# VIDEO-MEDIATED LISTENING PASSAGES AND TYPED NOTE-TAKING: EXAMINING THEIR EFFECTS ON EXAMINEE LISTENING TEST PERFORMANCE AND ITEM CHARACTERISTICS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

SECOND LANGUAGE STUDIES

AUGUST 2017

By

Justin Cubilo

Dissertation Committee:

James Dean Brown, Chairperson
Betsy Gilliland
Theres Grüter
Ronald Heck
Elvis Wagner

# ACKNOWLEDGEMENTS

The long, and sometimes arduous, process of completing this dissertation would not have been possible without the encouragement and support of those around me, and I would like to take this opportunity to thank those people that have helped. In particular, I would like to thank my committee. Without their expertise, encouragement, and patience, this study could have never been accomplished. First and foremost, I would like to thank my committee chair, Dr. James Dean Brown, for supporting and believing in the decisions I have made throughout this study, as well as for his valuable feedback and advice that has come at every step along the way. I am also grateful for the thorough feedback and critical advice received from Dr. Theres Grüter while working on this study. This dissertation was made better through her careful attention to detail and thoughtful questions. I would also like to thank Dr. Elvis Wagner and Dr. Betsy Gilliland, whose feedback, especially in the early stages of this project, and support have shaped this dissertation into what it has become today. Finally, I would like to thank Dr. Ron Heck for his valuable input while I was conducting my data analysis. Through his input, I was able to ensure that I was headed in the right direction, making a difficult process more manageable.

In addition to my committee members, I would also like to take the moment to thank Kenny Harsch, Priscilla Faucette, Joel Weaver, and Christine Guro for their support in data collection at the University of Hawaii. I would also like to thank my colleague A.C. Kemp at MIT for being so willing to introduce me to the different people who made it possible for me to complete my data collection after moving to Boston. Furthermore, I thank the instructors who let me visit their classes in my efforts to recruit participants, and I thank the participants for their willingness to take yet another test in addition to the ones I am sure they face over the course of their busy semesters. I thank my office mates Jeongyeong Park, Gordon West, and Shirley Hsu

for their support and camaraderie, advice, and support that we could provide each other while each of us was working on our own projects. Erika Lessien deserves my special appreciation for giving feedback on my writing along the way, for being willing to take extensive amounts of her time earlier in this project to help in filming lecture materials, and for volunteering to take even more of her time to serve as a second coder for my survey data. Without her willingness to do this, there may have been no lecture recordings to begin with. Finally, I would also like to thank Sarah Goodwin. Her help with some of the analysis in this study and her willingness to lend an ear to ramblings about this project as well as her camaraderie helped me to make it through the writing of this immense project.

Funding for this dissertation was funded through a Small Grant for Doctoral Dissertation Research from ETS, a Doctoral Dissertation Grant from The International Research Foundation for English Language Education, and the Elizabeth Carr Holmes Scholarship Fund from the Department of Second Language Studies at the University of Hawai‘i at Mānoa. I am extremely grateful to these organizations for their help in making this research possible.

Last but not least, I thank my parents for teaching me the value of hard work and perseverance and for showing me the importance of an education, which has served as a continuous source of encouragement while I have worked on this dissertation.

# ABSTRACT

Technology has created many implications for second language (L2) listening assessment, particularly as it relates to the role of visuals and typed note-taking. However, while previous research has investigated the effects of visuals and typed note-taking on listening test performance, the results of these studies have been contradictory at best, with research indicating that visuals and note typing both help and hinder performance. Therefore, the present study was designed to further investigate the role that visual and note-taking conditions have on L2 listening comprehension and item performance.

Two hundred L2 English learners participated in this study with each participant being randomly assigned to one of eight experimental groups in which they took two forms of a listening test exposing them to each of the input (video-based versus audio-only) and note-taking (handwritten versus typed) conditions. Data consisted of the test scores for the overall test, subscores for items targeting different listening subskills, and responses to an open-ended survey asking participants about their personal preferences for and perceptions of the different conditions.

Results revealed no significant effect of input or note-taking on overall test scores or on item difficulty. While items were slightly more difficult in video and typing conditions, these results did not significantly contribute to item performance. A path analysis investigating the relative relationship between input and note-taking conditions on listening subskills found that video made significant contributions to participants' abilities to identify details in the listening which potentially affected participants' abilities to identify the main ideas of the listening and make inferences. Qualitative analyses showed that participants preferred video-based listening texts and that note-taking preference tended to be a matter of comfort.

The findings offer several important implications for the development of L2 listening tests. While video may not significantly contribute to listening scores, it may impact certain listening skills, which may be grounds for using video-based passages. Additionally, while typed note-taking did not appear to impact scores, it did provide a sense of comfort to some participants, indicating that its affective benefits may be a reason for allowing test takers to take notes in this way.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1
# INTRODUCTION

Listening comprehension as a construct is one of the many skills targeted in language assessments and has been the subject of debate as technology has created new ways for developing and administering listening tests.[1] While traditional definitions of listening comprehensions previously formulated by assessment experts such as Lado (1961) initially defined listening comprehension as the pure transference and understanding of meaning in sound waves, more recent conceptualizations have sought to expand this definition, taking into account the use of non-verbal cues (i.e., those cues not transmitted via sound) in processes associated with comprehension. Such cues involving gestures, contextual visuals, and PowerPoint slides are seen in these definitions as being just as important in constructing meaning as the auditory signals collected by the listener, and such definitions have been provided for both L1 and L2 listening comprehension through a number of studies (Morrel-Samuels & Krauss, 1992; Sueyoshi & Hardison, 2005; Wagner, 2006).

Of course, defining the construct is not enough for developing a meaningful language test since the construct itself is somewhat dependent upon the context in which the listening takes place. In order for scores to be meaningful and useful to those who are meant to interpret and make decisions based on them, an appropriate context must also be chosen that is related to where specific skills will be used. The definition of this context, known as the *target language use* (TLU) domain (Bachman & Palmer, 2010), is important for adding such meaning. For

[1] While the author acknowledges that *listening* and *listening comprehension* are different in that *listening* is the reception of sound waves by an individual's ears and *listening comprehension* is the act of processing, interpreting, and understanding these sounds, for the purposes of the present study, *listening* and *listening comprehension* will be used interchangeably to mean *listening comprehension* unless otherwise noted in the text, as is standard practice in the listening assessment literature.

instance, in a listening test, contexts could be described as a lecture hall where there is a one-way interaction between listener and lecturer or as a class discussion between several students and the instructor, with other scenarios being possible. Each TLU domain will be associated with different types of comprehension skills that must be targeted by the items developed for the test as well as different test administration conditions. For instance, in the case of non-verbal cues, the TLU domain will determine what kinds of visual information will be appropriate by determining, for example, that a PowerPoint visual might be appropriate for a lecture-based listening passage, but not for a group conversation. In addition to the types of visuals the TLU domain allows, recent advances in technology affect the TLU domain in relation to other aspects of test administration. Because technology now allows laptops to be taken to classrooms, students commonly bring computers to class and many have become accustomed to typing notes rather than handwriting them. However, even though this shift to bringing technology in the classroom has occurred, a similar shift has not yet been seen in testing contexts in which listeners are still required to take notes by hand even though it may not be their preferred mode of doing so. This situation along with that of a lack of common non-verbal cues on tests of listening comprehension causes one to question whether a test that neglects conditions found in the classroom may hinder performance and risk misrepresenting the construct being tested.

This dissertation seeks to examine the impacts that testing conditions consisting of visuals and typed note-taking have on listening comprehension scores. By testing these conditions, the present study seeks to test expanded definitions of the listening comprehension construct, which is defined here as the ability of an individual comprehend auditory information through both verbal and non-verbal channels, within the TLU domain of academic listening. In defining the TLU domain, the focus of this dissertation is on listening material related to

academic lectures in which students have access to information presented on PowerPoint slides, lecturer gestures and lip movements, and both typed and handwritten note-taking abilities. In order to situate this study within the context of L2 listening assessment research, this chapter will provide a brief overview of the previous research conducted concerning the role of visuals and note-taking medium in the context of listening assessment followed by a discussion of the gaps in the present research and how this study seeks to fill in those gaps. Finally, this chapter ends by providing the research questions this study seeks to answer as well as an overview of the layout of this dissertation.

**Overview of Research on Visuals and Note-Taking in Listening Comprehension**

While listening is an essential part of second language (L2) communication and acquisition, researchers have never agreed on how the construct of L2 listening should be defined. While originally emphasized as simply the transfer of information through sound only (Lado, 1961), more recent definitions of the listening comprehension construct have sought to add the use of visual cues as important to the listening process. For instance, Rubin (1995) has defined L2 listening as "an active process in which listeners select and interpret information which comes from auditory and visual cues in order to define what is going on and what the speakers are trying to express" (p. 7). Studies have supported this newer definition, finding that nonverbal cues such as gestures, posture, facial expressions, and lip movements are important for promoting comprehension (Chung, 1994; Ockey, 2007; Sueyoshi & Hardison, 2005). Thus, these findings would suggest that it is important for test developers to consider visual input in listening and communication, especially when sources would normally have such input for the listener to attend to.

Technology has made it ever more possible to alter the way in which visuals are provided in language assessments, making the inclusion of video input for listening exams more and more possible. Several studies have examined the impact of including different kinds of visuals in tests of listening comprehension, with results finding that visuals have various effects on test scores. For instance, Gruba (1993) found no differences between scores from audio-only-based and video-based listening tests while Brett (1997) found that test takers who watched video listening passages scored higher on listening comprehension tests, but only on certain types of tasks. Suvorov (2009, 2015) found that performance on video-mediated listening tasks was significantly lower than on audio-only and photo-mediated tasks. Such results would seem to indicate that nonverbal cues do not improve listening comprehension and that they may actually serve as distractions. However, some studies have obtained results indicating the opposite, finding not only that video-mediated listening passages aid in comprehension, but also serve to increase listener confidence.

