty may not be being impacted only by play condition but also by the item text. These retired items may have been problematic in operational versions of the Michigan English Test; the fact that two quite proficient listeners both commented on this item may mean that its wording is unnecessarily confusing.

Xinyi and Roshan were two proficient listeners who both engaged in taking control of the audio play. According to the Qualtrics and ActivePresenter output, both students tended to click about two times per item. The SR interview findings revealed how they played the listening input audio. However, Qualtrics clicks and timing of other examinees around their same proficiency level differed widely, and examinees engaged in similar quantities of clicks or spent similar time on the same sets but had quite different ability levels. As shown previously, the click and timing data did not have a direct relationship with listening proficiency as measured by logit value, so a closer look at particular listening behaviors was necessary.

I next turn to the interviews I conducted with eight Intensive English Program students.

4.3.2 IEP student interviews

The other eight of the ten interview participants, IEP students who had already seen the 43 main items, saw 22 different items. As in the main test, at the beginning of the supplementary test, ten short warmup questions of similar format to the ten anchor items (short dialogues) were played once without examinee control permitted; I included them so that follow-up examinees would experience an equivalent introduction as on the main sets. All IEP interviewees then heard three sets, all monologues in the 1x, SPS, SPL order that contained four items in each set.

Table 10 provides information about the interviewees. Examinees are sorted by logit value from the main items, and the table also includes their score on the main items, Intensive English Program level, first language, gender, and score on the supplementary items.

Table 10 IEP student interviewee information

Name	*Logit value*	*Score on 43 main items*	*IEP level*	*Gender*	*L1*	*Score on 22 supplementary items*
Myung	0.10	55.81%	4	M	Korean	36.36%
Almas	0.23	58.14%	3	F	Arabic	54.55%
Octavia	0.49	62.79%	4	F	Spanish	63.64%
Gisela	1.06	72.09%	4	F	Spanish	86.36%
Yijun	1.56	79.07%	4	M	Chinese	68.18%
Estavan	1.56	79.07%	4	M	Spanish	77.27%
Celeste	2.19	86.05%	5	F	French	81.82%
Liliane	2.19	86.05%	3	F	French	86.36%
Average	**1.17**	**72.38%**				**69.32%**

Of these eight interviewees, five were female and three male. Two participants were enrolled in Level 3 (of 5 levels) courses in the Intensive English Program, five were in Level 4, and one was in Level 5. Their percentage scores follow a similar pattern for the main items as the supplementary items, though Gisela, a slightly above-average performer on the main items, scored exceptionally higher on the supplementary items. The eight interview participants' logit values on the main test ranged from 0.10 to 2.19. Their average logit value was 1.17 on the main items (the average logit value of the group of 100 examinees was 0.60). They had scored on average 72.38% on the main items and 69.32% on the supplementary items. Based on the listening tests, they can be characterized as lower-intermediate to advanced proficiency listeners.

I had planned to use screen capture for all interviewees. However, for Myung and Liliane, I was not able to video capture their detailed test behavior, so they participated in interviews without a stimulus for recall. Of the interview participants, they were the most proficient (Liliane) and least proficient (Myung) as measured by their score on the 43 main items. They offered their comments regarding the 22 supplementary items, which were similar format and themes as the main items; the setup concluded with two self-paced sets, one with a three-minute timer and the other no time limit. The SPS set input was a monologue of a tour guide discussing a rooftop sculpture garden, and the SPL input was a monologue of a zoo representative describing changes made to a penguin exhibit.

I next describe Liliane and Myung's interview comments, followed by Qualtrics click and timing data and stimulated-recall interview findings from the other six interviewees.

4.3.2.1 Time limit and control: Liliane and Myung's comments

Two key variables that likely impacted examinees' listening play behavior, according to almost all interviewees, were time limits of the SPS set and computer navigation; these caused people to interact with the input in different ways. Their preferences about play control did not always match their documented test behavior. For Myung, it was unclear whether he preferred having control over the audio play, since he was also focusing on familiarizing himself with the item format and concentrating on the listening audio. He did use play controls, but he said early in the interview that he did not like to take control because he couldn't focus on the input. Some of his comments, however, seemed to suggest that he preferred the version of the test with controls ("the better ones is, the rewind ... I think it is good. The listener can use that in the limitation of time" [Myung, interview comment]). Overall, although he used control for the self-paced sets, his interview responses indicated that if he wanted to take that control, he would also need to keep in mind the time limit and monitor his understanding of and concentration on the items.

Liliane also reported that she favored having a time limit on this test. She seemed to feel that a time limit made the test more challenging, and that she likes a challenge. For these sets, she preferred when the audio player was hidden and she didn't have to control it because she felt she could concentrate better. These monologues were 1 minute 43 seconds and 1 minute 26 seconds. Although Liliane stated that she preferred not to have control on these sets, she reported that she might want the option to be able to control the input on longer listening sets, such as the minilectures on the TOEFL listening section.

Navigating the computer interface also posed some difficulties for Liliane. She had some trouble clicking on the places of the screen she wanted, especially with moving the position of the audio slider bar to where she had wanted to review the input. Additionally, the mouse flickering during the screen capture process exacerbated this. During Liliane's interview, we stopped video capture after she commented that the mouse was flickering too much. Her comments indicated that she attempted to use play controls in order to check and confirm her listening test responses, but that she didn't prefer having control on these items. It is noteworthy that her sentiments were similar to those of Myung about using control but not preferring it, despite them being at different listening proficiency levels.

4.3.2.2 Interviewees with stimulus for recall

For the other six of the eight IEP interviewees, I conducted SR interviews during which they watched the video stimulus of their mouse clicks and movements from taking the test. Table

11 presents the SR participants' pausing and multiple-play behavior, sorted from lowest to highest score on the supplementary items.

Table 11 IEP student control use

Name	*SPS stopping*	*SPS pausing and un-pausing*	*SPS replay*	*SPL stopping*	*SPL pausing and un-pausing*	*SPL replay*	***TOTAL instan-ces of control***	*Score on 22 supp. items*
Almas	1	1	1	3	2	3	**11**	54.55%
Octavia	none	none	none	none	none	none	**0**	63.64%
Estavan	none	none	none	none	none	none	**0**	68.18%
Yijun	1	1	1	1	none	1	**5**	77.27%
Celeste	none	none	1	1	none	2	**4**	81.82%
Gisela	none	none	1*	1	none	1*	**3**	86.36%

*Gisela appeared to want to take play control in two instances but, due to the screen capture software causing the processor to lag, had some computer navigation difficulties.

There was a great deal of variation among use of play controls, from no control taken to 11 instances taken of pausing, starting/stopping, or replay. Octavia and Estavan did not engage in any replaying or pausing, while the other four did use some degree of play control. Control taken did tend to decrease by proficiency level as operationalized by the supplementary items, except for Estavan and Octavia. I wanted to observe not only whether candidates took play control but why, so I present the evidence from the interviews when examinees did explain why. The themes of timing conditions and computer interface manipulation were consistently present throughout the interviews, with many interviewees commenting on how those variables may have affected their listening performance.

Celeste tended to replay the self-paced sets more than once but not two full times. The SPS set given during SR, with its three-minute time limit, could not be played two full times because it was 1 minute 43 seconds in duration, which Celeste remarked on during the interview.

She said that she liked having play control because she tried to find the key words to be able to answer the items, strategizing about where in the audio clip she thought she had heard important information and replaying those pieces of the stimulus. Celeste highlighted a key feature of listening during her interview comments, which is that, unlike reading, one cannot usually go back:

> Yes. I think it's better because, when you're listening it's not_ it's not like when we're reading and we can go back and read. But when we are listening I think it's better when we are able to to go back and listen again. Because we can't think, while listening. So we want to listen then think and listen again. I think before reading. I don't know. And I think yes. (Celeste, interview comment)

From this and her other comments, it seems Celeste does tend to prefer having control over the audio stimulus play, particularly with respect to being able to review as she can when reading. I asked her whether she would prefer having control not only in this practice test but also in a standardized test situation, and she said yes but that she would still have to pay attention and listen carefully. She was also concerned about her reading abilities, commenting that she has to keep in mind what she reads during a test, as well as what she hears, so this would likely impact whether to not she chooses to take play control.

Almas, one of the less proficient listeners, also preferred having play control. She seemed to want to repeat the input play in order to be able to pay attention to the audio. She did, however, appear to have some trouble locating where in the audio input the tested information was. When I noticed on the video that she had paused close to the beginning of the audio play, I asked her why she repeated the play at the beginning, and she replied that she thought perhaps that was the location of some of the essential information required to key the items. The start of each input audio contained a scene-setter sentence delivered by a narrator, such as, "Listen to a tour guide at a museum," so she may have wanted to check that she understood the setting for the monologues presented. Almas preferred control and also frequently took that control. However, in contrast to Celeste, whose listening scores indicated she was understanding these monologues, Almas was perhaps still seeking where to find key idea units or tested information in the input.

Yijun scored around average and did take control on self-paced sets. In his interview, he commented that he would prefer listening just once. Like many others, however, he did actually engage in some replaying, listening again to confirm what he thought he had heard. Like Celeste, he didn't pay attention to the timer on the SPS set, so the interface bumped him to the next page after three minutes had passed, before he could give a response to all items. He pointed out a key issue with replaying, remarking, "I would prefer don't control it, because if you control it you can just check answer all the time, and you don't know where you are in your level or what you know" (Yijun, interview comment). This indicated that he perceived one audio play as the best listening setup to measure his listening ability. Regarding the interface design, I asked Yijun whether he used the numbers on the timers, either on the main timer or the audio player (on the player, a timer counting up was visible at the left, with a timer counting down visible at the far right). He responded that he didn't pay attention to the timer. He noted that I might want to mention the existence of the timers to test takers more explicitly so that they understand what they can do with the audio navigation and keep track of their time limit. He did give a caveat, though, that people may feel too stressed seeing a timer. These comments suggest that for Yijun, although he did sometimes want to check his responses, for him listening once was best.

Estavan, who did not take control of the input and scored about average, also had issues with the visible timer. He found it distracting seeing the timer on the SPS sets because he reported that it disrupted his concentration. He stated that he didn't replay because he only wanted to focus on listening first, then later remember what he had heard as he went on to respond to the items. However, at one point, he did mention that having the controls "could be a good option" (Estavan, interview comment) because sometimes he tried to anticipate what the tested information would be, based on what he had heard. Then, if the audio he heard didn't match those predictions, he would perhaps have missed key details, unless having the option to hear audio once more.

What happened with Estavan at one point during the SR added another level of replay. Because candidates watched the video of their screens from the test, they were able to hear items again after scores had been submitted, even though they were only designed to be played once during the practice test itself. Thus, at one point, on an item where he had to engage in inferencing behavior, he discussed what he had been thinking the conversation said, then explained that at the time of the interview he was more certain about his answer. This opportunity during the SR interview afforded a chance to discuss why replaying may have been helpful, although it was not possible for this item during the test itself. On the test, he had just listened to a short dialogue about purchasing movie tickets (F is a female speaker and M male):

> M: How was that new movie you went to see yesterday?
> F: You know, it's really funny. When we got there, all the shows were sold out, so we ended up getting tickets for later today.
> M: Wow. I didn't realize it would be so popular.
> F: We didn't either.

> D10 What can be inferred about the woman?
> a. She doesn't think the movie is funny.
> b. She'll watch the movie today. [intended key]
> c. She'll try to buy a ticket later.
> d. She doesn't want to see the movie. (CaMLA, 2014)

The transcript does not reflect clearly that the female turn "it's really funny" has a relatively low pitch, meaning the woman in the stimulus was commenting on the situation, not on the movie. As can be seen, though, the test developers had designed distractor "A" to play off a potential misunderstanding of that utterance. After I asked Estavan what he had been thinking, he responded:

> Because ... because I, when I listen I understand. In that part was yesterday and, I think that when she say that it's funny I think that she refers to the movie. And and and and, I say, oh, maybe she mm watch the movie today because she buy the tickets yesterday. (Estavan, interview comment)

With the possibility for Estavan to hear the item again, although it was not during the test itself but rather in the interview, he was able to confirm his hunches about his understanding of the utterance "it's really funny", not being distracted by the incorrect option for the comprehension item. Generally, though, he felt he did not need to replay; in fact, he did not replay the main audio at all during the supplementary test session.

Gisela, who was one of the highest scorers on the supplementary items, liked being able to take control:

> Mm, I prefer those when I can go back I_ go back. maybe I realize mm that in this time because when I, did the, the other audio in [instructor's name]'s class, I thought the opposite but now I realize that it's better if I can control, because maybe I want to clarify something and I can manage the time. (Gisela, interview comment)

Gisela was concerned about her performance on this test not only due to the time limits but also with the mouse flickering during ActivePresenter screen capturing. She stated that she had been more comfortable with the main test sets that she took in her IEP class; the difficulty being able to see the mouse on the screen made her feel less at ease with the supplementary sets. She had wanted to replay the SPS set, but by the time she wanted to listen to the audio again, it was time to move forward. On the SPL set, Gisela replayed in order to check key words that she thought she had heard, listening again to compare that what she thought was the right answer with what she listened to later. She also noted that the audio input was slightly longer during the supplementary items than the main test audio, so she also stated that she wanted to replay the supplementary audio input because the monologues were longer than in the main test, and she felt they may have contained more information. This parallels the comments Liliane made about desiring play control capability for longer listening input such as TOEFL minilectures.

Octavia's interview comments indicated that she does not prefer having control over the audio play. In fact, she did not take control. She said, "It is better for me focus on the lecture than focus if I have to change the, things for the volume or something like that" (Octavia, interview comment). For her, she may have interpreted "control" in the survey as volume control or another setting that could be changed, rather than main audio play position. Her interview comments suggest that she felt she could perhaps understand the topic better with another play, but this was not her preference. As Yijun also remarked, practicing listening once felt more effective for improving listening skills. Similar to other candidates, Octavia felt a great deal of time pressure. Sometimes she did not or could not replay because time on a set had run out. The time factor potentially interacted with candidates' ability to replay and perceptions about replaying.

4.3.2.3 Click and timing data from stimulated recall

I also examined the clicks and time spent per page from Qualtrics for the supplementary test interviewees. The data in Table 12 show the timing and clicks for the rooftop sculpture garden (SPS) set and the penguin exhibit update (SPL) set, as well as examinees' raw scores:

Table 12 Clicks and timing on supplementary sets

Participant	*SPS timing*	*SPS clicks per item*	*Score of 4 SPS items*	*SPL timing*	*SPL clicks per item*	*Score of 4 SPL items*
Myung	03:00	4.00	1	03:48	3.75	2
Almas	03:00	3.00	3	04:27	4.75	3
Octavia	02:55	2.00	1	02:38	1.75	3
Gisela	03:00	2.50	2	03:15	2.00	4
Yijun	03:00	1.00	1	02:07	2.25	4
Estavan	03:00	1.75	3	02:54	1.75	4
Celeste	03:00	3.75	3	02:28	2.75	4
Liliane	02:45	1.25	3	02:50	1.75	4
Average	**02:57**	**2.41**	**53%**	**03:03**	**2.59**	**88%**

Examinees clicked on average between two and three times per item on self-paced supplementary sets. The data are regarding participants' performance on just eight items, however, so any generalizations about self-paced listening proficiency or play tendencies would need to be made cautiously. The participants performed better on the SPL than the SPS set. On average, they only spent six more seconds on the SPL untimed set than the SPS set, where most used the whole three minutes they had been allotted; however, the SPS set had been 17 seconds longer than the SPL set, although input and items were overall designed to aim at similar proficiency levels.

As with the main items, there was not a clear pattern on the supplementary items between clicks or timing and examinees' scores. There tended to be fewer clicks and time spent with a higher number of correct item responses, but not for, for example, Celeste with regard to clicks. She tended to click more than participants with comparable proficiency levels, but this could have been due to considerations in reading the questions and options, or difficulties manipulating the mouse cursor and/or audio slider bar. The one self-paced item Celeste missed, the last item of the SPS set, was because she ran out of time on the three-minute countdown before she could respond to it. This was a vocabulary-in-context item about the exhibit being open "weather permitting":

> Listen to a tour guide at a museum.
> Now, for the final stop on our tour. The rooftop sculpture garden. And in my opinion, we've saved the best for last. This special collection of sculptures is complemented by this beautiful location. We're nearly twenty stories above the city streets and, as you can see, there's an unobstructed view of the skyline and the park below. It's especially beautiful at night, when the lights are twinkling, so I encourage you to come back some evening. The sculpture garden is open every night 'til ten P-M, weather permitting. Thanks to gifts from generous private benefactors and corporations, the museum was able to create this rooftop sculpture garden in nineteen ninety-seven. It always features three to four pieces of modern sculpture from the museum's permanent collection. Every six months, the exhibit is changed. Each exhibit displays the work of an individual artist, and what you'll see today is the sculpture of Michael Boyle. Boyle was an industrial engineer before becoming a sculptor. His works are made from scrap metal he finds in junkyards. As you look at the pieces, you'll notice objects such as machine parts, chains, wires and steel pipes. Boyle uses discarded metal objects to create symbols of nature. The four sculptures in this exhibit are simply titled Tree, Flower, Shell and Waterfall. This concludes the tour. Please walk around and enjoy the sculptures. If you'd like something light to eat or drink, the rooftop café is to your left, and serves beverages and snacks.
>
> SPS-Q1 What does the speaker recommend visitors do?
> a. return to the exhibit at night [intended key]
> b. take photographs
> c. donate money to the museum
> d. visit the gift shop

SPS-Q2 What type of material is used in the sculptures currently on display?
a. seashells
b. wood
c. metal   [intended key]
d. recycled paper

SPS-Q3 What does the speaker say the sculptures represent?
a. human figures
b. the natural world   [intended key]
c. machines
d. modern life

SPS-Q4 What does the speaker mean when she says: "the sculpture garden is open every night 'til ten P-M, weather permitting."
a. if the weather is good   [intended key]
b. if visitors call in advance
c. without exception
d. for visitors with a special pass (CaMLA, 2014)

In the stimulated recall, Celeste had asked me what the answer was for item SPS-Q4, and we had this exchange:

Sarah: for this exhibit 'cause, it's an outdoor exhibit, she says if the weather is good or_

Celeste: ohh! (if the) weather_

Sarah: she says "weather permitting" yeah.

Celeste: ohhh i see!

Sarah: that's like_ that's like a phrase, people will use for an event, if, if it's happening. yeah.

Celeste: "weather permitting" oh okay.

Sarah: that's_ it might be open if the weather is good, but if the weather is not good then the exhibit might be closed.

Celeste: i (wasn't_ didn't think) about weather, honestly. <LAUGH> weather.

Sarah: 'cause th- we also have W-H-E-T-H-E-R whether, like whether or not.

Celeste: (i was thinking "whether" yeah. that's what i said.) that's what i thought.

Celeste explained that she wasn't thinking about "weather" but rather "whether" when she heard the speaker. She did recognize the sound shape (Cauldwell, 2014) that had the phonological form /wɛðɚ/ but did not link it to the meaning of "weather" referring to the environment, specifically the possibility of rain impacting an outdoor exhibit being open. Thus, it makes sense that, in some instances, Celeste may have wanted to replay to be able to resolve incongruences in what she thought she had heard.

4.3.3 Justification for or against replay

Based on the eight IEP interviewees' comments, in Table 13 below, I summarized why they said they did or did not engage in multiple plays of the listening audio.

Table 13 Interview comment summary

Name	*Why replayed*	*Why didn't replay*
Myung	to effectively use time; to confirm responses	to be able to concentrate on the audio
Liliane	to confirm responses	to be able to concentrate on the audio
Celeste	to check key words; to follow the order of presented ideas; to understand the topic better	N/A
Almas	to pay attention to the audio; to locate key information	N/A
Yijun	to confirm responses	to avoid frequently checking his answers
Estavan	N/A	to be able to concentrate on the audio
Gisela	to check key words	N/A
Octavia	N/A	to be able to concentrate on the audio

Eight candidates gave some justification for why they replayed, while five gave reasons why they did not replay, with three of the interviewees describing their feelings about both. Generally, examinees were very focused on responding to the test items, a pragmatic reason for listening relevant to this setting. Being able to check or confirm what they had heard often were responses that occurred in some interviewees' remarks, for rationale both for replaying and for not replaying. For some participants' sentiments about not deciding to replay, it was often because they wanted to be able to focus more closely or effectively on the audio.

From the interview findings, it seems that examinees are balancing their own listening skills with reading skills, computer navigation, and timing in a test such as the ones they completed. They gave a number of comments regarding replaying, starting and stopping, and reading the item questions and options, all while attempting to understand what they had heard. It is evident that play condition is not the only variable impacting item results, so it is difficult to know how play condition alone is interacting with how examinees can best show their listening proficiency level.

In the next section, I return to the quantitative data to discuss the results of the Rasch analysis and post-test survey responses, as well as to integrate the qualitative data results from the video and interview observations.

4.4 Discussion of results

To summarize, self-paced items appear to be as valid and reliable as administrator-once-played items. Given the opportunity, many examinees did prefer to have the option to and/or take control of the audio play. Items were of similar difficulty and discrimination in 1x, SPS, and SPL conditions, with similar reliability values. Eighty percent of test takers said they preferred having play control on the main test. Although examinees tended to prefer having control, timing of the sets and examinees' computer use affected whether they could take that control. Timing limited how many times listeners could pause and play the input or check their responses. Those examinees less familiar with computer navigation may have been unable to show their listening ability as well as more proficient computer users.

4.4.1 Research question 1

RQ1 asked whether a self-paced (i.e., examinee control of playing, pausing, and audio position) play condition is as valid and reliable as one administrator-controlled play in a listening exam. To answer this, I consulted the quantitative data sources of the item responses, analyzed using both classical item analysis and Rasch measurement. I compared item difficulty and discrimination figures as well as item reliability. I also consulted examinees' survey responses as a measure of face validity.

My hypothesis for RQ1 was that, to bias for best, items in a self-paced listening assessment condition would permit listeners to show their ability better than on items presented just once, and examinees would prefer self-paced items and feel that they were a good measure of their listening proficiency. This hypothesis was confirmed; self-paced items performed as well as and often better than items with a once-played input condition. Main test items discriminated between test takers best (though not statistically differently) in the SPL (self-paced no time limit) condition. Items were easiest in the SPL condition, followed by SPS, then by 1x, as measured by logit value. SPL items were also easier than SPS for the stimulated-recall participants. Qualtrics click and timing data did not reveal significant correlations with proficiency level as measured by logit value, although candidates did spend slightly longer on SPL sets (3 minutes 2 seconds) as compared to 2:35 on 1x and 2:31 on SPS pages. Four of every five examinees on the main test said they preferred having control, potentially boosting the face validity of a self-paced listening test. Items had similar reliability, as measured by KR-21, across input play conditions.

4.4.2 Research question 2

RQ2 inquired whether, when, and why examinees take control over the listening audio play, and whether this varies by examinee listening proficiency level. To respond to this, I used the quantitative and qualitative data sources of the post-test survey responses, as well as the qualitative sources of the video observations and interview findings. Because simply offering a self-paced condition does not reflect whether examinees do use play controls, a closer look at examinees' play behaviors was necessary.

For basic or beginning English listeners, as measured by the main test, I had hypothesized that they would still have difficulty listening even in self-paced audio settings. However, I do not have substantial interview data to support hypotheses about beginners' play behaviors, and I cannot necessarily rely on their click or timing data as a proxy for pausing, unpausing, or replaying. On the main test, for all proficiency levels, not just beginners, there was little difference in performance on listening-once versus self-paced items.

For advanced listeners, I had imagined that they would not need to take advantage of the play controls; I thought listening once would be sufficient for them. Intermediate and advanced candidates varied in their preferences for and uses of control. Of the interviewees who took the supplementary items, there were two intermediate-to-advanced listeners who did not use the audio controls: Octavia and Estavan. Celeste, one of the most proficient interview participants, though, did use and like having the ability to start, stop, and repeat, so it is not necessarily the case that the more proficient the listener, the less she used control. Of the six stimulated-recall interviewees, one of the three lower-proficiency (intermediate to advanced) listeners took control of the audio play almost more often than the three higher-proficiency (advanced) listeners combined: Almas used play controls 11 times while Celeste, Gisela, and Yijun used control 12 times total among the three of them. However, these findings must all be interpreted with caution, as the interviewees' controls were observed separately from the main test. It is difficult to extend and connect the supplementary test findings to results of the main test without including the supplementary items in the statistical analysis and gathering more screen-capture and SR interview data. However, it did seem that repeated listens often appeared to help interviewees understand better what they had heard. Moreover, Estavan, an IEP level 4 interviewee, even commented during SR that hearing the audio again while evaluating his listening thoughts did help him comprehend part of it better, though he did not use the audio play controls during the test itself.

For intermediate and advanced listener interviewees, six of the eight did self-pace the audio input, even though I had not anticipated that at the highest levels. Even at the 2.19 logit values, the most advanced of the interviewees did occasionally take advantage of that control. Examinees did tend to take control, but strategic competence was very important to feeling successful on the listening test; test takers had to listen, read, and click on specific areas of the screen for successful responses. This means that the test may not be strictly a test of language proficiency but also a test of successful computer navigation, so if such an interface were implemented for high-stakes purposes, examinees would likely need a practice segment with unscored items to familiarize themselves with the format. Thus, for RQ2a, play control is part of a complex set of tasks examinees are contending with in a listening test, and its use is not necessarily connected strongly with proficiency level.