In one study investigating the effects of video-mediated listening passages on comprehension skills, Baltova (1994) found that videos helped listening comprehension and contributed to learners' confidence in their understanding of the message of the speaker. Similarly, Wagner (2010b) found that scores on video-based listening tasks were significantly higher than those stemming from audio- and picture-based listening tasks, which may have been due to the use of nonverbal cues by the speaker in the video. Further studies by Sueyoshi and Hardison (2005) and Wagner (2006, 2008) showed that individuals use video input differently, suggesting that the use of such visuals differs based on factors related to proficiency, listening context, listening content, and the actual task. Other studies conducted with native speakers (who process listening material in much the same way as L2 listeners) have also found that gestures

and other visual input are incredibly important in facilitating comprehension and avoiding misunderstandings (Hadar, Wenkert-Olenik, Krauss, & Soroker, 1998; Morrel-Samuels & Krauss, 1992). Thus, taken together, these studies would seem to suggest that video listening passages have some effect on listening test scores and may be an important aspect of listening comprehension to include in tests of listening comprehension.

With studies seeming to find contradicting results, the question still remains as to whether video should be included in listening tests. Buck (2001) states that some people are better than others at using visual cues and states that this is a separate talent from listening ability. Furthermore, he states that visuals may unfairly advantage those who are more adept at using nonverbal cues and, therefore, tests should focus solely on comprehension of auditory information. However, others have argued that taking away natural visual cues of communication creates unnatural conditions and that the fact that some are more adept than others at using visual cues results in construct-relevant variation (Raffler-Engel, 1980; Wagner, 2008, 2010b). Thus, the removal of such visual support could be said to lead to underrepresentation of the construct of listening comprehension.

In addition to the impact of visuals, the medium of note-taking could potentially have a significant impact on test-taker performance and could serve to interact with video-mediated listening. Research by Ladas (1980) and Teng (2011) has found that taking notes helps students to stay awake, concentrate, and pay attention to lectures. Furthermore, note-taking and visual cues have been found to interact, with research finding that paralinguistic cues (i.e., non-verbal information) can signal to learners what information is important to write down (Piolat, Olive, & Kellogg, 2005). Given limitations that have been found in terms of handwriting speed (Ladas, 1980) and the rising prevalence of the use of computers in the classroom for taking notes, it

would seem likely that major differences in test performance might arise that are dependent upon which note-taking medium the student is most accustomed. Several studies have already found differences in test performance based on whether notes are handwritten or typed, with some finding that handwriting notes leads to better performance (Smoker, Murphy, & Rockwell, 2009) while others have found the opposite (Peverly, Garner, & Vekaria, 2013). Therefore, such contradictory results highlight the need for further research investigating note-taking medium as a variable in terms of its influence on listening test performance. In addition, considering that note-taking aids in recall later and that it has been found that video-mediated passages appear to take up more attentional resources (Cubilo & Winke, 2013), it would be valuable to further investigate how these two variables interact to determine how test performance is impacted.

**Research Gaps in L2 Listening Assessment**

Until now, research conducted on the effects of visuals on L2 learners' performance in tests of listening comprehension has produced inconclusive results requiring further investigation into the effects they may have. Many previous studies have failed to take into account the different types of visuals (i.e., content or context visuals) that could possibly be used for comprehension, leading to results that may not fully represent the effects that visuals have on comprehension. In addition, while previous studies have focused on the overall effects that visuals may have on the composite test score, few, if any, studies examined the possible influences that these visuals may have on performance on questions attempting to test different comprehension skill types. Therefore, it is not yet clear if visuals have an overall effect on listening tests, or if they actually enhance or hinder performance in different listening skills areas. Finally, in relation to the general effect that visual-based listening passages have on item difficulty, with the exception of one recent study (Batty, 2014), there appear to be no studies that

have examined this effect to determine what item characteristics may or may not lead to enhanced performance in connection to visuals provided to the listener while listening. Therefore, the present study attempts to examine these issues in order to begin filling in this gap and to encourage further research in this area.

In addition to issues related to research on the role that visuals play in listening comprehension tests, there has been very little investigation into the role that note-taking plays. While several studies have showed the importance of note-taking in maintaining concentration and alertness while also showing how non-linguistic cues can actually interact with note-taking behavior, there has been a surprising dearth of literature investigating the role that different note-taking media play in test performance. In particular, since typing has the potential to allow the listener to more easily look at visual cues and still continue to type words into their word processing software, it could certainly be possible that typing may make for greater use of visuals and better test performance overall. Therefore, given the rise in computer use for note-taking in the classroom, it is certainly important to conduct research in this area to investigate the impact on test performance in order to ensure that construct misrepresentation is not present. Additionally, previous research has focused on the difference in typing and handwriting notes in relation to later test performance in which the lecture was separated from the test by several days (Piolat et al., 2005; Smoker et al., 2009). However, it is unclear how this difference relates to situations in which the test questions are presented immediately after the lecture. Therefore, it is worth considering whether there will be any impact on test performance in this particular situation given the lack of opportunity to fully review and study notes.

Finally, while more recent research in L2 listening assessment has utilized mixed methods research to collect data from various sources including test scores, stimulated recall,

surveys, and interviews, the relative frequency of research attempting to do so overall is still quite low. While test scores can provide a great deal of information regarding the impact of certain conditions on listening performance, they do not present a complete picture. As mixed methods research has further established itself as a research paradigm, it has become increasingly important to make use of these methods in order to provide complementary analyses to better understand what is happening in the test-taker's mind while they are presented with each of these conditions. Doing so serves not only to provide better analysis and explanations of the differences in test scores, but it can also be extremely useful in establishing arguments related to test validity. More recent studies from researchers such as Goodwin (2017), Suvorov (2013), and Wagner (2008) have made use of these methods as additional sources for explaining test data, demonstrating how the use of both quantitative and qualitative data in tandem leads to more robust interpretations and stronger claims. However, there is still a remarkable lack of such research given its usefulness. The present study seeks to continue to develop research along this line by using a quantitative-dominant mixed-methods approach through open-ended survey data to better explain observations from quantitative analyses of test scores, thereby providing stronger interpretations of test data and providing research that fills in part of the gap associated with the lack of mixed methods research.

**Goals of the Present Study**

Based on the gaps in the current research related to this topic, the present study attempts to further investigate the role of visuals in listening comprehension by specifically targeting the role of content visuals and the influence they have not only on overall test scores, but also on individual item difficulty and on certain listening skills assessed by different test items. Additionally, the present study also seeks to fill in gaps associated with note-taking research by

investigating differences in handwritten and typed note-taking conditions in relation to test scores, item difficulty, and listening skill. This is all framed within a mixed methodological research framework in which a conversion research design for the collection of quantitative and qualitative data is performed, and both sets of data are used in conjunction with each other to arrive at more robust interpretations than what would be possible if each method were to be used in isolation of each other.

**Overview of Dissertation**

This dissertation is organized into five chapters. The first chapter has served to introduce the most important issues in research on the role visuals and note-taking play on listening test performance while also introducing the questions investigated in the current study and the research gaps that these questions intend to fill.

Chapter 2 reviews both the previous literature and methodological frameworks that will be used in this study. The chapter first opens by discussing issues related to the definition and modeling of listening comprehension and then moves on to discuss the different types of visuals used in listening passages as well as the research that has been conducted investigating the role that these visuals play in comprehension. The chapter then goes on to describe research regarding the role of note-taking in comprehension and the theoretical issues related to construct validity, specifically the way in which the listening construct and target language use (TLU) domain are defined. Chapter 2 then closes with a discussion of the mixed methods framework utilized in the data collection and analysis of this study.

Chapter 3 presents the research design, describing the participants and discussing the materials and instruments used in the study. The chapter discusses the piloting of the test as well as the way in which the test was administered to participants. The chapter then ends with a

discussion of the specific data analyses that were used to answer each of the research questions presented above. Following Chapter 3, the fourth chapter provides the results from each of the analyses conducted in this study, starting by providing the results of the quantitative analysis and then providing qualitative results obtained from participants' answers to the open-ended survey questions. Finally, Chapter 5 provides a discussion of the key results of the previous chapter and explains their meaning while at the same time indicating the implications for each of the findings of the study related to L2 listening test development (specifically related to issues of construct validity). The chapter ends by providing a discussion of both limitations and directions for future research.

# CHAPTER 2
# LITERATURE REVIEW

This chapter consists of several sections examining the literature related to test validation via issues related to defining the target language use (TLU) domain and construct and the role that visuals and note-taking play in comprehension of listening material presented to learners on tests of listening comprehension. The first section of this chapter discusses issues related to test validation and the pivotal role that construct and TLU domain definitions play in this validation. It then goes on to discuss common definitions of listening and listening comprehension, models of listening comprehension, and listening comprehension skill taxonomies found in the academic context, ultimately arriving at a definition of the L2 academic listening construct used in this study. The second and third sections of this chapter provide overviews of the role of visuals in listening comprehension and the role that note-taking plays in both L1 and L2 academic listening comprehension and test performance. Finally, the chapter closes by providing a brief overview of the methodological framework used in this study and the literature relevant to it.

**The Academic Listening Construct and TLU Domain**

In order for assessment measures to provide meaningful results that score recipients can interpret and use, developers must not only work carefully to formulate items that possess appropriate characteristics and demonstrate adequate functioning, but (more importantly) they must also carefully define the construct and TLU domain associated with the test. This section seeks to explain the importance of the role these two definitions play in the validation process and to explain how the definitions used for the current study were formulated based on previous research examining listening comprehension.