For RQ2b, when listeners do take control of the audio play, I thought they might do so when they wanted to confirm their hunches about a selected choice and/or when they wanted to hear a specific detail again. Interviewees described why they self-paced for more reasons than I expected, though I did not have an extensive list of possible reasons catalogued in advance. They took audio control in order to effectively use time, to confirm responses, to find key words, to follow the order of presented ideas, to understand the topic better, to pay attention to the audio, to locate key information, to not have to read items simultaneously, to verify that no information was missed, or just to hear the audio again because it seemed intriguing. There were no clear patterns by proficiency; however, the lower-intermediate-level interviewees appeared to take listening play controls not just to check or confirm their understanding but rather to identify where in an audio clip they believed they had heard something about tested information. No participants paused or unpaused without also replaying some of the input. Some interviewees also provided reasons they did *not* use play controls, which included being able to concentrate or focus better on the audio or reading texts, or helping avoid frequently checking answers. It is intriguing that examinees reported wanting to maintain focus or helping themselves learn; these were both offered as reasons for *and* for not using controls. Thus, there may be connections with examinees being given a choice and biasing for best in test development, as some candidates chose to take play controls while others did not, selecting what they felt best reflected their listening ability. In a standardized listening assessment, if the items are equally valid, the decision of whether or not to offer play controls must also be balanced with construct validity and practicality considerations. If test centers were to adopt items such as these, computer familiarity would have to be considered in the exam design as well as what type of a time limit to place on individual sets or the entire listening section.

The concluding remarks are presented in the last chapter, including implications for and limitations of the study as well as future directions.

5 CONCLUSIONS

By combining multiple-choice listening comprehension test item analyses with post-test survey responses, video capture data of test navigation behavior, and interview data, I determined whether candidates took or did not take control over input audio and how this impacted test performance. Examinees did often use play controls but not always, and they did perform better on multiple-choice test items in a self-paced condition with no time limit given. In this chapter, I describe implications of the findings, limitations, and suggestions for further research.

5.1 Implications of the study

As opposed to current L2 English listening tests in which a recording is played just once to the audience, the study focused on what occurs when audio play control is put in the hands of examinees. Rather than only including traditional comparisons of language learners' test totals or measurement tools such as analyses of variance, this study employed Rasch analysis to compare items by the listening variable of self-pacing with and without a time limit. It introduced an interactive test interface to adult language learners in a program of English for academic purposes, many of the items discriminated best in self-paced conditions, and participants generally felt the test showed their listening ability well. The practice exam helped many students feel that they could take control over the listening audio play; this may help students become more able to take control of their own language learning and engage in listening more actively. I employed verbal protocols because they allowed me to gather information, specifically about why candidates used play controls, that I would not have been able to gain only from responses to multiple-choice test items; this underscores the value of listening to stakeholders', especially examinees' opinions about a test.

A self-paced test format with no time limit may make items slightly easier for examinees (.48 logits easier in the present study) than once-played with exam center control. However, this slight difference in item facility is not concerning because it is not more than half a logit or more than three times the standard error overall. Items also generally discriminated better in self-paced conditions, though item results should be interpreted cautiously due to some measurement error.

To ensure that equal or comparable reliability and validity are maintained, examiners must clearly define the listening construct best for their needs, considering the impact of pausing and multiple plays on listening performance. Test developers must also consider the purpose for listening and whether that fits with the opportunity to take play control. If pausing and replaying are permitted, more comprehension items may be designed that require examinees to, for example, listen for very detailed information instead of simply listening for the gist or highly salient idea units. Thus, testing professionals must carefully consider the type(s) of listening they are intending to assess and design the input and expected responses accordingly.

The self-paced conditions, whether timed or not, may introduce some variability in test taker behavior and thus on test results, if used in high-stakes scenarios. If test developers would want to allow self-pacing, a condition with timing by set does not appear to cause items to become substantially easier than in a once-played condition. It would be the decision of examination boards to decide if this slight variability is appropriate to the construct for listening, as well as how variable listening audio play may impact test results and hence the inferences made from those results.

Strategic competence may play a large role in a test such as this, to the point of potentially giving some unfair advantages. For example, some examinees more keenly identified which audio position was necessary to move to in order to hear where in the stream of audio an item's key was located. This variability is a type of construct-irrelevant variance, a factor not necessarily involved in L2 listening proficiency that must be minimized when considering test design, especially for an exam that may be used for high-stakes purposes.

The results confirm the existing findings of language assessment work which has shown that more than one play makes a listening test easier (Berne, 1995; Brindley & Slatyer, 2002; Buck, 2001; Chang & Read, 2006; Jensen & Vinther, 2003; Ruhm et al., 2016). The twice-played testing conditions in these studies, though, were almost all administrator controlled. Twice-played audio, moreover, does not necessarily reveal that examinees did listen two times. Taking play control does not even necessarily mean that listening occurred. Because listening is internal, it is impossible to truly uncover whether examinees did listen more than once. However, these results indicate that many examinees seemed to have benefited from having the ability to access play controls. Being given the option to have control improved item discrimination and caused items on average to become slightly easier, without significant differences. Examinees tended to prefer the option of control and generally felt they did better on the test when able to have playing and pausing capabilities. Self-pacing, specifically the ability to control the audio play, as in Zhao (1997), resulted in stronger performance for examinees. Proficiency effects, as in Blau (1990), may have played a role, but the results of the present study are rather limited with regard to being able to make any generalizations about play control use. There did seem to be some interactions with play control and proficiency level, as in Roussel (2011); in this study, some interview participants at lower proficiency levels did utilize play control more often than more proficient listeners. However, examinees at similar listening ability levels used play controls differently, and no systematic patterns were found in Qualtrics click or timing data with respect to listening proficiency level. This indicates that examinees are controlling various aspects of their strategic competence in a listening examination situation.

With such a listening test as this in which self-pacing is permitted, there may be more instances of active listening and feeling more accomplishment in the listening process. Standardized tests of listening, due to practicality considerations, are often not able to include the perspective of the listener, except in face-to-face tasks involving listening combined with speaking or other modalities. They do not often consider the L2 learner as a contributor to an interaction, which is an important consideration in conceptualizing social factors (important in listening models from Rost, 2014; Weir, 2005; and language learning models from Gardner & Lambert, 1959, 1972; Pavlenko, 2002). When students listen, they should be encouraged to consciously improve their abilities (Rost, 2014); the self-paced test items in this study may help L2 listeners not only gauge their proficiency but also locate areas for listening focus and improvement. Tests can have a powerful influence on L2 learners (Bailey, 1999), so a self-paced format may also lead to positive washback, or impact on teaching, learning, and future assessment. Also borne out here was the warning from Chapelle and Douglas (2006), who commented on the potentially mediating or interfering variable of computer control on examinees' ability to show their language proficiency. Computer-based controls would need to be carefully evaluated if used as part of large-scale standardized listening examinations.

5.2 Limitations

Due to the data collection timeline and relatively low enrollment in the Intensive English Program, the low number of participants limited the inferences I am able to make about the item performance and hence listening proficiency. There are many potentially variable features of the test format relating to the presentation of audio, questions, and options; these features, aside from who holds audio play control, may impact examinee performance. The text of items was presented to candidates immediately, which is another potentially confounding variable in nearly any standardized listening test. This means that some listeners may have tried to engage in simultaneous written as well as spoken language processing, causing them to become frustrated, anxious, or distracted. Self-paced conditions may reduce the pressure of having to read items at the same time as listening, but reading may have interacted with or had effects on listening or concentration.

Listening studies that use multiple-choice items are sometimes criticized for lack of authenticity. Although responding to questions in that format is inauthentic compared to real-world activities, these items tap the *processes* involved in listening in the real world and the EAP classroom. Also, sets had been developed to have a certain level of lexicogrammatical features, but some audio recordings may have contained more propositions or idea units than others. Sets were aligned with TLU (target language use) domain types (e.g., listening to a lecture in the main test was directly connected to students' real listening needs); nevertheless, there were other domains (occupational or public/personal settings for monologues or dialogues) that were potentially irrelevant to students' listening needs. In any high-stakes standardized test, construct as well as content validity would need to be taken into consideration.

Test format familiarity and computer navigation skills also may have impacted exam results. For some users, navigating technology created construct-irrelevant variance, affecting their ability to show their listening proficiency. Technology manipulation should not necessarily be part of a listening construct. However, as we see more computer- or handheld-device-based language learning software approaching the forefront, such as online course management interfaces, technology may become more integrated with language acquisition and testing practices.

In order to examine connections with face validity, or the perception that the test was a good assessment of one's listening skills, I had included a post-test survey. In hindsight, I wish I had given more explanation of the survey in the main test class sessions. Not everyone may have interpreted "1 strongly *disagree*" to "6 strongly *agree*" in the same manner. Additionally, the only options for responses to preferring control were Yes and No; I could have included a scale to gauge the strength of people's preference. I am also not certain whether all examinees understood the item asking about preferring to have control; some who clicked Yes wrote open-ended comments that indicated that they actually did *not* prefer having control, or vice-versa. For the main test, I also wished I had asked candidates to note whether (and, if so, when) they utilized play controls, such as by having them type in a comment box or indicate the number of plays and/or number of times they paused or started/stopped. It could be that some examinees did not understand the items and/or interpreted them in different ways. For example, in a post-test survey response, one respondent commented that she changed the audio volume, and although this is a type of control, it is not what I was investigating.

With responses from just ten total interview participants on two different tests, the extendibility of claims that can be made about the qualitative data is limited. I did not calculate correct and incorrect answers or check examinee item responses prior to each interview, so I did not ask examinees in detail why they selected the choices they did. Had I done so, when examinees selected the incorrect choices, we could have discussed more about when and why their listening comprehension had broken down. In some of the interviews, I attempted to gather follow-up information about why a misunderstanding had occurred, if an interviewee did bring up a point of confusion, but I did not do this systematically for all participants or test items.

For the stimulated recall and interviews, I did not have many lower proficiency learners. I did not offer the interview in languages other than English; had I done so, candidates may have felt more comfortable to introspect. However, one pilot participant did tell me in the past, since the test is in English, it would be unusual to then have to discuss the test in her first language. Such a design with L1 interviews in languages other than the tested language would need to be carefully planned. I also wondered whether any language learners would have engaged in more frequent use of play controls, especially three, four, or more plays of the audio, without me sitting near them in the linguistics laboratory. Moreover, the test is still largely under administrator control, aside from the audio play itself, so there are many controlled and variable elements of the test. Roussel (2011) hypothesized that mouse motions and clicks were indicative of listeners' metacognitive processing such as planning or monitoring; however, I used the video-captured movements to investigate why people used play controls, not necessarily to tap into other types of metacognition about listening. Future studies with the opportunity to take listening play control could potentially focus more closely on metacognitive strategies and specific behaviors to listen successfully.

5.3 Future directions

With the study, I sought to uncover not only whether self-pacing (the opportunity to pause and replay) impacted item performance but also whether candidates took play control, and whether they felt they did better on the items with self-pacing. It is my hope that the results will make contributions to the existing work on L2 learning and research as well as listening assessment.

Further listening test analyses with many-facet Rasch measurement, videos, and/or interviews could tell us more about how else we can bias for best in listening assessment situations. For example, timing seems to be a crucial variable for listening. It is connected to possible expectations of automaticity for proficient listeners. In this study, I did not measure response speed because I wanted test takers to consider the items carefully and do their best. In a timed condition, the faster correct responses are made, the sooner the examinee can proceed in the test, if she is a proficient enough language user. Hence, test developers would have to consider whether to prioritize a subconstruct such as automaticity or speed of response as an indicator of "fluent" listening; compare, for example, timed writing having been used to operationalize writing fluency (Wolfe-Quintero, Inagaki, & Kim, 1998). Purpura (2004) and Read (2015) note that response speed is usually not a factor included in scoring language proficiency tests, as it may unduly boost test takers' anxiety.

There are also face validity connections with this project, in which I wish to underscore the value of gaining feedback about examinees' views. As a language tester, listening to test takers' and other stakeholders' opinions ensures fairness of test procedures and format, as well as learning more about the impact of testing on learning and instruction. I need to design exams to ensure that inferences made from test results are defensible, and that for the given purpose, test takers can best show their language proficiency level. A future direction in assessment research, specifically the testing of listening, might be opportunity to take play control of not just audio-based but also video-based listening and its impact. There, a test must be carefully designed because video introduces additional variables that test takers may be impacted by (Batty, 2015; Feak & Salehzadeh, 2001; Ginther, 2002; Ockey, 2007; Suvorov, 2009, 2013; Wagner, 2008, 2010a, 2010b, 2013). Play control, if relevant to the construct being assessed, would need to be examined in conjunction with features of the input or expected response potentially impacting test performance.

Undoubtedly, there are other phenomena occurring in this research setting besides just audio play. For example, some examinees may still be building their listening processing in English. They have different levels of experience and familiarity with test taking behaviors and strategies. Such a listening interface that allows them to take control may help them feel more confident and less anxious, especially in computer-assisted language learning situations in and out of the classroom, or in self-directed learning activities, if not enrolled in a formal course of study. Students and prospective test takers, if preparing for listening in a self-paced format, can be taught how to take advantage of their strategic competence in order to become more successful listeners. For example, with a computing-device-assisted listening interface, examinees need to know about how controls function and how the position of an audio slider bar can show the progression through a recording.

If high-stakes tests add self-pacing for listening, this would have implications on not only L2 learning but also classroom teaching. The idea of students being able to *listen again* is relevant to both live and recorded speech. If live speech is presented to students, and the context is appropriate, they could be encouraged to request clarification or repetition and identify where misunderstandings occurred. If speech is recorded, teachers would have to know how to use any playback devices and be able to show L2 listeners how to operate controls. This idea of making pausing and replay practices authentic when they were formerly not as authentic (Robin, 2007) is connected to motivating learners to take control and could transfer to an assessment situation.