**The role of construct and TLU domain definitions in test validation.** To fully appreciate the importance of the domain and construct definition, it is useful to discuss what current notions of validity are and where the domain and construct definitions fit into them. The notion of what validity is and how to assess the validity of a given measure has undergone several changes over the past half century. Early conceptualizations of validity focused on the notions of criterion, content, and construct validity as more or less separate models. However, it has been recognized that criterion and content validity, while useful, are limited in what they can provide as supporting evidence for establishing validity since when they are used individually they only address a smaller portion of what needs to be considered for assessing the validity of a measure. This led some theorists such as Loevinger (1957) to suggest that criterion and content validities were simply parts of validation which fell under the umbrella of construct validation. Based on this view of validation, Messick (1989) proposed a unified model of validity, which included empirical methods for construct validation and consequences for test interpretation and use. At this time, Messick (p. 13) defined validity as:

> An integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. [italics in original]

Thus, with his definition Messick removed the test itself from being the focus of validation and instead placed the focus on the score interpretation and use. This would ideally be accomplished through the construction of a logic-based validity argument by gathering the necessary evidence for and against the proposed interpretation or use of the test score and the inferences that are associated with these interpretations. Kane (2006) outlines such an argument-based approach and Chapelle, Enright, and Jamieson (2008) have expanded on it to include considerations of the

construct definition and its relation to how interpretations of scores are extrapolated to performance within a TLU domain.

According to Kane (2006), validation consists of two types of arguments, an interpretive argument and a validity argument. The interpretive argument is built upon a number of inferences and assumptions that are meant to justify score interpretation and use whereas the validity argument evaluates the interpretive argument in terms of how reasonable and coherent it is as well as how plausible the assumptions are (Cronbach, 1988). Development of such arguments requires the use of a clear structure on which the argument may be based. For this reason, those who work on developing interpretive and validity arguments (Kane, 2001; Mislevy, Steinberg, & Almond, 2003) base their arguments on Toulmin's (1958, 2003) framework for creating informal arguments, which essentially requires that a chain of reasoning be established that is able to build a case towards a final conclusion, which in this case would be to determine the plausibility and reasonableness of score interpretations and uses.

Toulmin's (2003) argument structure is built on several components, which include the grounds, claim, warrant, backing, and rebuttal. As it relates to test score interpretation and use, the claim of an argument is the conclusion one draws about an individual based on test performance whereas the grounds serve as the data or observations upon which the claim is based. For example, one may make the claim that an individual learning English has inadequate listening comprehension abilities for studying at an English medium university based on the grounds that they received a low score on a multiple-choice listening comprehension test consisting of a series of lectures utilizing academic vocabulary and structures. However, the inference linking the grounds to the claim is not given and therefore justification is needed in the form of a warrant. The warrant in Toulmin's model is considered to be a rule, principle, or

inference-license that is meant to provide justification for the inference connecting the grounds to the claim. Warrants in turn need backing which comes in the form of theories, research, data, and experience. In relation to the example provided above, the warrant justifying the inference between the grounds and the claim would be that performance on the listening comprehension tasks reflect relevant and necessary language abilities needed in an academic context. This warrant would then be supported by backing that might say that individuals with low-level listening ability generally have difficulty understanding academic words, making inferences or predictions from what a speaker has said, or poor knowledge of signal words and phrases meant to hint at main ideas or important points and that such deficiencies lead to poor performance in an academic English-speaking context. Finally, while warrants and backing justify the inferential link between the grounds and claim, rebuttal data can serve to weaken the initial argument by providing evidence or possible explanation which may call into question the warrant. Going back to the previous example, a possible rebuttal may be that several of the topics presented in the lectures may have been too technical or abstract, the vocabulary may have consisted primarily of less commonly or frequently used academic vocabulary, or even that the audio qualitymay have been poor. Such data would serve to weaken the inference connecting the grounds and claim and would either have to be investigated further or accepted by the test developer with the knowledge that it places a limit on the argument. Thus, these components are all connected with each other and are essential for establishing an inferential connection between the claims and grounds.

In order to establish a connection between the claims and grounds, Kane (1992) stated that multiple inferences of different types must be used in a chain to connect observations and conclusions. Therefore, Kane, Crooks, and Cohen (1999) developed a three-bridge model for the three types of inferential bridges they thought were essential for linking arguments together in

order to move from observation (i.e., the grounds) to score interpretation (i.e., the claim). Each inference is in turn based on a series of assumptions, each of which requires support. These three inferences were identified as evaluation, generalization, and extrapolation inferences. The evaluation inference refers to the score that is assigned to an individual's performance on a measure with the underlying assumption that appropriate criteria are used to score the performance, that they have been applied as planned, and that the conditions under which the performance took place match the intended score interpretation (Kane, 2002b, 2013; Kane et al., 1999). Following the evaluation inference, the generalization inference refers to the use of an observed score as a way of estimating future performance or scores of a test taker if given parallel tasks or test forms. Finally, following generalization is the extrapolation inference that refers to predictions of how the expected score is to be interpreted as an indication of performance and scores that the individual would receive in the TLU domain. An important assumption of extrapolation is that test tasks are authentic relative to tasks test takers would be expected to perform in the TLU domain.

In applying the bridge model to language testing, Chapelle, Enright, and Jamieson (2008) describe three further inferences in their validity argument for the TOEFL iBT that can be used to strengthen the connection between the grounds and claim and these are labeled as the explanation, domain description, and utilization inferences. The explanation inference describes the relationship between the observed test performance and a theoretical construct (e.g., a construct of second language listening). The domain description inference refers to a detailed description of the TLU domain and is meant to provide a link between performances in the TLU domain and observed performance on the test. Finally, the utilization inference provides the link between the target score that has been obtained for the test taker and the decisions that will be

made about the test taker in relation to policy. Taken together, these six inferences along with their assumptions and support, which is obtained through a variety of methods, are able to provide a chain of arguments that can support the link between the grounds and claims of the overall validity argument.

Therefore, given the details presented above, central to laying the groundwork for a valid assessment is to have a clear and well-defined TLU domain and construct definition that encompasses the many variables that one will encounter in the process being tested. Doing so ensures that assessment scores can be connected to performance in the real world (Bachman & Palmer, 2010). However, it should come as no surprise that the way in which a construct or domain related to listening comprehension is defined is not necessarily agreed upon in the field of L2 assessment given the contradictory findings that have been obtained from the different studies reviewed below.

**Defining listening comprehension**. The ability to comprehend what one is listening to is undoubtedly a valuable skill that L2 learners must master in order to successfully acquire and interact with their L2 and, as such, listening comprehension has been the subject of a vast array of studies among both native and nonnative speakers of English and other languages. However, even though it is widely recognized as an important skill, L2 listening comprehension is still under-researched (Harding, 2012), the least understood of the different language skills (Vandergrift, 2010), and difficult to assess (Buck, 2001; Wagner, 2006). Part of the reason for these issues is the difficulty that plagues researchers in accurately defining and targeting listening comprehension skills since the processes of listening are not directly observable and are incredibly complex, requiring a careful consideration of what listening comprehension is in order to best develop and explain results from certain listening comprehension measures.

As definitions of listening comprehension have evolved, numerous researchers have discussed the necessity of establishing a widely-held definition of listening comprehension as a language skill; however, this definition has yet to be realized and accepted among L2 researchers and practitioners in the teaching and testing communities. One of the main sources of this problem, as Wagner (2002) suggests, is that L2 listening comprehension relies on many different processes. Rost (2011), for instance, describes listening as consisting of neurological (i.e., the physical structures of the ear and the way they transmit sound to nerve regions in the brain), linguistic (i.e., the decoding of phonological rules and parsing of syntax and prosodic units), semantic (i.e., the formation and activation of mental models, schema, and memories as well as the processes of learning), and pragmatic (i.e., inferring speaker intention, formulating responses, and consideration of social roles) processes that must work in concert with each other in order to provide the listener with meaningful input. With such an extensive number of processes involved, creating a single all-encompassing definition of listening comprehension is quite difficult. In addition to the processes involved, researchers (Bloomfield et al., 2010; Buck, 2001; Rubin, 1994; Wagner, 2002) have discussed a number of other factors that strongly influence listening comprehension and, therefore, further serve to complicate efforts to develop a universal definition of listening comprehension. These factors include the context of the situation, the purpose or context of the listening (e.g., academic listening, social interactions, listening for information), the characteristics of the listener (e.g., working memory capacity, affective features, L2 proficiency), characteristics of the speaker (e.g, accent and speech rate), task characteristics (e.g., factors affecting note-taking, time limit, types of questions associated with the text, control over playback), and text characteristics (e.g., length and complexity, visual cues, organization). In order to develop a universal definition of L2 listening comprehension, it is

necessary to take each these processes and factors into account, making the process of developing such a definition a daunting task.

As a result of each of these factors and processes, the exact definition of what listening comprehension consists of is not necessarily agreed upon by those who conduct research on this subject, leading the definition of listening comprehension as a skill to evolve over time with many conflicting ideas of what should and should not be included within the definition. For instance, earlier definitions of listening comprehension are in stark contrast to many of the more current definitions. These earlier definitions, such as one put forward by Lado (1961), placed the transference of sound and the information it brought with it as the main component of listening comprehension. In this definition, listening comprehension was strictly related only to the reception of sound waves by the listener and did not take visuals of any sort into consideration.

However, as time passed, definitions began to move away from such a confining conceptualization of listening comprehension and have come to incorporate more and more variables affecting comprehension in an effort to more accurately and effectively research, teach, and test listening comprehension. For instance, Rubin (1995) defined listening as "an active process in which listeners select and interpret information which comes from auditory and visual cues in order to define what is going on and what the speakers are trying to express" (p. 7). Similarly, Wolvin and Coakley (1996) define listening as "the process of receiving, attending to, and assigning meaning to aural and visual stimuli" (p. 69). Chung (1994) provides additional features in his definition by stating that messages that listeners hear have three types of information associated with them: oral (verbally transmitted information from speaker to listener), paralinguistic (body language, gestures, posture, facial expression, voice pitch, and rate of speech), and the visual context (items present in the environment of the conversation).

Furthermore, verbal and non-verbal elements are recognized by the International Listening Association (1995) as being important in the process of "receiving, constructing meaning from, and responding to spoken and/or non-verbal messages" (p. 4). However, as Suvorov (2013) points out, what is meant by the term "non-verbal" is open to the interpretation of the reader. Given the evolution of how researchers and practitioners define listening comprehension, it is highly apparent that the definition of listening comprehension has gone beyond the more simplistic definition put forth by Lado and has since come to acknowledge the importance in the role of the contextual factors (i.e., non-verbal cues), though, as Wagner (2007) and Olson (2003) state, there is still a place for strictly auditory delivery of sound given that certain situations are still routinely encountered by L2 speakers that do not provide non-verbal input (e.g., telephone calls or listening to the radio for information) and, therefore, the incorporation of non-verbal stimuli in such tasks would be highly unrealistic. Therefore, there must be some flexibility in the overall definition of listening comprehension to accommodate all possible situations that learners may encounter.

Since listening comprehension can be defined in such a way as to include verbal and non-verbal cues, it is possible to further define it as a communication activity (Suvorov, 2008). In the process of listening, the listener takes all the aspects of the situation, both verbal and non-verbal, into account and acquires some sort of meaning from them, possibly using the non-verbal information as support for making inferences about what is being said by the speaker. Many researchers have investigated the role of the different factors mentioned above in influencing this acquisition of meaning. For example, Ockey (2007) cited a number of studies in which it was found that such factors as prosody, rate of speech, background knowledge, and rhetorical cues have an impact on an individual's ability to listen. The use of non-verbal cues has also been

found to have an effect on listening comprehension. Sueyoshi and Hardision (2005) found that both lip movements and gestures are able to aid in the comprehension of a listening task. Ockey (2007) and Rubin (1995) found similar results suggesting that body movements, gestures, and facial expressions affect listening comprehension among learners. Findings such as these indicate the importance of considering both auditory and visual cues in defining listening comprehension. The question to consider next is how these two channels of input come together to allow the learner to create meaning from the input and convert that into an appropriate response.

**Models of listening comprehension.** Based on definitions of listening comprehension created by previous scholars, a number of models of listening comprehension have been proposed to better understand the way in which different factors interact in order to make input meaningful to the listener and where limitations prevent comprehension. While some of these models may have been developed by scholars specializing in L1 or L2 listening comprehension, it should be noted that the overall process of making meaning from aural input is essentially the same between the two groups, with L2 listening comprehension seeming to experience delays in processing because of interference from the L1 due to more ready access to phonological, rhythmic, or other characteristic patterns associated with it (Cutler, 2012). Therefore, for the purposes of this study, the theories upon which the models are based are considered as being equivalent for both L1 and L2 listening comprehension.

One example of such a model of listening comprehension is one discussed by Flowerdew and Miller (2010). They discuss three cognitive models for the listening process: the bottom-up model, the top-down model, and the interactive model. The bottom-up model states that listeners start by receiving an auditory signal and construct meaning from this signal by starting with individual phonemes and using them to build words and increasingly larger units of meaning. In

contrast, the top-down model states that learners use their prior knowledge in approaching a comprehension task. Vandergrift and Goh (2011) state that this knowledge can be experiential knowledge, pragmatic knowledge, cultural about the L2, or discourse knowledge, all of which is stored in long-term memory as schemata and can be drawn upon to aid in comprehension of aural input. Finally, the interactive model states that the bottom-up and top-down models rarely function independently of each other and, therefore, they function together to promote comprehension of aural input.

Using these models as a starting point and Levelt's (1993) model of speech production, Vandergrift and Goh (2011) developed their own cognitive model of listening comprehension, which takes into account not only the input that must be processed and comprehended by the listener, but also the response that the input requires. Therefore, they have developed an interactive model that can be used to explain both one-way (e.g., lecture-based) and two-way (e.g., conversation-based) listening comprehension. In their model, top-down processing provides the listener with the necessary schemata for interpreting input in a meaningful way and determining the appropriate response to the input. In addition, bottom-up processing allows the listener to build up comprehension by decomposing it into smaller phonemes and building them up to longer meaningful units. While these processes are going on, Vandergrift and Goh add a metacognition component that allows for parallel processing of input and output so that they can monitor the ways in which they are interpreting input and articulating output and adjust accordingly.

Adding to these models, Gruba (1999) wrote that a connectionist cognitive processing model of listening comprehension is most defensible. Drawing on the construction-integration model of comprehension put forth by Kintsch (1998), which described comprehension as a

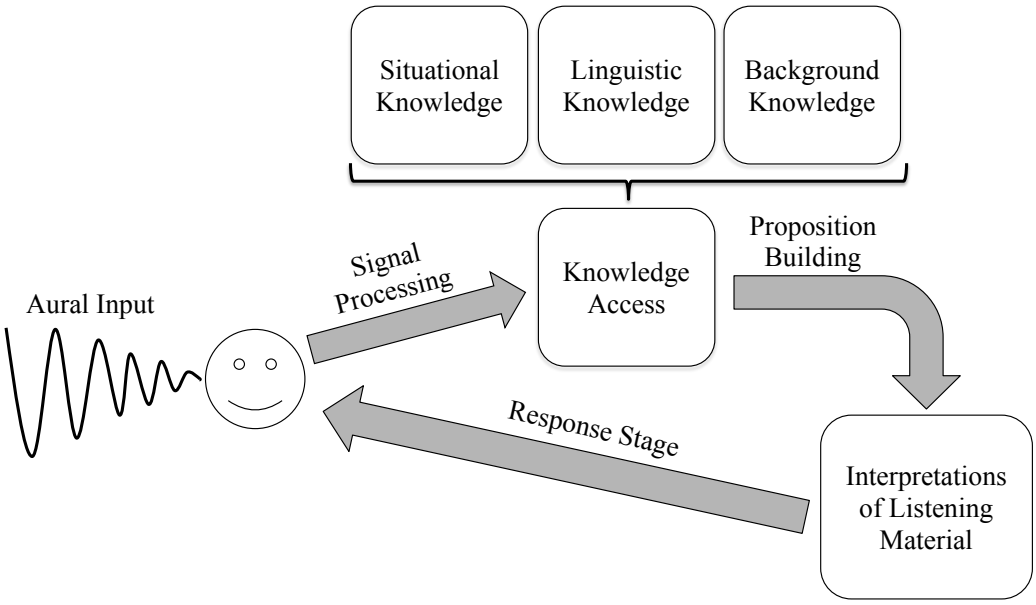



*Figure 2.1.* Connectionist framework of listening comprehension. Adapted from "Redefining the L2 Listening Construct Within an Integrated Writing Task: Considering the Impacts of Visual Cue Interpretation and Note-Taking," by J. Cubilo and P. Winke, 2013, *Language Assessment Quarterly, 10,* p. 373. Copyright 2013 by Taylor and Francis. Reprinted with permission Taylor and Francis LLC, (http://www.tandfonline.com).

process in which understanding is developed as a series of proposals or propositions that are modified within the emerging context of the text that the learner is attempting to understand (with many of these propositions being built based on the top-down and bottom-up processes described above), Gruba explains that this framework is complex enough to fully incorporate the diverse behaviors and factors that exist among listeners. Bejar, Douglas, Jamieson, Nissan, and Turner (2000) took this connectionist approach further and, since listening comprehension generally requires some form of interaction between the listener and speaker, modeled listening comprehension by splitting it into two stages: the listening stage and the response stage (this model is illustrated in Figure 2.1). Three types of knowledge need to be accessed during the listening stage in real time: situational knowledge, linguistic knowledge, and background knowledge. Each of these types of knowledge is accessed whenever an incoming acoustic signal

is received as the signals are processed. This culminates in a set of propositions being produced allowing the individual to switch into the response stage in which they use the propositions in order to formulate a response that can come in the form of writing, an oral response, or a selection from among a set of choices.

Other models of listening comprehension have also been proposed by Nagle and Sanders (1986) and Johnson-Laird (1983). Nagle and Sanders (1986) proposed a model in which an acoustic signal is received and processed via automatic and controlled processes in such a way as to produce comprehension of the listening text. In their model, they discuss three types of memory and their roles in processing the received acoustic information: echoic memory, working memory, and long-term memory. They state that acoustic signals are held briefly in echoic memory when they are first encountered. If the signal has a characteristic of interest to the individual it is processed in working memory, which uses both automatic and controlled processes to transfer the signal to long-term memory. In long-term memory, linguistic knowledge, general knowledge, and contextual knowledge map onto the incoming stimulus and are used to interpret it. One aspect of this model that makes it helpful for explaining L2 listening comprehension in particular is that it considers the greater amount of controlled processing (as opposed to automatic processing) that individuals sometimes need to perform in their L2, accounting for slower processing speeds.