As examinees' class materials and learning resources are increasingly available via computing devices, assessment needs may have to become more in line with classroom and real-world listening scenarios. If students are permitted to self-pace in classroom and real-world listening opportunities, it may follow that a testing situation could match those experiences, if the exam setting is practical, items are equally or even more valid and reliable, and the construct being assessed warrants such a format. The future of the field of listening assessment is bright with these rich opportunities for exam format innovation and new perspectives on the adult L2 listening construct.

REFERENCES

Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *The Modern Language Journal*, *78*(2), 155-168.

Alderson, J. C. (1990). Learner-centered testing through computers: Institutional issues in individual assessment. In J. H. A. L. De Jong & D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 20-27). Clevedon: Multilingual Matters.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: Freeman.

Anderson, J. R. (2009). *Cognitive psychology and its implications* (7th ed.). New York: Worth Publishers.

Arnold, J. (2000). Seeing through listening comprehension exam anxiety. *TESOL Quarterly, 34*(4), 777-786.

Aryadoust, V., & Goh, C. C. M. (2014). Predicting listening item difficulty with language complexity measures: A comparative data mining study. *CaMLA Working Papers*, *2014-02*, 1-39. Retrieved from http://www.cambridgemichigan.org/wp-content/uploads/2014/12/CWP-2014-02.pdf

Aryadoust, V., Goh, C. C., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*(4), 361-385.

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., & Palmer, A. S. (2010). *Language assessment in practice*. New York: Oxford University Press.

Baddeley, A. (1986). *Working memory*. Oxford: Clarendon Press.

Bailey, K. M. (1999). *Washback in language testing* (Research Report RM-99-4). Princeton, NJ: Educational Testing Service.

Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, *30*(1-2), 8-24.

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.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101-118.

Berne, J. E. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, *78*(2), 316-329.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. London/New York: Pearson.

Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly*, *24*(4), 746-753.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369-394.

Brown, G., & Yule, G. (1983). *Teaching the spoken language*. Cambridge: Cambridge University Press.

Brown, S. (2011). *Listening myths*. Ann Arbor, MI: University of Michigan Press.

Buck, G. (1990). *The testing of second language listening comprehension* (doctoral dissertation). University of Lancaster, England. Retrieved from the British Library EThOS (Electronic Theses Online Service).

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, *8*(1), 67-91.

Buck, G. (1995). How to become a good listening teacher. In D. J. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 113-131). San Diego, CA: Dominie.

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

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Cambridge Michigan Language Assessments (CaMLA, 2014). *MET Support Materials*. Retrieved from http://www.cambridgemichigan.org/resources/met/support-materials

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Candy, P. (1988). On the attainment of subject-matter autonomy. In D. Boud (Ed.), *Developing student autonomy in learning* (pp. 59-76). London: Kogan Page.

Cauldwell, R. (2014). *Phonology for listening: Teaching the stream of speech*. Birmingham, UK: Speech in Action.

Chafe, W. (1985). Linguistic differences produced by differences between speech and writing. In D. R. Olsen, N. Torrance, & A. Hilyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 105-123). Cambridge: Cambridge University Press.

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.

Chang, A. C.-S., & Read, J. (2008). Reducing listening test anxiety through various forms of listening support. *TESL-EJ*, *12*(1), 1-25.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chaudron, C., & Richards, J. C. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics*, *7*(2), 113-127.

Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, *26*(2), 345-374.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cross, J. (2014). Promoting autonomous listening to podcasts: A case study. *Language Teaching Research*, *18*(1), 8-32.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*, 133-142.

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.)., *Cognition and second language instruction* (pp. 125-151). Cambridge: Cambridge University Press.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

East, M., & King, C. (2012). L2 learners' engagement with high stakes listening tests: Does technology have a beneficial role to play? *CALICO Journal*, *29*(2), 208-223.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*.

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, *89*(2), 206-220.

Elliott, M. & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (*Studies in Language Testing*, Vol. 35; pp. 152-241). Cambridge: Cambridge University 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.

Ericsson, K., & Simon, H. (1996). *Protocol analysis: Verbal reports as data* (3rd ed.). Cambridge, MA: MIT Press.

Faerch, C., & Kasper, G. (1987). From product to process: Introspective methods in second language research. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 5-23). Clevedon/Philadelphia: Multilingual Matters.

Feak, C. B., & Salehzadeh, J. (2001). Challenges and issues in developing an EAP video listening placement assessment: A view from one program. *English for Specific Purposes*, *20*, 477-493.

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *System*, *32*, 363-377.

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

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (*Studies in Language Testing*, Vol. 35; pp. 77-151). Cambridge: Cambridge University Press.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension: An overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7-29). Cambridge: Cambridge University Press.

Flowerdew, J., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes*, *16*(1), 27-46.

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). Oxford: Wiley-Blackwell.

Fox, J. (2004). Biasing for the best in language testing and learning: An interview with Merrill Swain. *Language Assessment Quarterly*, *1*(4), 235-251.

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.

Gardner, R., & Lambert, W. (1959). Motivational variables in second-language acquisition. *Canadian Journal of Psychology*, *13*, 266-272.

Gardner, R., & Lambert, W. (1972). *Attitudes and motivation in second language learning*. Rowley, MA: Newbury House.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Cambridge Research Notes*, *32*, 2-5.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing*, *19*(2), 133-167.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*(1), 55-75.

Goh, C. C. M., & Aryadoust, S. V. (2010). Investigating the construct validity of the MELAB listening test through the Rasch analysis and correlated uniqueness modeling. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *8*, 31-68.

Goh, C. C., & Aryadoust, V. (2016). Learner listening: New insights and directions from empirical studies. *International Journal of Listening*, *30*(1-2), 1-7.

Goldman, S. R., Hogaboam, T. W., Bell, L. C., & Perfetti, C. A. (1980). Short-term retention of discourse during reading. *Journal of Educational Psychology*, *72*(5), 647.

Graham, S. (2006). Listening comprehension: The learners' perspective. *System*, *34*(2), 165-182.

Grgurović, M., & Hegelheimer, V. (2007). Help options and multimedia listening: Students' use of subtitles and the transcript. *Language Learning & Technology*, *11*(1), 45-66.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, *29*, 163-180.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston: Heinle & Heinle.

Hegelheimer, V., & Tower, D. (2004). Using CALL in the classroom: Analyzing student interactions in an authentic classroom. *System*, *32*(2), 185-205.

Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly*, *20*(3), 559-562.

Hulstijn, J. H. (2003). Connectionist models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning*, *16*(5), 413-425.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System, 34*(3), 317-340.

Jensen, E. D., & Vinther, T. (2003). Exact repetition as input enhancement in second language acquisition. *Language Learning*, *53*(3), 373-428.

King, P. (1994). Visual and verbal messages in the engineering lecture: note taking by postgraduate L2 students. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 219-238). Cambridge: Cambridge University Press.

Lawless, K. A., & Brown, S. W. (1997). Multimedia learning environments: Issues of learner control and navigation. *Instructional Science*, *25*(2), 117-131.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

Linacre, J. M. (2014). Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: Winsteps.com.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes*, *10*(2), 79-88.

McBride, K. (2011). The effect of rate of speech and distributed practice on the development of listening comprehension. *Computer Assisted Language Learning*, *24*(2), 131-154.

McNamara, D. S., & Shapiro, A. M. (2005). Multimedia and hypermedia solutions for promoting metacognitive engagement, coherence, and learning. *Journal of Educational Computing Research*, *33*(1), 1-29.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Miller, L. (2009). Engineering lectures in a second language: What factors facilitate students' listening comprehension? *Asian EFL Journal*, *11*(2), 8-30.

Murphy, J. M. (1996). Integrating listening and reading instruction in EAP programs. *English for Specific Purposes*, *15*(2), 105-120.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (TOEFL Research Reports, 51). Princeton, NJ: Educational Testing Service.

Norton Peirce, B., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics*, *14*(1), 25-42.

O'Bryan, A., & Hegelheimer, V. (2007). Integrating CALL into the classroom: The role of podcasting in an ESL listening strategies course. *ReCALL*, *19*(2), 162-180.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing*, *24*(4), 517-537.

Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening*, *30*(1-2), 84-98.

Papageorgiou, S., Stevens, R., & Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Language Assessment Quarterly*, *9*(4), 375-397.

Patall, E. A., Cooper, H., & Robinson, C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, *134*(2), 270-300.

Pavlenko, A. (2002). Poststructuralist approaches to the study of social factors in second language learning and use. In V. Cook (Ed.), *Portraits of the L2 user*, Vol. 1 (pp. 277-302). Clevedon: Multilingual Matters.

Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, *31*(6), 459-470.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Qualtrics Research Suite (2016) [computer software]. Provo, Utah: qualtrics.com.

R Core Team (2016). R: A Language and Environment for Statistical Computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press (original work published 1960 by the Danish Institute for Educational Research).

Read, J. (2002). The use of interactive input in EAP listening assessment. *Journal of English for Academic Purposes*, *1*(2), 105-119.

Read, J. (2015). *Assessing English proficiency for university study*. London: Palgrave Macmillan.

Read, T., & Barcena, E. (2016). Metacognition as scaffolding for the development of listening comprehension in a social MALL App / La metacognición como andamiaje para el desarrollo de la comprensión oral en una App de MALL social. *Revista Iberoamericana de Educación a Distancia* [*Iberoamerican Journal of Distance Education*], *19*(1), 103.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595-626.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(1), 31-65.

Robin, R. (2007). Commentary: Learner-based listening and technological authenticity. *Language Learning & Technology*, *11*(1), 109-115.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39-59.

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

Rost, M. (2014). Listening in a multilingual world: The challenges of second language (L2) listening. *International Journal of Listening*, *28*(3), 131-148.

Roussel, S. (2011). A computer assisted method to track listening strategies in second language learning. *ReCALL*, *23*(2), 98-116.

Ruhm, R., Leitner-Jones, C., Kulmhofer, A., Kiefer, T., Mlakar, H., & Itzlinger-Bruneforth, U. (2016). Playing the recording once or twice: Effects on listening test performances. *International Journal of Listening*, *30*(1-2), 67-83.

Sage, K., Bonacorsi, N., Izzo, S., & Quirk, A. (2015). Controlling the slides: Does clicking help adults learn? *Computers & Education*, *81*, 179-190.

Sasaki, M. (2014). Introspective methods. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1340-1357). Hoboken, NJ: John Wiley & Sons.

Sawaki, Y., & Nissan, S. (2009). *Criterion‐related validity of the TOEFL iBT listening section*. (Research Report RR-09-02). Princeton, NJ: Educational Testing Service.

Scheiter, K., & Gerjets, P. (2007). Learner control in hypermedia environments. *Educational Psychology Review*, *19*(3), 285-307.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing*, *8*(1), 23-40.

Smidt, E., & Hegelheimer, V. (2004). Effects of online academic lectures on ESL listening comprehension, incidental vocabulary acquisition, and strategy use. *Computer Assisted Language Learning*, *17*(5), 517-556.

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). Ames, IA: Iowa State University.

Swain, M. (1983). Large-scale communicative language testing: A case study. *Language Learning and Communication*, *2*, 133-147.

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.

Tsui, A. B. M. (1996). Reticence and anxiety in second language learning. In K. M. Bailey and D. Nunan (Eds.), *Voices from the language classroom* (pp. 145-167). Cambridge: Cambridge University Press.

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, *19*(4), 432-451.

Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning*, *53*(3), 463-496.

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, *26*(1), 70-89.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191-210.

Vandergrift, L. (2010). Researching listening. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 160-173). London and New York: Continuum.

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

Wagner, E. (2004). *A construction validation study of the extended listening sections of the ECPE and MELAB* [Research Report]. Spaan Fellow Working Papers in Second or

Foreign Language Assessment. Ann Arbor, MI: English Language Institute, University of Michigan.

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.

Wagner, E. (2010b). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493-513.

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.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263-287.

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

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

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (No. 17). Honolulu, HI: University of Hawaii Press.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In R. M. Smith (Ed.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 1-24). Maple Grove, MN: JAM [Journal of Applied Measurement] Press.

Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*(1), 21-44.

Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics*, *18*(1), 49-68.

APPENDICES

Appendix A: Sample item sets

F = female speaker; M = male speaker; introductory scene-setter line spoken by a male narrator.

Correct answers are indicated with bolded and underlined text.

Sample dialogue

Listen to a telephone conversation.
F: Hello, I'm wondering if it's too late to arrange for a next-day delivery to Los Angeles?
M: It's too late to schedule a pickup. Our drivers are already out making their final rounds. But if you can bring your package to one of our offices, we can still guarantee next-day delivery.
F: Okay, great, um, where's your office?
M: We have several in the area. Where are you?
F: I'm on Fifty-third Street and Fifth Avenue.
M: Let's see. Fifty-third and Fifth. Uh, one moment. Our closest office is on Fifty-ninth Street and Seventh Avenue. And uh, they close at, they close at six-thirty this evening.
F: Oh, really? Great! I have plenty of time then. I was afraid I wasn't gonna be able to get this out today. Um, okay, well, I guess that's all. Thanks for your help.
M: No problem. Have a good day.

Q31 What does the woman want?

- ○ a. to pick up her package tomorrow
- ○ b. to have the man pick up her package tomorrow
- ○ **c. to have a package delivered the next day**
- ○ d. to deliver a package to the man tonight

Q32 What does the man tell the woman about the delivery?

- ○ a. He will contact a driver to deliver the package tonight.
- ○ b. She will get a refund if the package is delivered late.
- ○ **c. She must bring the package to an office for next-day delivery.**
- ○ d. Her package will be delivered before noon.

Q33
Test takers hear but do not read:
Listen to a part of the conversation again. Then answer the question.
"Oh, really? Great! I have plenty of time then."

Test takers hear and read:
Why does the woman say:

Test takers hear but do not read:
"Oh, really?"

- ○ a. She wants the man to continue speaking.
- ○ b. She agrees with the man.
- ○ c. She does not understand what the man said.
- ○ **d. She is surprised by what the man told her.**

sample monologue

Listen to a researcher giving a presentation to his colleagues. He is talking about a research study.

I wanna share with you a study we did, on how teens use the internet to find out about music. We interviewed over eighteen hundred teens to ask them about their internet use, especially how they use the internet with regard to music. We spoke to all these young people face-to-face. What we learned is that teenage girls are more likely than boys to use the internet to research a musician or a band. Girls are also more likely to go online to listen to music and watch music videos. About the only music-related activity that boys seem to do more of is downloading music to copy to CDs. Another interesting thing we found is that teenage girls who spend a significant amount of time online, about half of them spend at least a hundred dollars a year on buying music, buying music from online retailers. These girls actually prefer to get their music this way, rather than going to the store. In our study, we also identified those teens who are the so-called music influencers, the ones who other people turn to for advice or opinions about music, the ones who seem to know what's new or cool in music. Music influencers also tend to be teenage girls. And these girls spend nearly a third more money on music than average teens, which makes sense. They're influencers because they listen to more music. They tend to have a wide range of musical tastes. And because they're spending so much time listening to music, they wind up buying more music.

Q44 What was the research study about?

- ○ a. how often teenagers use a computer
- ○ b. where teenagers buy their CDs
- ○ c. what kind of music teenagers like
- ○ **d. how teenagers use the Internet**

Q45 How was the information for the study collected?

- ○ a. e-mail questionnaires
- ○ b. telephone surveys
- ○ **c. live interviews**
- ○ d. website forms

Q46 What does the speaker say about teenage girls who spend a lot of time online?

- ○ **a. They often use the Internet to buy their music.**
- ○ b. They spend lots of money on music magazines.
- ○ c. They like buying music at the store.
- ○ d. They attend a lot of live concerts.

Q47 What kind of people tend to be music influencers?

❍ a. teens who watch a lot of music videos
❍ b. teenage boys who spend a lot of money on music
❍ c. teenage boys who download music from the Internet
❍ **d. teenage girls who often listen to music**

Appendix B: Stimulated recall research protocol

Research Protocol

Listening test and stimulated recall

Date: ____________________Time: ______________________

Special notes: ____

____

____

I. Consent and introduction

Researcher introduces herself and reads aloud the consent form to the participant.

Researcher asks whether participant agrees to be video and audio recorded. Researcher starts recording devices only if participant gives consent.

II. Test administration

Researcher gives participant pen, scrap paper, and on-screen and verbal instructions for the listening test.

"Next, a listening test will be given. There are directions on the screen for each section, which I'll read with you to check whether you have any questions or concerns.

"During the test, your screen will be video-recorded. Please tell me at any time if you have concerns or questions, or want to stop the study."

Researcher starts screen-capture software. Examinee takes listening test. Researcher stops screen-capture recording.

Examinee takes post-test survey while researcher saves video file.

III. Stimulated recall interview

Researcher sets up video for participant to be able to control.

"For this part, I'd like you to watch the video of your mouse movements and clicks while taking the test. You can pause the video at any time if you want to tell me something. I would like you to think about how you felt while you were listening at that time, and pause the video whenever you want to tell me something. You can discuss information that you did OR did not understand during the test. Remember to stop or pause the video whenever you have something you want to say. Please tell me at any time if you have concerns or questions, or want to stop the study."

Participant plays video and talks while researcher listens and takes notes.

(Researcher asks follow-up questions if needed, based on what participants have said.)

IV. Wrap-up

"Do you have any questions or comments about the test?"

"Do you have any other questions for me?"

"Thank you very much for your time. This was very helpful for me to be able to hear your opinions."

Researcher stops recording devices and collects and locks materials away securely.

Appendix C: Classical item facility and point biserial values

Appendix C.1: Classical facility and discrimination totals

The information is sorted by item ID, with percentage correct and item discrimination (point biserial) values shown in the second and third columns. The anchor items are listed first, followed by variable items divided into sets by horizontal lines. All item conditions (1x, SPS, SPL) were calculated together for these figures:

Item	*Facility*	*Point Biserial*
A01	23.76	34.89
A02	45.54	41.46
A03	91.09	20.46
A04	86.14	43.57
A05	85.15	33.60
A06	25.74	23.69
A07	71.29	53.42
A08	42.57	41.95
A09	79.21	44.21
A10	82.18	47.99
Q23	70.30	43.84
Q24	71.29	61.69
Q25	89.11	18.64
Q26	61.39	36.16
Q27	73.27	28.73
Q28	45.54	53.89
Q29	56.44	62.44
Q30	47.52	49.65
Q31	65.35	34.66
Q32	81.19	48.61
Q33	85.15	33.24
Q34	75.25	52.52
Q35	67.33	47.31
Q36	85.15	48.11
Q37	76.24	36.32
Q38	72.28	18.87
Q39	31.68	39.88
Q40	77.23	34.12
Q41	63.37	34.99
Q42	35.64	36.95
Q43	42.57	45.86
Q44	70.30	35.37
Q45	58.42	58.42
Q46	85.15	26.35
Q47	65.35	29.24
Q48	51.49	57.96
Q49	90.10	22.93
Q50	68.32	53.26

Item	*Facility*	*Point Biserial*
Q51	79.21	36.90
Q57	45.54	29.81
Q58	15.84	35.12
Q59	22.77	40.92
Q60	32.67	46.18

Appendix C.2: Classical facility and discrimination by condition

The facility and discrimination values separated *by condition* are shown here. Columns 2 and 3 present classical item analysis for the 1x condition, columns 4 and 5 for SPS, and columns 6 and 7 for SPL. The point biserial, the right-hand value in each column pair, is bolded if it is the highest for that condition:

	1x		*SPS*		*SPL*	
Item	*Facility*	*Point Biserial*	*Facility*	*Point Biserial*	*Facility*	*Point Biserial*
Q23	0.66	0.37	0.76	0.478	0.71	**0.480**
Q24	0.66	**0.72**	0.79	0.54	0.69	0.63
Q25	0.88	0.14	0.94	**0.43**	0.86	0.11
Q26	0.44	0.19	0.61	**0.69**	0.77	0.33
Q27	0.57	0.31	0.81	0.28	0.82	**0.32**
Q28	0.43	**0.58**	0.38	0.53	0.58	0.44
Q29	0.51	0.63	0.63	0.55	0.58	**0.70**
Q30	0.34	**0.57**	0.41	0.44	0.70	0.46
Q31	0.41	**0.46**	0.79	0.15	0.77	0.33
Q32	0.75	**0.55**	0.82	0.54	0.86	0.47
Q33	0.94	0.41	0.73	0.32	0.91	**0.47**
Q34	0.71	**0.55**	0.72	0.50	0.82	0.51
Q35	0.69	0.39	0.63	**0.51**	0.70	0.50
Q36	0.80	0.47	0.84	**0.54**	0.91	0.35
Q37	0.67	**0.58**	0.89	0.29	0.72	0.34
Q38	0.70	0.12	0.77	0.19	0.69	**0.30**
Q39	0.33	0.30	0.23	0.36	0.41	**0.48**
Q40	0.85	**0.52**	0.63	0.17	0.84	0.38
Q41	0.58	0.29	0.66	**0.54**	0.69	0.23

	1x		*SPS*		*SPL*	
Item	*Facility*	*Point Biserial*	*Facility*	*Point Biserial*	*Facility*	*Point Biserial*
Q42	0.33	0.37	0.46	**0.46**	0.28	0.38
Q43	0.42	0.26	0.46	0.57	0.41	**0.61**
Q44	0.53	0.27	0.85	0.38	0.74	**0.44**
Q45	0.44	0.45	0.64	**0.70**	0.69	0.63
Q46	0.84	-0.01	0.91	0.28	0.83	**0.46**
Q47	0.75	0.13	0.55	0.25	0.69	**0.55**
Q48	0.54	**0.69**	0.44	0.50	0.55	0.60
Q49	0.94	0.08	0.91	**0.35**	0.88	0.23
Q50	0.71	0.42	0.56	**0.63**	0.79	0.39
Q51	0.71	**0.55**	0.75	0.37	0.91	0.04
Q57	0.33	0.08	0.54	0.23	0.50	**0.63**
Q58	0.18	0.29	0.11	0.33	0.19	**0.45**
Q59	0.36	0.41	0.14	0.38	0.19	**0.41**
Q60	0.45	0.53	0.17	0.15	0.38	**0.72**
AVG	**0.59**		**0.62**		**0.67**	