While Nagle and Sander's (1986) model is useful, Buck (2001) states that it has a shortcoming in that it does not actually explain the way in which the text's meaning is constructed in memory. Buck (2001) states that this shortcoming is addressed in mental models of listening comprehension described by Johnson-Laird (1983). According to this model, an individual takes the information that is received and transforms it into propositions or mental

models. Johnson-Laird (1983) states that propositions are simple concepts or ideas associated with verbal or textual ideas while a mental model is a mental representation of events described in discourse and of the state of affairs in the world often associated with a picture or mental diagram. Because propositions alone would create a heavy cognitive burden on memory, Johnson-Laird (1983) suggests that individuals rely more heavily on mental models as a method of easing this burden and focusing on the larger situation. As a result, a listener will process input as a mental representation rather than as a linguistic representation, explaining why the listener will often remember the overall meaning of a text but not the exact language used in the text (Buck, 2001). However, even though all of the models above are useful for explaining the underlying processes that occur while trying to comprehend auditory signals, they still fail to take into account the non-verbal cues that may help in constructing meaning from the input provided to the listener. Therefore, other models must be examined to help explain the role of visuals in processing auditory information.

In addition to models of listening comprehension that focus on the processing of the auditory information that individuals receive, there have been other models proposed by those utilizing cognitive load theory that consider the way in which visual and auditory information interact while information is processed in the brain and how this can be best utilized to enhance comprehension and learning. For instance Mayer (2005) has put forth a cognitive theory of multimedia learning which assumes that both an auditory and visual subsystem exist in the individual. Based on Baddeley's (1992) model of working memory, Mayer (2005) states that the two subsystems exist within working memory and that information is processed through both of these channels. The theory also assumes that either channel can only process a limited amount of information at any given time. This is based on research conducted by Chandler and Sweller

(1991) that demonstrated the impact that unintegrated instructional materials can have on learning, showing that limited cognitive resources are overburdened when material is unnecessarily untintegrated due to limitations in working memory, leading to reduced comprehension and retention of information. The final assumption that Mayer (2005) makes is that active processing, which involves accessing one's prior knowledge and the available external information, is required for comprehension and learning to occur (similar to the concept of top-down processing described above). Mayer's (2005) theory overlaps with Johnson-Laird's (1983) theory in that both essentially propose that information is used to construct mental models. However, Mayer (2005) seeks to explain more of the cognitive limitations in comprehension and learning by proposing that both visual and auditory channels take information into working memory, where the limitations of working memory capacity can limit the extent to which the two are able to be integrated into a mental model, making it important that instructional design and every day listening tasks have information integrated where necessary so as to avoid creating unnecessary cognitive load through splitting the individual's attention.

Mayer's (2005) theory has been highly influential in relation to the field of cognitive load theory, as is evident in the theory Schnotz (2005) developed, which incorporates Mayer's (2005) theory along with those of several others. As a result, Schnotz (2005) developed the integrated model of text and picture comprehension, which, while similar to the cognitive theory of multimedia learning in several ways, differs in the fact that while Mayer's (2005) model assumes that separate verbal mental models (i.e., models based on acoustic information) and pictorial mental models (i.e., models based on pictures, written texts, or diagrams) are created prior to their arrival in working memory where the working memory capacity determines how

extensively they are integrated, Schnotz (2005) assumes that only one mental model (rather than two) is created in working memory based on information that it receives from different channels and that the accuracy of this model is dependent upon whether the information presented to the learner is within the capacity constraints of a given channel. Additionally, the integrated model of multimedia learning and comprehension also assumes asymmetric comprehension between verbal and pictorial processes. While verbal comprehension (e.g., comprehension of acoustic sounds) may rely more on proposition formation, pictorial comprehension (e.g., comprehension of diagrams) will rely more heavily on mental model formation, showing that Johnson-Laird's (1983) model may account for different types and channels of processing based on the input presented to the listener and the degree to which certain combinations of incoming information are integrated in such a way as to promote or hinder comprehension.

A key assumption of both of these models is related to the limitations associated with the sensory channels by which information is presented to the learner and the capacity of their working memory for integrating the information received from these channels in such a way so as to build schemata for comprehension. While the sensory channels found in both models may have high capacity, the working memory has limited storage capacity, therefore making it difficult for learners to work with input from multiple modalities (especially input that may be redundant or extraneous). Horz and Schnotz (2010) explain that if input is redundant (i.e., the same information is presented in text and spoken form), this essentially results in working memory being required to formulate one cognitive representation of information for one input source before having to move on to formulate yet another representation of the same information for the other input source. Since the same information must be processed at the same time, the listener misses key pieces of information because they find themselves processing information

that they may have already processed in the other format when switching between the two modalities. Additionally, extraneous information results in further straining of working memory capacity, as it requires the learner to put effort into processing information that is unnecessary for understanding learning material that they are being presented with. In both of these situations, the split-attention effect, which is the result of situations in which individuals' cognitive capacity is overburdened due to unnecessary processing, is present, hindering learner performance (Ayres & Cierniak, 2012). Rather, Horz and Schnotz (2010) state that the incorporation of multimodal instruction through both visual and auditory channels should be complimentary, effectively leading to the expansion of working memory capacity since this creates a situation in which visual and auditory channels share the burden, easing the processing burden of the individual and allowing them to more fully integrate information that they have received. Therefore, not only do these models agree with Johnson-Laird (1983) in showing that model building (through visual channels) and proposition building (through acoustic channels) leads to an easing of cognitive burden, but they also seem to suggest that multimodality is an important aspect of comprehension in general, signaling that the process of listening comprehension may utilize both the visual and auditory channels as a means of enhancing comprehension by reducing such burden.

Upon examining the multitude of models attempting to describe the process of listening comprehension, it is not surprising that there is, as of yet, no universally accepted theory that explains the process of listening comprehension (Ockey, 2007). While many theories focus on how input is processed, they consider this by focusing on different aspects. Some focus specifically on the way specific information is used in determining the meaning of a text (Flowerdew & Miller, 2010) while others focus on how information is stored and processed in

memory (Nagle & Sanders, 1986). In addition, while many researchers focus only on the input and processing of a listening text, some view the actual use and response of that information as an important part of the listening comprehension process (Bejar et al., 2000).

Based on these models and definitions of listening comprehension, a general definition of the listening construct can be formed for the present study. While each model and definition focuses on a different aspect of listening comprehension, they are not necessarily exclusive of each other and serve to offer explanatory power for observations accounted for in other models. For this study, the listening construct is defined as the comprehension of information that is transferred through both auditory and visual channels. Additionally, the process of comprehending this information is defined as being based on models put forth by Bejar et al. (2001) and Schnotz (2005) (which expand upon the other models presented here) in which relevant world and linguistic knowledge is used to make propositions that result in a response to a stimulus, with stimuli across different channels (i.e., visual and auditory) used in tandem having the potential to either aid or hinder processing of listening material for comprehension.

**The L2 academic listening domain and construct definition**. As was discussed above, the listening process is a complex interplay between a number of factors, and each of these factors must be taken into consideration when creating an assessment measure. In addition to defining the general construct to be tested, the context that the assessment is meant to represent (i.e., the TLU domain) must also be determined so that the general construct definition can be made more specific. In the case of listening tests, a number of target domains could be tested. Two examples of this would include a service worker domain in which the listener must be able to comprehend the types of requests or complaints customers will make, or an academic domain in which the listener could encounter any combination of lectures, student-instructor

conversations, and student-student conversations. Each of these domains will require listening to different registers, a knowledge of different vocabulary, and a different set of comprehension skills. The present study focuses on the latter of these two examples by focusing on lecture-based listening passages, which requires a number of considerations informed by previous theoretical contributions.

Since many large-scale listening tests are meant to predict the ability of a test taker to comprehend what they hear in an academic setting, academic listening has been the chief focus in much of the literature pertaining to L2 listening comprehension (Chaudron, Loschky, & Cook, 1994; Flowerdew, 1994; Smidt & Hegelheimer, 2004). Academic listening can be classified as two-way interactions in which a professor and a student have a discussion or in which two students are having a discussion, which they would experience when working in groups (Lynch, 2011). One-way listening (Flowerdew, 1994), which can be argued to be the most commonly experienced academic listening type (especially in the first year of study), is another mode of academic listening in which the listener receives information from the speaker without responding or would respond to at a later time in the form of a test or assignment, as one would see in a lecture. Based on these different types, it is clear that different levels of formality exist in the language with which the listener is presented, with academic lectures often providing the greatest level of formality. This register can pose a number of challenges and the nature of listening associated with academic lectures is complex, making it an important skill to test because many ESL students studying at university will be required to take classes in which English is the medium of instruction (Flowerdew, 1994). In addition, because lectures are non-interactional by their general design, this provides learners with fewer opportunities for clarification in the moment (Smidt & Hegelheimer, 2004), making it necessary that students be

well acquainted with and have the appropriate skills to succeed in such an environment. With this information in mind, the TLU domain for this study was determined to be an academic listening context in which the primary focus was on one-way, lecture-based interactions that require test takers to comprehend information presented in a formal register that requires comprehension of slightly more technical vocabulary.

Once the TLU domain is defined, it is necessary to determine what kind of listening comprehension skills are representative of this domain and how they will be represented within the test itself. Several classification systems have been proposed for how listening comprehension skills should be categorized. For instance, King and Behnke (1989) offer an early classification system in which they group listening skills into three categories, including comprehensive listening (i.e., listening to understand the gist of the message), interpretive listening (i.e., listening to draw inferences from the text), and short-term listening (i.e., listening to process information over a short period of time). Wolvin and Coakley (1996) follow this classification with their own proposed typology, stating that five types of listening exist, which include discriminative, comprehensive, therapeutic, critical, and appreciative listening. Bejar et al. (2000) provide an additional system in which they classify listening purposes or skills as listening for specific information, for basic comprehension, for learning, and for integrating information. Field (2008) adds yet another classification system by classifying listening skills or purposes into two broad categories: listening for global goals and listening for local goals.

The number of taxonomies available for listening skills does pose a problem in that the agreed upon definition of how skills or listening purposes should be classified is unclear. However, while these systems do have their differences, it is still clear that these are all necessary listening skills that must be utilized in academic listening. Therefore, in order to define

the TLU domain and how this will affect the construct definition, one must first come to a clear understanding of how targeted listening comprehension skills will be determined and the terminology that they will use for these skills. Without doing so first, there can be no clear definition of what the test is trying to assess and how this will relate to performance in the TLU domain. While the differences in terminology exist, it is clear that there are similarities in terms of the exact skills that should be tested. For instance, the majority of the classifications above mention listening comprehension skills related to identifying the gist and details of a listening passages as well as making inferences and understanding the attitude of a speaker, and these appear to be aligned with comprehension skills that many tests of academic listening comprehension used by various academic departments test, including the *TOEFL iBT* (ETS, 2012) as well as the *IELTS* (Cambridge English, 2012). Thus these are the primary skills focused on in the present study.

Based on the similarities between taxonomies and the prevalence of these skills across commonly used tests of academic listening comprehension, the present definition of the TLU domain, and by extension the construct definition of this study, is narrowed to focus on these aspects of listening comprehension. Considering aspects of the TLU domain described above, the general construct of listening comprehension described in the previous section can be further narrowed to focus on the specific construct targeted in the current study. The more specific construct of academic listening used in this dissertation can be stated as the following:

> The ability to not only comprehend the overall purpose and main details of a lecture-based listening passage, but also to make inferences about a speaker's statements and understand a speaker's overall attitude toward the information from it as well. This understanding will come from both auditory and visual channels in an academic lecture-

based environment in which listeners are exposed to a formal register with somewhat more technical vocabulary seen in first-semester university lectures.

While this definition is specific and allows for a more carefully crafted test that targets appropriate academic listening skills, elements within the academic listening environment that some may include in their construct definition are not without controversy. In particular, debate still exists over whether visual information should be part of any definition of a listening construct (as is included here), and there is some question as to how advances in technology could affect the TLU domain as it relates to test administration conditions such as note-taking conditions. Such issues are discussed in the following sections through a review of the views expressed and data obtained by researchers investigating these aspects of the listening construct.

**Visuals and the Validity of the Academic Listening Construct**

As mentioned earlier, the amount of processing required in one-way listening found in lecture-style passages is challenging and should provide an idea of just how efficiently a test taker listens to incoming stimuli. In addition to determining the type of academic listening that a test should focus on in defining the L2 academic listening construct, it is also important to consider the role that visuals play in this definition. As stated above, visuals have a somewhat controversial reputation in terms of their effect on listening comprehension. However, even with this controversy, the question still remains as to whether visuals should be incorporated into a new definition of the listening construct. This question has been the focus of a great deal of discussion in the past and continues to be so.

While it has been suggested that video should be used in tasks of listening comprehension which are based on audio that originated with video (Buck, 2001), test developers have frequently rejected the use of video in their listening tests (as reported in Wagner, 2008). While

limitations related to inadequate video or recording devices may make the inclusion of videos in in-house tests in university departments less feasible, some have stated that decisions to exclude video in situations where these limitations do not necessarily exist are still made, even when material comes from video-based material (Coniam, 2001). This raises serious questions of the construct validity of these tests (i.e., are the tests measuring the full range of listening ability?), the answer to which has a significant impact on the definition of the L2 academic listening construct. If research points to the fact that listeners make use of certain kinds of visuals in processing aural stimuli, then it stands to reason that excluding such visuals would lead to construct underrepresentation and, thus, less valid results.

Just as research has supported both sides of this issue, scholars have also weighed in to express opinions arguing on both sides. Buck (2001) was concerned that research has shown that people differ in their abilities to use visual cues while listening. Therefore, he has argued that inclusion of visual cues in an assessment may create a situation in which certain test takers who are more adept at using non-verbal cues are given an unfair advantage over those who are not particularly adept at using them. He has therefore argued that it is better for listening assessments to focus on the comprehension of strictly auditory information. In addition to Buck, Gruba (1993) expressed concern with how the use of visual information would affect the overall construct validity of listening tasks, thus seeming to agree with Buck (2001) in the idea that the verbal aspects of listening tasks, and communication in general, are more important than the non-verbal aspects.

On the opposite side of the argument, many scholars have taken the stance that visuals, particularly videos, that are naturally connected to the audio being presented actually help the construct validity of a listening test. Without the presence of video, the construct validity of a

listening test would be endangered. Von Raffler Engel (1980, p. 235), for example, suggested that taking the visual cues of communication created, "an unnatural condition which strains the auditory receptors to capacity." In other words, removing the visual channel creates unnecessary strain for the test taker and does not accurately reflect the natural environment in which test takers would use their listening ability. Bachman and Palmer (1996) expanding on Messick's (1989) statement that tests are valid only if they reflect the context of learning, proposed the idea of the TLU domain. This would encompass the issues mentioned above regarding the different skills and purposes that exist for listening as well as the type of academic listening one encounters. In addition, their proposal states that the language tasks should reflect what the test taker will encounter outside of the testing domain (the extrapolation inference mentioned above). Therefore, if a learner who goes to lecture has access to visual information through the lecturer's gestures, writing on the board, or through slides, then the listening task should also include these features. Without such features, the test validity is threatened due to construct underrepresentation (Wagner, 2006).

Adding to these arguments, other researchers have come to conclusions stating the importance of including visuals as a part of the listening construct, asserting that their inclusion is an important part of the test. For instance, Cubilo and Winke (2013), in their study investigating the impact that visuals played on an integrated writing assessment, argued that, while some students may have found the inclusion of videos on the listening task to be distracting or confusing, the vast majority still appreciated them and stated that they helped. They took this information a step further to assert, in opposition to Buck (2001), that removing visuals from a listening task would unfairly disadvantage those who do not know how to use visual information effectively. They stated that students are expected to sit in lecture where they

are presented with a variety of visual information such as gestures, graphs, embedded videos, and writing on the board. Therefore, if a student cannot efficiently make use of this information while concentrating on the aural input from the instructor who is teaching them in their second language, they risk poorer performance than those who are able to attend to all of this information. With this in mind, they argued that it is necessary to include visuals in a listening assessment since being able to manage these different input sources is an essential part of performance in a lecture setting and should be included within the listening construct.

In addition to Cubilo and Winke (2013), Suvorov (2015) also responded to Buck's (2001) claims. In his study tracking the eye movements and attention to visuals that participants exhibited in a video listening task, he found that test takers, even when they had the opportunity to look away from visuals to help themselves to concentrate better, interacted extensively with the videos in his listening assessment. He interpreted this to mean that students chose to use both visual and aural information while listening to the academic lectures he presented to them. Using these results, Suvorov stated that Buck's (2001) belief that L2 listening assessments should focus only on the processing of auditory information since visuals would "serve to increase the cognitive load of the test taker, and that may interfere with the testing process" (p. 254) is outdated. This conclusion on his part comes from the idea that if the visuals were indeed causing excessive cognitive burden on the test takers, they would have chosen to avoid making eye contact with the videos to focus only on note-taking or close their eyes to concentrate. Cubilo and Winke (2013) actually observed such behaviors in a minor portion of the students, supporting Suvorov's (2015) claim.

Discussion of issues related to the role of visuals in the definition of the listening construct has led to a great amount of research in this area. However, it is evident based on the

discussion above that there is still much to examine before test developers can have any hope of establishing a clearer answer. While there have been extensive studies examining the role that different visual conditions play in test taker performance overall, little research has examined the effect that such conditions play in relation to item performance. This is especially important to consider when one attempts to define the TLU domain and realize that an academic listening task is meant to test a variety of skills or purposes. Having knowledge on which types of visuals have the greatest impact on item performance and what types of skills are most affected is important for more fully understanding the conceptualization of how visual input fits into the listening process. At present, few studies have attempted to examine such effects, with the present author finding only one (Batty, 2015) that has examined item bias for visual format. However, even this study did not fully take into account the different types of listening skills or purposes that a test of academic listening examines. Therefore, the present study attempts to fill this gap by more fully examining this issue. In addition to a lack of studies examining the effect of visual presentation on individual items, there is a surprising lack of research into the role that note-taking using different media plays in the comprehension process. The following section provides an overview of the research in this area and its application for listening assessment.

**Types of visuals**. With several researchers (Hadar et al., 1998; Wagner, 2013) making convincing arguments that visual cues play just as important a role in listening comprehension as the auditory stimuli that individuals receive, it has become important to define and classify the types of visuals that exist clearly so as to avoid ambiguity in assessment development and making it possible to more definitively determine the effects that visuals have on listening comprehension. These visuals that the listener has access to while trying to process and respond to auditory input are often divided into two primary types: content visuals and context visuals

(Bejar et al. 2000; Ginther, 2002). Content visuals carry information relevant to the content of the spoken stimulus while context visuals carry information about the context or situation in which the speech act is taking place (Ginther, 2002). Bejar et al. (2000) further break each of these categories down into subcategories of visuals. For content visuals, they describe four classifications according to their function relative to the auditory stimulus: (a) visuals that replicate the oral stimulus; (b) visuals that illustrate the oral stimulus; (c) visuals that organize information in the oral stimulus; and (d) visuals that supplement the information from the oral stimulus. The first type of content visual would be the inclusion of words written on a board or projector slide that match the aural input exactly. An example of the second type of content visual would be one in which the presenter provides a picture to the listener that portrays the information being spoken while an example of the third type of content visual would be one in which the speaker provides a diagram that presents information in a different way than the auditory information is presented. Finally, the fourth type of content visual would be a visual that provides information that has not been included in the auditory input. Bejar et al. (2000) hypothesized that, while the first three types of content visuals in their classification would be helpful in L2 listening comprehension, the fourth type would actually risk creating more difficulty.