Appendix D: FACETS output for Rasch analysis

Appendix D.1: Examinee measurement report

```
Examinees Measurement Report  (arranged by mN).

+-----------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |                     |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Num Examinees       |
|--------------------------------+--------------+----------------------+------+-------------+---------------------|
|    41      43       .95    .98 |   3.55   .75 | 1.14   .4  7.39  2.5 |  .74 |  -.03   .26 |  61 61              |
|    40      43       .93    .96 |   2.97   .63 | 1.02   .1  1.08   .4 |  .95 |   .22   .29 |  17 17              |
|    39      43       .91    .95 |   2.75   .57 |  .96   .0  2.09  1.2 |  .94 |   .31   .37 |  83 83              |
|    39      43       .91    .95 |   2.72   .56 |  .88  -.2   .39  -.6 | 1.15 |   .47   .35 |  36 36              |
|    38      43       .88    .94 |   2.45   .52 |  .91  -.1   .43  -.7 | 1.15 |   .50   .40 |  81 81              |
|    38      43       .88    .94 |   2.45   .52 |  .95   .0   .45  -.6 | 1.12 |   .48   .40 |  90 90              |
|    38      43       .88    .93 |   2.42   .52 | 1.04   .2  1.36   .6 |  .90 |   .29   .38 |  37 37              |
|    38      43       .88    .93 |   2.33   .51 | 1.21   .7  2.08  1.4 |  .76 |   .12   .35 |   4 4               |
|    37      43       .86    .92 |   2.19   .49 | 1.12   .4   .78  -.1 |  .96 |   .39   .43 |  70 70              |
|    37      43       .86    .92 |   2.19   .49 |  .90  -.2   .61  -.4 | 1.12 |   .50   .43 |  75 75              |
|    37      43       .86    .92 |   2.19   .49 |  .77  -.7   .48  -.7 | 1.23 |   .57   .43 |  93 93              |
|    37      43       .86    .92 |   2.17   .49 |  .81  -.6   .42  -.9 | 1.23 |   .56   .41 |  38 38              |
|    37      43       .86    .92 |   2.17   .49 |  .75  -.8   .40 -1.0 | 1.27 |   .58   .41 |  67 67              |
|    37      43       .86    .91 |   2.09   .48 |  .74  -.8   .45 -1.0 | 1.26 |   .58   .38 |   2 2               |
|    36      43       .84    .90 |   1.96   .47 |  .75  -.9   .51  -.8 | 1.26 |   .60   .45 |  74 74              |
|    36      43       .84    .89 |   1.87   .45 | 1.13   .5  1.12   .3 |  .88 |   .29   .40 |  13 13              |
|    35      43       .81    .88 |   1.76   .45 | 1.13   .5  1.53  1.0 |  .81 |   .33   .47 |  97 97              |
|    35      43       .81    .88 |   1.74   .44 |  .87  -.4   .86  -.1 | 1.12 |   .51   .45 |  68 68              |
|    34      43       .79    .86 |   1.56   .43 |  .91  -.2   .72  -.5 | 1.12 |   .54   .48 |  79 79              |
|    34      43       .79    .86 |   1.56   .43 |  .85  -.6   .73  -.5 | 1.17 |   .55   .46 |  45 45              |
|    34      43       .79    .86 |   1.56   .43 | 1.02   .1   .98   .1 |  .97 |   .44   .46 |  50 50              |
|    34      43       .79    .85 |   1.50   .42 | 1.13   .6  1.28   .7 |  .82 |   .31   .43 |   3 3               |
|    34      43       .79    .85 |   1.50   .42 |  .93  -.2   .85  -.2 | 1.09 |   .47   .43 |   5 5               |
|    34      43       .79    .85 |   1.50   .42 |  .90  -.3   .84  -.2 | 1.11 |   .48   .43 |   7 7               |
|    34      43       .79    .85 |   1.50   .42 | 1.18   .8  1.50  1.1 |  .75 |   .26   .43 |  10 10              |
|    34      43       .79    .85 |   1.50   .42 | 1.25  1.0  1.34   .8 |  .69 |   .22   .43 |  11 11              |
|    33      43       .77    .83 |   1.38   .41 |  .65 -1.7   .43 -1.6 | 1.46 |   .71   .48 |  34 34              |
|    33      43       .77    .83 |   1.38   .41 |  .78  -.9   .83  -.3 | 1.23 |   .59   .48 |  46 46              |
|    33      43       .77    .82 |   1.33   .40 | 1.04   .2   .82  -.3 | 1.01 |   .44   .44 |   6 6               |
|    33      43       .77    .82 |   1.33   .40 |  .79  -.9   .58 -1.1 | 1.31 |   .61   .44 |  12 12              |
|    32      43       .74    .81 |   1.22   .40 | 1.35  1.5  1.38  1.0 |  .58 |   .28   .50 |  78 78              |
|    32      43       .74    .80 |   1.17   .39 | 1.00    .0   .90   -.1 | 1.02 |    .45   .45 |  24 24                |
|    31      43       .72    .78 |   1.06   .40 | 1.18    .8  1.12    .4 |  .79 |    .39   .51 |  69 69                |
|    31      43       .72    .78 |   1.06   .40 |  .73  -1.3   .58  -1.2 | 1.37 |    .68   .51 |  73 73                |
|    31      43       .72    .78 |   1.06   .40 | 1.34   1.5  1.59   1.5 |  .53 |    .27   .51 |  88 88                |
|    31      43       .72    .78 |   1.02   .38 | 1.23   1.1  1.06    .2 |  .73 |    .32   .46 |   1 1                 |
|    30      43       .70    .76 |    .91   .38 |  .58  -2.4   .56  -1.5 | 1.57 |    .75   .50 |  53 53                |
|    30      43       .70    .76 |    .91   .39 |  .77  -1.1   .92   -.1 | 1.25 |    .62   .51 |  71 71                |
|    30      43       .70    .76 |    .91   .39 | 1.01    .1  1.44   1.3 |  .90 |    .46   .51 |  72 72                |
|    30      43       .70    .76 |    .91   .39 |  .79  -1.0   .70   -.9 | 1.30 |    .64   .51 |  86 86                |
|    30      43       .70    .76 |    .91   .39 | 1.16    .8   .95    .0 |  .86 |    .44   .51 |  94 94                |
|    29      43       .67    .73 |    .77   .38 | 1.36   1.7  1.91   2.5 |  .33 |    .21   .51 |  39 39                |
|    29      43       .67    .73 |    .77   .38 | 1.15    .8  1.05    .2 |  .82 |    .42   .51 |  40 40                |
|    29      43       .67    .73 |    .77   .38 | 1.15    .8  1.24    .8 |  .75 |    .39   .51 |  49 49                |
|    29      43       .67    .73 |    .77   .38 | 1.02    .1  1.21    .7 |  .92 |    .47   .51 |  54 54                |
|    29      43       .67    .73 |    .76   .38 |  .62  -2.2   .50  -1.9 | 1.58 |    .76   .52 |  87 87                |
|    28      43       .65    .70 |    .63   .37 |  .75  -1.4   .92   -.2 | 1.33 |    .64   .51 |  35 35                |
|    28      43       .65    .70 |    .63   .37 |  .92   -.3   .77   -.8 | 1.18 |    .58   .51 |  42 42                |
|    28      43       .65    .70 |    .63   .37 | 1.04    .2  1.00    .0 |  .94 |    .48   .51 |  48 48                |
|    28      43       .65    .70 |    .61   .36 |  .76  -1.5   .62  -1.6 | 1.49 |    .67   .48 |  20 20                |
|    27      43       .63    .67 |    .49   .37 |  .78  -1.2   .67  -1.3 | 1.40 |    .66   .52 |  43 43                |
|    27      43       .63    .67 |    .48   .37 |  .85   -.8  1.06    .3 | 1.19 |    .58   .52 |  84 84                |
|    27      43       .63    .67 |    .48   .37 | 1.01    .1  1.02    .1 |  .97 |    .50   .52 |  89 89                |
|    27      43       .63    .67 |    .48   .37 |  .86   -.7   .80   -.7 | 1.25 |    .61   .52 | 101 101               |
|    26      43       .60    .64 |    .36   .36 |  .98    .0  1.02    .1 | 1.01 |    .51   .52 |  63 63                |
|    26      43       .60    .64 |    .35   .35 | 1.22   1.3  1.29   1.2 |  .53 |    .30   .48 |  25 25                |
|    25      43       .58    .61 |    .23   .36 |  .88   -.6   .78   -.9 | 1.26 |    .60   .52 |  60 60                |
|    25      43       .58    .61 |    .23   .35 |  .83  -1.1   .72  -1.3 | 1.40 |    .62   .48 |   8 8                 |
|    25      43       .58    .61 |    .23   .35 | 1.05    .3   .98    .0 |  .94 |    .45   .48 |  16 16                |
|    25      43       .58    .61 |    .22   .36 | 1.34   1.8  1.38   1.4 |  .37 |    .29   .52 |  80 80                |
|    24      43       .56    .58 |    .11   .35 |  .67  -2.4   .60  -2.2 | 1.74 |    .73   .48 |  18 18                |
|    24      43       .56    .58 |    .11   .35 |  .73  -1.9   .65  -1.9 | 1.62 |    .69   .48 |  22 22                |
|    24      43       .56    .58 |    .10   .36 |  .77  -1.4   .66  -1.6 | 1.48 |    .67   .52 |  41 41                |
|    24      43       .56    .58 |    .09   .36 |  .83  -1.0   .78   -.9 | 1.34 |    .62   .52 |  85 85                |
|    24      43       .56    .58 |    .09   .36 | 1.21   1.2  1.20    .8 |  .60 |    .37   .52 |  96 96                |
|    23      43       .53    .55 |   -.01   .35 | 1.04    .3   .98    .0 |  .95 |    .46   .48 |  33 33                |
|    23      43       .53    .55 |   -.02   .36 |  .85   -.9   .74  -1.2 | 1.35 |    .62   .51 |  56 56                |
|    23      43       .53    .55 |   -.02   .36 |  .97   -.1   .89   -.4 | 1.10 |    .54   .51 |  57 57                |
|    23      43       .53    .55 |   -.03   .35 | 1.04    .3  1.12    .5 |  .89 |    .47   .51 |  77 77                |
|    23      43       .53    .55 |   -.03   .35 | 1.20   1.2  1.11    .5 |  .65 |    .39   .51 |  91 91                |
|    23      43       .53    .55 |   -.03   .35 |  .95   -.2   .87   -.5 | 1.13 |    .55   .51 |  95 95                |
|    22      43       .51    .52 |   -.15   .35 | 1.34   2.0  1.52   2.1 |  .23 |    .25   .51 |  65 65                |
|    22      43       .51    .52 |   -.16   .35 | 1.39   2.3  1.58   2.2 |  .08 |    .22   .51 | 100 100               |
|    21      43       .49    .49 |   -.25   .35 | 1.06    .4  1.16    .8 |  .80 |    .41   .48 |  19 19                |
|    21      43       .49    .49 |   -.25   .35 |  .75  -1.8   .66  -1.8 | 1.62 |    .67   .48 |  21 21                |
|    21      43       .49    .49 |   -.27   .35 |  .96   -.2   .96   -.1 | 1.08 |    .52   .51 |  58 58                |
|    21      43       .49    .49 |   -.27   .35 |  .93   -.4   .86   -.5 | 1.18 |    .55   .51 |  59 59                |
|    20      43       .47    .46 |   -.37   .35 |  .89   -.7   .86   -.6 | 1.27 |    .55   .48 |  14 14                |
|    20      43       .47    .46 |   -.39   .35 | 1.10    .6  1.07    .3 |  .80 |    .43   .50 |  44 44                |
|    19      43       .44    .43 |   -.49   .35 | 1.37   2.3  1.75   2.9 | -.06 |    .13   .47 |   9 9                 |
|    19      43       .44    .43 |   -.49   .35 | 1.15   1.0  1.23   1.0 |  .62 |    .34   .47 |  23 23                |
|    19      43       .44    .43 |   -.52   .35 |  .92   -.4   .82   -.6 | 1.21 |    .55   .50 |  47 47                |
|    19      43       .44    .43 |   -.52   .35 |  .98   -.1   .89   -.3 | 1.08 |    .51   .50 |  52 52                |
|    19      43       .44    .42 |   -.53   .35 | 1.05    .4  1.23    .9 |  .79 |    .42   .49 |  98 98                |
|    18      43       .42    .40 |   -.61   .35 |  .88   -.8   .79   -.9 | 1.32 |    .56   .47 |  15 15                |
|    18      43       .42    .39 |   -.65   .35 | 1.21   1.4  1.88   2.7 |  .33 |    .28   .49 |  76 76                |
|    18      43       .42    .39 |   -.65   .35 |  .73  -2.0   .60  -1.6 | 1.66 |    .67   .49 |  92 92                |
|    17      43       .40    .37 |   -.73   .35 |  .89   -.7   .76  -1.0 | 1.32 |    .56   .46 |  26 26                |
|    17      43       .40    .37 |   -.73   .35 |  .90   -.6   .81   -.7 | 1.26 |    .54   .46 |  31 31                |
|    17      43       .40    .37 |   -.77   .36 | 1.19   1.3  1.26    .9 |  .54 |    .33   .48 |  51 51                |
|    16      43       .37    .35 |   -.86   .35 | 1.34   2.1  1.59   2.0 |  .16 |    .16   .45 |  27 27                |
|    16      43       .37    .35 |   -.86   .35 | 1.08    .6  1.04    .2 |  .83 |    .39   .45 |  29 29                |
|    16      43       .37    .34 |   -.89   .36 | 1.49   2.9  2.18   3.1 | -.30 |    .05   .47 |  55 55                |
|    16      43       .37    .34 |   -.90   .36 |  .80  -1.4   .82   -.5 | 1.40 |    .58   .47 |  99 99                |
|    15      43       .35    .31 |  -1.03   .36 | 1.24   1.5  1.32   1.0 |  .48 |    .29   .46 |  82 82                |
|    14      43       .33    .28 |  -1.15   .37 | 1.25   1.5  1.73   1.9 |  .36 |    .22   .45 |  62 62                |
|    14      43       .33    .28 |  -1.15   .37 |  .90   -.6   .73   -.7 | 1.25 |    .53   .45 |  66 66                |
|    12      43       .28    .24 |  -1.38   .37 |  .89   -.6   .79   -.5 | 1.21 |    .49   .41 |  28 28                |
|    11      43       .26    .21 |  -1.52   .38 |  .90   -.5   .76   -.5 | 1.19 |    .48   .40 |  32 32                |
|     9      43       .21    .16 |  -1.89   .41 |  .88   -.5   .65   -.6 | 1.20 |    .47   .38 |  64 64                |
|--------------------------------+--------------+------------------------+------+-------------+-----------------------|
|    27.0    43.0     .63    .65 |    .60   .40 |  .99    .0  1.05    .0 |      |    .46      | Mean (Count: 100)     |
|     7.6      .0     .18    .21 |   1.13   .07 |  .20   1.1   .75   1.2 |      |    .16      | S.D. (Population)     |
|     7.7      .0     .18    .21 |   1.13   .07 |  .20   1.1   .75   1.2 |      |    .16      | S.D. (Sample)         |
+----------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .40  Adj (True) S.D. 1.05  Separation 2.61  Strata 3.81  Reliability .87
Model, Sample: RMSE .40  Adj (True) S.D. 1.06  Separation 2.62  Strata 3.83  Reliability .87
Model, Fixed (all same) chi-square:  677.0  d.f.: 99  significance (probability): .00
Model,  Random (normal) chi-square:  86.5  d.f.: 98  significance (probability): .79
------------------------------------------------------------------------------------------------------------------------
```

Appendix D.2: Condition measurement report

```
Condition Measurement Report  (arranged by mN).

+----------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit     |Estim.| Correlation |                      |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | N Condition          |
|--------------------------------+--------------+----------------------+------+-------------+----------------------|
|  1282    2100       .61    .65 A    .00   .06 | 1.03   .9  1.07  1.1 |  .96 |   .56   .57 | 1 onceplayed         |
|   679    1100       .62    .69 |   -.18   .08 |  .97  -.9   .93  -.8 | 1.06 |   .57   .56 | 2 SPS3minlimit       |
|   738    1100       .67    .75 |   -.48   .08 |  .96 -1.0  1.15  1.6 | 1.06 |   .56   .54 | 3 SPLnotimelimit     |
|--------------------------------+--------------+----------------------+------+-------------+----------------------|
|   899.7  1433.3     .63    .69 |   -.22   .07 |  .98  -.3  1.05   .6 |      |   .56       | Mean (Count: 3)      |
|   271.4   471.4     .03    .04 |    .20   .01 |  .03   .9   .09  1.1 |      |   .01       | S.D. (Population)    |
|   332.4   577.4     .03    .05 |    .24   .01 |  .04  1.1   .11  1.3 |      |   .01       | S.D. (Sample)        |
+-------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .07  Adj (True) S.D. .18  Separation 2.61  Strata 3.81  Reliability .87
Model, Sample: RMSE .07  Adj (True) S.D. .23  Separation 3.27  Strata 4.70  Reliability .91
Model, Fixed (all same) chi-square:  25.0  d.f.: 2  significance (probability): .00
Model,  Random (normal) chi-square:  1.8  d.f.: 1  significance (probability): .17
---------------------------------------------------------------------------------------------------------------------
```

Appendix D.3: Items measurement report

```
Items Measurement Report  (arranged by mN).

+-------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit     |Estim.| Correlation |                      |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Nu Items             |
|--------------------------------+--------------+----------------------+------+-------------+----------------------|
|    16     100       .16    .12 |   2.86   .30 |  .99   .0  1.02   .1 | 1.01 |   .38   .38 | 41 Q58               |
|    23     100       .23    .18 |   2.31   .26 | 1.00   .0  1.04   .2 |  .99 |   .41   .42 | 42 Q59               |
|    24     100       .24    .23 |   2.03   .26 | 1.05   .4  1.19   .8 |  .91 |   .37   .42 |  1 A01               |
|    26     100       .26    .25 |   1.90   .25 | 1.20  1.5  1.37  1.5 |  .67 |   .25   .43 |  6 A06               |
|    32     100       .32    .28 |   1.75   .24 | 1.01   .1  1.07   .4 |  .96 |   .43   .45 | 35 Q39               |
|    33     100       .33    .30 |   1.69   .24 |  .99   .0  1.01   .1 | 1.01 |   .45   .45 | 43 Q60               |
|    36     100       .36    .33 |   1.52   .23 | 1.07   .6  1.18  1.1 |  .82 |   .38   .46 | 38 Q42               |
|    43     100       .43    .42 |   1.15   .23 | 1.00   .0  1.01   .0 |  .99 |   .46   .47 | 39 Q43               |
|    46     100       .46    .45 |   1.01   .23 |  .91  -.9   .84 -1.2 | 1.24 |   .55   .48 | 23 Q28               |
|    46     100       .46    .46 |   1.00   .23 | 1.18  1.8  1.21  1.5 |  .56 |   .32   .47 | 40 Q57               |
|    42     100       .42    .46 |    .99   .23 | 1.04   .4  1.04   .3 |  .91 |   .43   .47 |  8 A08               |
|    48     100       .48    .48 |    .90   .23 |  .93  -.7   .87  -.9 | 1.20 |   .54   .48 | 25 Q30               |
|    46     100       .46    .51 |    .79   .23 | 1.06   .6  1.03   .2 |  .87 |   .42   .47 |  2 A02               |
|    51     100       .51    .52 |    .75   .23 |  .85 -1.6   .81 -1.4 | 1.37 |   .59   .47 | 29 Q48               |
|    57     100       .57    .59 |    .44   .23 |  .83 -1.9   .74 -1.9 | 1.44 |   .60   .47 | 24 Q29               |
|    59     100       .59    .62 |    .35   .23 |  .82 -2.0   .74 -1.8 | 1.44 |   .60   .46 | 19 Q45               |
|    61     100       .61    .64 |    .24   .23 | 1.04   .4  1.11   .7 |  .88 |   .41   .45 | 14 Q26               |
|    64     100       .64    .68 |    .08   .23 | 1.09   .9  1.04   .2 |  .82 |   .37   .43 | 37 Q41               |
|    66     100       .66    .70 |   -.02   .24 | 1.07   .7  1.05   .3 |  .86 |   .38   .44 | 15 Q31               |
|    66     100       .66    .70 |   -.02   .24 | 1.20  1.9  1.16   .8 |  .61 |   .29   .44 | 21 Q47               |
|    67     100       .67    .71 |   -.10   .24 |  .97  -.2   .88  -.5 | 1.08 |   .47   .44 | 27 Q35               |
|    69     100       .69    .74 |   -.21   .24 |  .90  -.9   .79 -1.0 | 1.21 |   .51   .43 | 31 Q50               |
|    71     100       .71    .76 |   -.31   .24 | 1.01   .0   .85  -.6 | 1.04 |   .43   .42 | 11 Q23               |
|    71     100       .71    .76 |   -.31   .24 |  .79 -2.0   .62 -1.9 | 1.41 |   .59   .42 | 12 Q24      overfit  |
|    71     100       .71    .76 |   -.31   .24 | 1.05   .4  1.30  1.3 |  .84 |   .34   .42 | 18 Q44               |
|    72     100       .72    .77 |   -.37   .24 | 1.23  1.9  1.40  1.6 |  .56 |   .20   .40 | 34 Q38      misfit   |
|    73     100       .73    .78 |   -.45   .25 | 1.12  1.0  1.08   .3 |  .83 |   .33   .41 | 22 Q27               |
|    71     100       .71    .79 |   -.53   .24 |  .86 -1.3   .75 -1.2 | 1.28 |   .53   .41 |  7 A07               |
|    75     100       .75    .80 |   -.57   .25 |  .87 -1.1   .71 -1.2 | 1.24 |   .52   .40 | 26 Q34               |
|    76     100       .76    .81 |   -.62   .25 | 1.02   .2   .91  -.3 | 1.00 |   .38   .38 | 33 Q37               |
|    77     100       .77    .82 |   -.69   .26 | 1.02   .1  1.16   .6 |  .94 |   .34   .38 | 36 Q40               |
|    79     100       .79    .84 |   -.84   .27 |  .97  -.2   .98   .0 | 1.03 |   .39   .38 | 32 Q51               |
|    81     100       .81    .86 |   -.97   .28 |  .85 -1.0   .68 -1.0 | 1.20 |   .48   .36 | 16 Q32               |
|    79     100       .79    .87 |  -1.04   .27 |  .91  -.5   .83  -.5 | 1.12 |   .43   .37 |  9 A09               |
|    82     100       .82    .89 |  -1.26   .28 |  .85  -.9   .68  -.9 | 1.18 |   .46   .35 | 10 A10               |
|    85     100       .85    .90 |  -1.32   .30 |  .84  -.8   .63  -.9 | 1.18 |   .46   .34 | 28 Q36               |
|    86     100       .86    .90 |  -1.39   .31 | 1.01   .1   .86  -.2 | 1.00 |   .31   .32 | 17 Q33               |
|    86     100       .86    .90 |  -1.39   .31 | 1.09   .5  1.08   .3 |  .91 |   .24   .32 | 20 Q46               |
|    86     100       .86    .92 |  -1.60   .31 |  .86  -.7  1.25   .7 | 1.11 |   .40   .31 |  4 A04               |
|    86     100       .86    .92 |  -1.60   .31 |  .99   .0  1.05   .2 | 1.00 |   .31   .31 |  5 A05               |
|    89     100       .89    .93 |  -1.70   .34 | 1.07   .3  4.26  4.0 |  .84 |   .14   .29 | 13 Q25               |
|    91     100       .91    .94 |  -1.97   .37 | 1.11   .5  1.05   .2 |  .92 |   .19   .27 | 30 Q49               |
|    91     100       .91    .95 |  -2.16   .36 | 1.06   .3  1.00   .1 |  .96 |   .21   .26 |  3 A03               |
|--------------------------------+--------------+----------------------+------+-------------+----------------------|
|    62.8   100.0     .63    .66 |    .00   .26 |  .99   .0  1.05   .0 |      |   .40       | Mean (Count: 43)     |
|    20.7      .0     .21    .24 |   1.23   .04 |  .11  1.0   .53  1.1 |      |   .11       | S.D. (Population)    |
|    21.0      .0     .21    .24 |   1.24   .04 |  .11  1.0   .54  1.1 |      |   .11       | S.D. (Sample)        |
+-------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .26  Adj (True) S.D. 1.20  Separation 4.58  Strata 6.44  Reliability .95
Model, Sample: RMSE .26  Adj (True) S.D. 1.22  Separation 4.64  Strata 6.52  Reliability .96
Model, Fixed (all same) chi-square:  843.3  d.f.: 42  significance (probability): .00
Model,  Random (normal) chi-square:  40.0  d.f.: 41  significance (probability): .51
```

Appendix D.4: Logit and standard error values by condition

Negative values = easier items; positive values = more difficult items

	1x		*SPS*		*SPL*	
Item	*Logit*	*SE*	*Logit*	*SE*	*Logit*	*SE*
Q23	-0.13	0.42	-0.49	0.44	-0.49	0.42
Q24	-0.13	0.42	-0.69	0.46	-0.32	0.41
Q25	-1.68	0.57	-2.26	0.75	-1.53	0.52
Q26	1.04	0.41	0.35	0.39	-0.85	0.44
Q27	0.31	0.39	-1.13	0.49	-0.91	0.48
Q28	1.05	0.39	1.38	0.42	0.51	0.39
Q29	0.60	0.39	0.04	0.41	0.51	0.39
Q30	1.53	0.41	1.21	0.41	-0.13	0.41
Q31	1.21	0.41	-0.69	0.46	-0.85	0.44
Q32	-0.69	0.45	-0.91	0.48	-1.53	0.52
Q33	-2.51	0.75	-0.31	0.43	-2.18	0.63
Q34	-0.49	0.42	-0.49	0.44	-0.91	0.48
Q35	-0.32	0.41	0.04	0.41	-0.13	0.41
Q36	-1.06	0.46	-1.38	0.52	-1.79	0.63
Q37	0.04	0.41	-1.83	0.56	-0.49	0.44
Q38	-0.13	0.41	-0.85	0.44	-0.31	0.43
Q39	1.74	0.41	2.27	0.46	1.21	0.41
Q40	-1.16	0.51	0.00	0.39	-1.38	0.52
Q41	0.51	0.39	-0.16	0.40	-0.31	0.43

	1x		*SPS*		*SPL*	
Item	*Logit*	*SE*	*Logit*	*SE*	*Logit*	*SE*
Q42	1.74	0.41	0.90	0.39	1.93	0.44
Q43	1.26	0.39	0.90	0.39	1.21	0.41
Q44	0.54	0.41	-1.16	0.51	-0.67	0.43
Q45	1.04	0.41	0.20	0.40	-0.32	0.41
Q46	-1.38	0.52	-1.79	0.63	-1.28	0.48
Q47	-0.69	0.45	0.66	0.39	-0.32	0.41
Q48	0.46	0.39	1.04	0.41	0.66	0.39
Q49	-2.66	0.75	-2.03	0.63	-1.44	0.56
Q50	-0.49	0.42	0.38	0.41	-0.69	0.46
Q51	-0.49	0.42	-0.69	0.45	-1.79	0.63
Q57	1.74	0.41	0.46	0.39	0.71	0.41
Q58	2.70	0.48	3.32	0.59	2.59	0.50
Q59	1.57	0.40	3.01	0.54	2.59	0.50
Q60	1.11	0.39	2.73	0.51	1.38	0.42

Appendix E: Post-test survey open-ended responses

Examinees' verbatim responses to "Why did you choose Yes or No? [to ' Did you prefer being able to have control over the audio?']" are presented first under Yes, then under No, with examinees ranked by logit value numerically from lowest (less proficient) to highest (more proficient):

Appendix E.1: "Yes" responses to preferring control

-1.89	because I think this to give confident .
-1.52	Not easier to control audio and answering on both time

-1.38	because I understand
-1.15	because the audios are fast and some times I can't remember it.
-1.03	because , if i can control over the audio, it is easy for me to understand.
-0.90	Because it's easy to use.
-0.89	can listen to many times
-0.86	Because, is difficult to understand from first time also If I lesson many times I able to discover the new vocabulary and i can fox to the correct pronunciation .
-0.73	Because I was able to control the audio.
-0.65	because it helps to refer to the missed parts during listening
-0.65	IT WILL HELP MORE TO DEVELOP MY MEMORY TO REMEMBER EASILY AND EFFECTIVELY..
-0.61	i can listen clearly
-0.53	because I can't understand and I need to start again
-0.52	because i can read the equation first
-0.52	because i could hear some details that i can guess before i decide to choose.
-0.49	I try to practice the audios; I mean iI can try to listen and answer the conversations in the audios. / I'm really so excited the practice listen test that is amazing.
-0.39	Because it can flexible
-0.37	i do not prefer
-0.27	I can listen many times
-0.25	I prefer being able to have control over the audio beacause when you are 100 person on the room its hard to listen every thing .
-0.25	To read the question before we listen and understand it.
-0.15	Because you can listening again if you don't stay sure
-0.03	Yes because we can listening carefully every section... with timer its OK...
-0.03	because they are clear , but some conversation i misunderstand
-0.02	to improve my ability
-0.02	to listen again
-0.01	because some of them i don't understand so i prefer if i can listen two times .
0.09	I CAN LISTEN AGAIN
0.10	can be more concentrate / simple
0.11	I had to guess a lot of answers, but the audio is pretty clear.
0.11	because it will help me to control the audio or if i want to listen again to understand more
0.23	because I don't know what are the questions
0.23	because i want to read the questions before i listen to the lecture.
0.23	if I can control the time, it will be easy for me to understand what are they talking about.
0.35	I choose yes
0.36	for this moment, I prefer to have control and practice my audio for next time . I'm sure I can do but but step by step.
0.48	because sometime we might need to hear it twoice
0.48	Yes because you can be able to listen again to the lecture if you miss some thing.
0.61	i choose yes because sometimes i want confirm if what i thought is the right things.
0.63	it is more flexible.