Bejar et al. (2000) also classified context visuals into three different types: (a) visuals with information about the setting that are either relevant or irrelevant to the information provided to the listener, (b) visuals with information about the participants in the oral stimulus; and (c) visuals that provide information about the text type. Context visuals falling into the first category would include pictures that portray a lecture hall where a lecture is given or a coffee shop where two friends are having a discussion. Context visuals falling into the second category,

on the other hand, would include some mark to the listener indicating whether the speaker is a professor, student, or a group of friends talking to each other. Finally, an example of a context visual in the third category would be a visual of a professor at a podium giving a lecture.

Beyond being distinguished as content or context visuals, researchers have also further classified visuals in terms of their mode of delivery or format. For instance, visuals can be delivered as single still images, a series of still images, or as videos (Ockey, 2007). Other content-supporting visuals described by Ginther (2002) have included charts and graphs, flow charts, or other drawings. This classification of visuals into static versus dynamic animation (McCuistion, 1991) has become somewhat standard in the area of L2 listening comprehension and continues to be useful for better understanding the effects of visuals on listening comprehension.

Additional visual classifications have been described based on how they are used in different fields of expertise as well as the types of body language that is used. Rost (2011) described a classification system in which visual signals are designated as being either kinesic or exophoric. Kinesic signals are related to body movements and gestures that can be used by the listener as a signal meant to emphasize the importance of information or to draw attention to certain aspect of the listening text. These signals include baton signals, which are head and hand signals that emphasize key aspects of the verbal message, directional gaze, which consists of eye movements or eye contact with members of the audience, and guide signals, which consist of any gesture or movement of any part of the body that draws attention to or emphasizes specific points in an oral message. In contrast to kinesic signals, exophoric signals are external to the speaker and include references to the oral input such as writing that has been done on the board or points found in a PowerPoint presentation.

In addition to Rost's (2011) proposed classification system, Desnoyers (2011) proposed a classification system that was focused specifically on visuals that speakers in the sciences generally use. Desnoyer (2011) divided these visuals into three categories: (a) cosmograms, (b) typograms, and (c) analograms. Cosmograms consist of picture-based visuals that are meant to signify objects or environments. Examples of cosmograms would be photographs of objects, diagrams of buildings, or maps. Typograms are language-based visuals with text and numbers that could be visuals such as flow charts or tables. Finally, analograms consist of visuals that represent data using graphics. These would be graphs such as scatterplots or line charts, pie charts, or circle diagrams.

Given the number of classification systems, a single visual is capable of being classified in a variety of ways depending on the focus of the classification system being used and according to the purposes of the test developer or researcher. While these classifications are widely utilized, particularly the content-context distinction, the fact that some visuals can be classified in multiple ways leads to some ambiguity, leading to the conclusion that classifications can be somewhat arbitrary in nature. For instance, a video-based listening passage may have both content and context visuals present simultaneously, raising questions of how to best control for one or the other. Furthermore, while having these different classification systems is useful, research is necessary to see how these different visual classification affect listening comprehension and learner interaction with the material.

**Role of visuals in listening comprehension**. A number of researchers have sought to examine the influence that visuals have in the listening comprehension process, with findings generally establishing that non-verbal information can have a significant impact on the overall process. Evidence for this influence can be seen not only in L2 listening comprehension, but has

also been seen consistently in L1 listening comprehension (which is not surprising given that the same processes are used in the L1 and L2). For example, Ochs and Schiefflin (2009) found that in L1 acquisition children rely extensively on non-verbal cues when they are developing their L1 speech perception and interacting with caretakers. In addition, Morrel Samuels and Krauss (1992), in a study investigating the interplay of gesture and speech in interaction, found that gestures actually serve to facilitate speech production and can even be an aid to listeners who are listening to a speaker of their native language. Likewise, Hardar, Wenkert-Olenik, Krauss, and Soroket (1998) found that gestures are able to help native speakers negotiate the meaning that the speaker is attempting to convey in instances where there may be misunderstanding and that they actually aid native speakers of a language by helping them recall lexical items more quickly. Furthermore, McGurk and MacDonald (1976) demonstrated the importance of lip movement in their study in which they provided listeners with a recording of a speaker using lip movements of "ga" while having a soundtrack for the sound "ba." The result of this experiment showed that the vast majority of participants perceived the actual sound as "da," an intermediate sound between the two actually present in the video suggesting the importance of lip movements as a visual in listening comprehension. Based on these studies, it appears that the presence of visuals play an important role in listening comprehension in L1 listening comprehension.

In addition to their effects on L1 listening comprehension, visuals have also displayed a number of influences on L2 listening comprehension. As mentioned above, they have been demonstrated to help the listener identify the context and speaker's role (Bejar et al., 2000; Ginther, 2002). In addition, Ockey (2007), in a study investigating differences in learner performance between different visual types, found that visuals play a key role in promoting learner access to background knowledge about the listening material. Additional benefits of

visuals have been found by Sueyoshi and Hardison (2005) in a study looking at the way in which lip movement and gestures affected the comprehension process. In their study, they found that listeners with access to the visual channel led to increased listening comprehension while also finding that proficiency level played an important role in what visual cues L2 listeners attended to. In particular, they found that lower proficiency learners make use of body language and gestures in order to aid listening comprehension, while high proficiency learners make use of lip movements to aid their comprehension. These findings support findings by other researchers (Kellerman, 1992; Rost, 2011) that also indicated that non-verbal cues help learners to fill in gaps in listening comprehension. In addition to these findings, Sueyoshi and Hardison (2005) also found that access to visuals helps learners develop more positive attitudes towards L2 listening tasks, supporting similar findings made by Progrosh (1996) and Wagner (2010b).

Although several researchers have found positive effects of visuals on listening comprehension, others have also argued that they may lead to negative effects on listening performance. Bejar et al. (2000) determined in their study that visuals displaying information not related to the auditory input may lead to confusion and poorer comprehension and Rubin (1995) stated that visuals that do not fit into the L2 listeners' expectations seem to have a similar effect. Moreover, since individuals have certain cognitive load capacities that restrict their ability to fully attend to all of the input around them efficiently and effectively (Moreno & Park, 2010) and, since visuals are viewed by some to increase the already high cognitive load of an individual functioning in their L2, visuals may only serve to further slow down processing and interview with L2 listening comprehension (Vanderplank, 2010). Finally, while visuals may have some positive effects, the way in which they are executed could serve to make an otherwise effective visual something that harms the listeners' performance. For instance, Pettersson (2002)

states that factors related to size, shape, color, light and shadows, composition, quality, format, pace, and editing can all work against the listener to some extent if executed poorly.

Visuals clearly play a significant role in both L1 and L2 listening comprehension, even though the actual effect that they have on one's ability to comprehend incoming oral stimuli is not always entirely clear. While many argue that the presence of visuals allows for more positive attitudes towards the listening experience and can help the listener to fill in gaps in information they may have missed, the delivery of visuals could have significantly negative influences on individuals' abilities dependent upon how the visuals are presented. Given that visuals could deliver both positive and negative influences, it is essential that test developers conduct research on the effects that visuals have in the testing environment prior to including visuals on their tests.

**Research on Visuals in L2 Listening Assessment**. Based on the different classifications systems of visuals and what is known about their role in listening comprehension, researchers have conducted a number of studies investigating the effects of visuals on L2 listening test performance in the past several years. However, while many studies have indicated that visuals do have an effect on test taker performance, the exact nature of their role is still unclear. In some of the studies, results have indicated that visuals provide support for listening comprehension, thus leading to improved performance on listening tasks (Baltova, 1994; Chung, 1994; Wagner, 2010b). In contrast, other studies have suggested that there is either no effect on performance in listening tasks (Gruba, 1993; Baltova, 1994; Londe, 2009) or that visuals or certain types of visuals (most commonly videos) can have negative effects on tasks (Coniam, 2001; Brett, 1997; Suvorov, 2009).

One study that found two different results was performed by Baltova (1994) in which she conducted two separate experiments in which Canadian students learning French were assigned

randomly to groups. In the first experiment, the students were assigned to sound-only, video-and-sound, silent viewing, and no-story groups (i.e., answering questions with no exposure to any element of the listening passage) and were asked to complete a multiple-choice test after the listening passage was completed. The findings showed that those in video-and-sound and silent viewing groups actually scored almost twice as high as those who were in the sound-only group. In the second experiment, Baltova (1994) then separated students into sound-only and video-and-sound groups, asking them to complete a slightly longer multiple-choice test after the listening passage. The findings of this experiment found that there was no significant difference between the groups. While the findings from these two experiments are contradictory, it is important to note that a pre-test was not used and, therefore, the learners' ability levels were not controlled for in any manner, leading to potentially inaccurate results.