0.63	Yes because it can help you to replay the discussion. also to listen better
0.63	i choose YES because it helps me more to develop my listening skills.
0.76	to have chance to read the material before to understand well
0.77	to answer the question
0.77	YES, Because sometimes I miss some words, listen again is good for me.
0.77	i CHOOSE YES BECAUSE IT IS MORE COMFORTABLE WHEN YOU ARE ABLE TO CONTROL IT, THEN YOU CAN LISTEN AGAIN
0.91	i preferred being able to have control over the audio because i can listen many time
0.91	to be sure.
0.91	the sound too low, i need to replay it.
0.91	I very sure that the result will show all. When i can control it will be better to prepare before listen and during the listening i can maintain my time. Moreover, Its hard to listen and read all the information in one time. Its also very easy to be confused to follow the question and follow the listening.
0.91	easy to listen
1.02	Because I can understand
1.17	because the lecture is too fast and there are some words I didn't hear.
1.22	Because I will find what I miss in the audio.
1.33	Yes, because sometimes i do not understand some words or expressions.
1.33	because you let me paper and ready for the questions
1.38	Because I can remember what they said.
1.38	it is much easier to get information
1.50	i wanna know my scord
1.50	Because I could listen again and make sure about the answer.
1.50	Because if it is long, I can pause it and going answering some of the questions.
1.50	Because I feel more comfortable, less stressfull when I can control
1.56	Because sometimes it is difficult to me identify some words.
1.56	Because if i'm not sure about questions, I'm able to check it again and again to make sure answer correctly.
1.56	because i can read the question before, so i can form a little bit about the the information i need to find
1.74	It makes me feel more comfortable. Also, I could have more time to prepare for the next test.
1.76	Because i can control the volume as i can
1.87	I had to chose one so i chose the one that is more helpful for a test context.
2.09	I chose yes because if you don't understand what was said in the audio you can go back to listen again.
2.17	Since the efficient way to answer the questions is to knowing questions before you listen to the audios. Moreover, some times I lost the key words.
2.17	Because I can listen to find the missing information
2.19	I choose yes above because I strongly believe that hearing is better to improve our English skills. To explain clearly, if we want to be comfortable in English, the pronunciation of words seems to be important for me. By having control over the audio, we can easily recognize words. In addition, reading questions during a test might take more time than when you hear it. According to these two main point, I will

	prefer being able to have control over the audio.
2.19	Because I can listen again what information I miss
2.33	If I can control the audio, I can replay it to make sure that I won't miss important information.
2.42	more convenient
2.45	Because i can go back if i did not understand something.
2.45	I think, it's better if the time count when I click "start" on the audio bar, it's help me to control the exam and feel more comfortable, easier to focus on the listening part.
2.72	I prefer being able to have control over the audio because sometimes the conversation is too long for me to remember all of its contents to answer the related questions, even I can hear and understand it. So I have to re-play some parts of the conversation if needed.
2.75	for replay the audio
2.97	as i am not a native English speaker, sometimes it is difficult to understand some speech when i heard them one time. so i have to replay for better understanding

Appendix E.2: "No" responses to preferring control

-1.15	Because i feel very fast ,I don't understand.
-0.86	because i could understand
-0.77	The audio was so fast.
-0.73	many new word I do not know, so sometimes I can not understand all of the audio. However, I think this is a interested test I have.
-0.49	I am not listening is not good. I tried to listen it.
-0.27	because i did not have enough time to read the question, and the audio was running
-0.16	Speaker talk very fast , I can't hear clearly by one time.
-0.03	Because I have to try to understand at first like in the real life.
0.09	I choose no because in is batter to practice lestening.
0.22	i think before the audio starts, it should let student have a little time to read the question and answers. So that student can understand and prepare for listening and respond the question better.
0.48	Because some time I need to list ion to the recorder many time to understand.
0.49	Because I prefer to focus in answer the questions
0.77	because the fact that i did not have the control helped me to test my abilities to listen quickly
1.06	When we do not control our time, we think quickly.
1.06	I would say yes because if there's an audio control it'd be somewhat easier, but I chose No instead. The main reason for this conflict is that I want to test my listening skill thoroughly. Without the control, exam takers cannot replay the audio, which reflects their true hearing ability as well as their concentration. / In short, for the test: no replay button or so. / / Have a good day ☺
1.06	Because having the control of the audio would not help me a lot to check my level of understanding to a lecture.
1.50	Because when I know that I can control the audio, I can't concentrate on the lecture. I

	am under stress if I can't control over the audio, therefore, I will do better.
1.96	I do not want to have control over the audio because it is more challenging and it is the best way for me to improve my skills , and to know if there are some improvements or not.
2.19	It`s more challenging when I can`t control it. It`s more helpful
3.55	I didn't have to control the audio. I had enough of time

Appendix F: SR video observations and commentary

The text and items for Sets 5 and 6 can be seen in Appendix A. The left column shows graduate students' behavior with instances of play control highlighted with gray background. Where possible, the right column shows a relevant interview excerpt or notes ("R:" is the researcher):

Appendix F.1: Xinyi play behavior and interview comments

SPS started around 11:11 in MP4

Xinyi Set 4 of 9: **Items 37-40** *SPS laptop warranty*	*why controlled? based on transcript, etc.*
item 1 clicked C	
item 2 clicked A	
item 2 changed to B once the woman said "the hard drive failed"	
w/ 1:48 remaining on timer, played main audio from beginning	said he was trying to discern the man's feeling transcript p. 8 yeah i_ I repeat this (uh recording) yeah, because I, I think I, I (concerned) his his feeling.
item 3 clicked C	
then paused main audio with 1:30 timer left	**[likely to play replay-context item]**
played replay-context item 4 twice before choosing B	
Xinyi Set 5 of 9: **Items 41-43** *SPS retiring coworker*	
item 1 clicked B	
item 3 clicked A	
w/ 1:45 remaining played whole set again	p. 11 it's because I wanted to, check the (answer).
item 2 hovered over A a while before choosing it	

16:09 in MP4: with 00:31 left on timer, replayed just :59 to 1:02 in stimulus	
Xinyi Set 6 of 9: **Items 57-60** *SPS climate change lecture*	
item 1 clicked C	
item 2 clicked C	
checked timer with 1:15 left	
scrolled back up to items	
at 18:11 in MP4: clicked play and moved audio to 00:20-00:21 position right away	p. 11 the third and the third and fourth question of this set I feel is uh, difficult for me.
checked timer with :54, :53 remaining - also seemed to scroll up to check timer with :42-:17	
with :17 left on timer, moved audio position from 1:07 to 1:18	p. 12 the time <LAUGH> (is xx). I think I have not enough time. (xx right), to to... (xx) the recording but the remaining part is you know the (fifteen) seconds [R: mhm] so <LAUGH> I yeah (I noticed that xx) [R: mhm] more nervous than (xx) so I, I think I have lose my mind at that point. <LAUGH> **[may have had trouble with mouse]**
clicked on item 4 C then item 3 C with less than 1 second to spare	

SPL started around 19:30 in MP4

Xinyi Set 7 of 9: **Items 23-26** *SPL apartment*	
clicked items 1, 3, 4 on 1st play	
item 1 clicked C	
item 3 clicked A	
item 4 clicked B	

restarted audio a couple of times	p. 14 I wanted to check. yeah check something. yeah so I can check every one. <LAUGH> (there was) no time limit and I feel, (I just) feel comfortable. <SPEAKERS LAUGH> (I can) check it yeah.
replayed from beginning then clicked item 2 C during 2nd play ~00:11 in	[so likely remembered it from first play]
went on after "...moved out after the 28th or 29th..." during 2nd play	
Xinyi Set 8 of 9: **Items 31-33** *SPL next-day delivery*	
started around 22:07 in MP4	
played 00:00 to 00:13	
clicked back to beginning of slider bar, played 00:00 to 00:16	p. 15 (xx) for_ I remember for the first time I didn't find the answer of_ the other two, question I think. and then I, repeated (probably. you see.)
item 1, clicked C during 2nd play	
clicked back to beginning of slider bar again	
during 3rd play, item 2, hovered over C for some time	
4th play clicked item 2 C	
scrolled to top of page at 00:41 audio position, paused main audio, scrolled back down to click replay-context item 3 D, then went on	
Xinyi Set 9 of 9: **Items 44-47** *SPL music influencers research study*	
00:19 in, clicked item 1 D	
item 3 clicked A	
item 4 clicked D	

played from beginning	pp. 16-17 but, but I think uh, this question has some problem uh uh I mean, maybe you can say how teenagers use the internet to explore the music. [R: mhm] maybe it's more, correct. (xx question.) yeah. but uh I think uh this (xx) he talk more about how the teenagers use (it) to explore music [R: mhm] (xx music) band or something like that so [R: yeah] (xx) the general (talk about it). so I actually for the first question I get confused. I really get confused about (the answer). [R: mhm] I think there's no answer to this question. p. 17 yeah I I I I didn't hear the, information so I.
at 00:21, paused 2nd play	pp. 17-18 I wanted to refresh my memory. <LAUGH> I, uh you see I changed my answer (over) the first time so I, [R: mhm] so I, I just get confused about the (xx). yeah. but I feel, that the A is also, general. (it's) "how to use a computer". [R: mhm] "how to use the internet". very general right?
resumed play, changed 1 from D to A	
clicked item 2 C	
changed item 1 back to D	

Appendix F.2: Roshan play behavior and interview comments

SPS began ~07:38 on MP4 file

Roshan Set 4 of 9: **Items 37-40** *SPS laptop warranty*	*why controlled? based on transcript, etc.*
listened once without clicking anything	
clicked 1 C	
clicked 2 B	
clicked 3 C and hovered over that choice a while	transcript p. 10 So that's the problem, I can't remember the exact meaning of the word.
played item 4 replay-context item audio slider	
clicked 4 A	
(didn't scroll back up to check timer)	
Roshan Set 5 of 9: **Items 41-43** *SPS retiring coworker*	
listened once without clicking anything	
clicked 1 B	
clicked 3 A	
replayed stimulus audio	transcript p. 11 I s- I think it's the answer but I want to be sure, you know, [R: mhm] I don't want to loss my credit. you know if_ I answer that I jump from this question to another . You know this time I know that it's not a good time, you know, I should go fast but I am afraid that pass from these questions so I see the time. p. 12 ... And because I haven't enough time I never you know, I'm not sure (xx) what time (xx). (we can see.) And I haven't enough time to, I don't know which part of the listening is depend on this. Because it's inference, I'm always afraid of inference questions. <LAUGH>

jumped from 00:02 to 00:22 in stimulus	
paused stimulus at 00:46	
clicked 2 A	
(clicked to the next screen before time expired, both for Set 4 and Set 5)	
Roshan Set 6 of 9: **Items 57-60** *SPS climate change lecture*	
listened once without clicking anything	
clicked 1 C	
clicked 2 B	
clicked 3 A	
clicked 4 B	
timer ran out	

<u>SPL</u>   ~16:21 on MP4 file

Roshan Set 7 of 9: **Items 23-26** *SPL apartment*	
listened once without clicking anything	
replayed audio from beginning to 00:08	p. 14 I changed my plan. First I want to listen to the question and I think okay there is the limited time okay it's the time to time to write the question. (xx) I can return. <LAUGH>
clicked 1 C	
clicked 2 A	
clicked 3 A	
clicked 4 B	
Roshan Set 8 of 9: **Items 31-33** *SPL next-day delivery*	
paused at 00:02 on audio stimulus	p. 15 Because I want again to read the question, then.
scrolled up and down (to view all items?)	
clicked play on audio stimulus from paused point and played stimulus up to 00:15	
clicked back to 00:02 and played whole stimulus from there	
clicked 1 C	
hovered over some options for item 2	
clicked to 00:04 on player	

clicked to 00:06	
played stimulus from 00:06 to 00:19	
clicked 2 C	
clicked replay-context item audio	
clicked 3 D	
Roshan Set 9 of 9: **Items 44-47** *SPL music influencers research study*	
paused at 00:12 on audio stimulus	
scrolled down then back up	
moved audio stimulus slider back to 00:01, played from there	p. 17 And because it's interesting about you know I play it again and again because it's a very interesting topics. The girls, they download more music, they pay more money. So it means if I was a musician, so it means that my fans perhaps the eighty percent are girls, so I should concentrate for this population.
hovered over 1 D	
played stimulus from beginning to 00:13	p. 18 Because you know I want to be sure what I (chose as my answer. Internet so music is make like xx)
played stimulus from paused point (00:13)	
clicked 1 D	
paused stimulus at 00:16	
clicked 2 C	
played stimulus from paused point	p. 19 Yes, just for fun because. <LAUGH>
clicked 3 A	
clicked 4 D	