Similar to findings of Baltova's (1994) first experiment, Chung (1994) conducted an experiment in which 75 participants grouped as advanced and intermediate learners and non-French Learners (serving as a comparison group) were presented with four different French dialogues with increasing amounts of visual information. Dialogues in this study ranged from being audio only to one still picture to multiple still pictures to moving-video images. Among many of the his findings, Chung's key results indicated that visuals almost always improved comprehension of the dialogues, with video conditions showing the greatest improvement. Additionally, he found that multiple still images were distracting in certain circumstances and that paralinguistics (e.g., body language and gestures) aided in interpretation of the dialogues, especially for those at higher proficiency levels. This finding is in contrast to Baltova (1994) who found no significant contribution of video-based visuals to test performance; however, unlike Baltova (1994), Chung (1994) took proficiency level into account. Additionally, the findings of

this study support those of Sueyoshi and Hardison (2005) who found that higher proficiency learners tend to use non-verbal cues in a more effective manner. Interestingly, Chung (1994) also found that test takers tended to perceive the video-based audio passages as faster (a finding similar to that found by Cubilo and Winke (2012)) even though the dialogues were carefully monitored to ensure similar speeds across conditions. This indicates that the presence of visuals, while not necessarily having a negative effect on comprehension, may serve to increase the overall cognitive load on the L2 learner.

In addition to these two studies, research conducted by both Wagner (2010b) and Ginther (2002) also supports claims that visuals have positive effects on learners' listening comprehension. For instance, Wagner (2010b) split 202 participants into two groups in which one group watched a video listening passage and another watched an audio-only version of the same passage. His results showed that individuals in the video group scored 6.5 percent higher than participants in the audio-only group, which proved to be significantly different. Ginther (2002) found similar results from a study in which she used the TOEFL listening section scores of 160 participants to examine the role played by content and context visuals (with the visuals being static pictures) on test performance. Her results indicated that content visuals that were complementary to information from the listening text significantly increased scores when compared to those participants who did not have access to visuals while context visuals did not seem to influence scores significantly. Thus, based on these studies, it would seem that evidence exists for the positive role that visuals play in the listening process.

However, while the studies mentioned above indicate that visuals have a positive influence, several other studies have produced results that indicate that visuals may have no effect at all. For instance, Gruba (1993) conducted a study comparing the listening test results of

91 ESL students placed in video or audio-only groups finding that no significant difference in scores existed between the two groups. However, no pre-test was delivered prior to the start of the test to establish a baseline, similar to Baltova's (1994) second experiment, and the test had a low reliability that may have affected test scores. Similarly, Coniam (2001) tested 104 English language teachers on their listening ability by placing them in audio-only and video test groups that were presented with a talk show discussion on a current education topic, finding no significant difference in scores. Coniam added to his results by providing a questionnaire to the test takers and found that the vast majority of the participants (82 percent) did not actually find the video helpful in comprehending the text.

Similar to Coniam (2001) and Gruba (1993), Londe (2009) compared ESL students on their listening test performance based on whether they were in an audio-only, a talking head video (i.e., a close-up of the professor's face), or a full body video (i.e., a video of the professor's full body as well as the board and some of the students in the classroom) group. Her findings showed no significant differences between the three groups, indicating that visual mode did not affect comprehension in either a positive or negative manner.

Finally, Suvorov (2008, 2009) investigated the impact of context images and videos on ESL listening performance as opposed to their performance on audio-only listening measures. While he found that no statistically significant difference in performance arose between the audio-only and context images sections of his listening test, he did find that the context videos actually seemed to have a negative effect on the test takers' performance. Thus the findings of this study combined with those of Coniam (2001), Gruba (1993), and Londe (2009) seem to give credence to claims that visuals have no significant impact on listening and that they may even have detrimental effects.

As is evident from the studies described here, the effects that visuals (specifically video-based visuals) have on listening comprehension are far from clear and appear at times to be contradictory. However, a number of explanations exist for these contradictory findings. For instance, while much of the research has been comparing test taker performance on audio-only and a visual-based listening test, the types of visuals being investigated differ. While many studies investigate the comparison between audio-only and video-based listening passages (Coniam, 2001; Gruba, 1993; Londe, 2009; Wagner, 2010a), some studies in L2 listening assessment focus on differences in performance between audio-only passages and still pictures (as is the case with Ginther's (2002) study) or all three visual conditions (Suvorov, 2008, 2009). In addition, not all studies are careful to make a distinction between context and content visuals, making it unclear which types of visuals were most prevalent in the videos they were showing to test takers. Therefore, as an area of study, the role of visuals in L2 listening comprehension is only starting to be understood and will require further investigations into this matter.

Furthermore, as Wagner (2010b) describes in his overview of studies investigating the impact of visuals on L2 listening comprehension, studies that are currently available can differ markedly in relation to the length and difficulty of the listening passages, the groups that are being tested, and the procedures that each study used. As such, it would seem that some of the inconsistencies and contradictory findings between the studies described here may be due to the fact that the types of listening texts and their length vary too much and that visuals may be more effective with certain text types or lengths but not others. Moreover, not all researchers have published their reliability statistics for the tests they created. For instance, while Brett (1997) found that there was a significant difference making those with visuals perform better, and Coniam (2001) found the opposite result, neither of these studies published their test reliability,

leading to uncertainty as to whether the results are adequate measures of comprehension or if error has led to the results that the researchers obtained.

In addition to these factors, issues related to the item types being used also call into question the results of these studies. Most studies may have had a multiple-choice component to their tests, but many also included other item types such as true and false questions, open-ended response questions, and fill-in-the-gap questions. Without having a standardized item type across every study, it is difficult to draw any certain conclusions for the efficacy of visuals on listening performance. Finally, related to test items, there have been limited studies thus far that have investigated the differential functioning of different items based on the presence or absence of visuals. The reason that some of these tests may be showing differing results may be that some items are more biased towards the use of visual information than others. For instance, Batty (2014) found that several of the test items in his listening measure were biased towards one visual format over another. Therefore, if some tests display items that have bias only towards the audio-only format, then there will obviously be no significant performance enhancement when visuals are displayed. Tests with items that can be enhanced by visual displays need to be developed to see if learners make use of information such as that which they may find in the lecture environment in the form of diagrams, pictures, or lecture slides with linguistic information provided on them. At present, based on the studies above, it appears that many tests used in research still rely heavily on prior conceptualizations of the listening comprehension construct and need to be adapted to encompass the construct that testing experts are trying to argue for if they want to adequately see what effects visuals will truly have on performance.

**Note-Taking and Listening Comprehension**

Research has shown that students take notes during lectures for a variety of reasons.

These reasons include helping them to stay awake, aiding in concentration, or helping them to pay attention to the aural input they are receiving (Ladas, 1980; Teng, 2011). Based on research conducted by Baker and Lombardi (1985) and Palkovitz and Lore (1980), it is clear that taking notes serves another purpose: to aid in the recall of information. Both of these studies found that native-speaking students who recorded the tested information in their notes were between two and seven times more likely to answer an item on a test or quiz correctly, a finding more recently corroborated by a study conducted by Asl and Kheirzadeh (2016) in which they found that L2 learners in listening and note taking groups outperformed listening-only groups on a listening test. This is somewhat surprising given the motor-processing limitations observed by Ladas (1980), which showed that students are only able to write notes down by hand at a rate of 20 words per minute, thus making it difficult for students to have elaborate notes. With notes playing such an important role in test performance, it is essential to fully consider the role that note-taking plays in the test taking process in terms of how the visuals presented on an assessment interact with note-taking practices, how individual factors related to the test taker affect note-taking practices, and how instruction in note taking and medium of note taking (i.e., handwritten versus typed) affect the quality of the notes and test performance.

Several studies related to both L1 and L2 academic note-taking have examined the role note taking plays in listening test performance. Hartley and Davies (1978) provide an excellent overview of 35 such studies examining both L1 and L2 note-takers in which they found that 17 of the studies reported that note takers performed better, 16 reported no difference in performance, and two reported that note taking seemed to interfere with test performance. However, while these results indicate an unclear effect of note-taking, they do state that many of these studies did not take student or lecturer differences or different visuals or handouts into

account. More recently, Hayati and Jalilifar (2009) conducted a follow-up study in which they found that when students are instructed how to take notes during a TOEFL practice test versus when they received no such instruction or were told not to take notes, they performed better on the test. Additionally, they noted that not all students received adequate instruction in note taking while listening, concluding that such practices were an acquired skill that would be essential in determining what from the input is important to note down.

In addition to the role that instruction plays on note taking and test performance, individual differences have also been found to play a significant role in the efficacy that note-taking plays in test performance. For instance, Asl and Kheirzadeh (2016) in their study of Iranian students in the EFL context found that working memory correlates significantly with listening comprehension regardless of whether students were allowed to take notes. These findings support previous findings by Dunkel et al. (1989) who found that both native and nonnative speakers of English with higher short-term memory capacity did better in their note taking. Moreover, nonverbal cues play an important role in signaling what to write down in one's notes (Piolat, Olive, & Kellogg, 2005). While individual differences may limit test taker's performance or note taking abilities, instruction on how to utilize these nonverbal cues may help to overcome potential deficits in working memory or to enhance short-term memory capacity by helping learners identify what cues signal important information (Hayati & Jalilifar, 2009). English (1982, 1985) explored such instruction by teaching visual cue interpretation to English-language learners and how this instruction affected academic listening comprehension. She found that instruction helped learners identify important information for their notes based on nonverbal cues and that fewer notes were taken, signifying more careful selection of information for inclusion in notes. Similar results were found in Cubilo and Winke's (2013) study in which

they found that learners took fewer notes during video-based tasks most likely due to inclusion of nonverbal information. Therefore, based on these studies, academic note taking appears to be a highly complex and learned skill that is dependent upon a number of factors, such as short-term and working memory, writing speed, the lecturer's nonverbal cues, and even fatigue or alertness, and it should be more carefully considered as a potential aspect of the listening construct given its potential effect on listening comprehension.

In addition to the cognitive factors related to note-taking, another important consideration that must be taken into account is the medium through which the listener takes notes. With the rapid development of technology, it is not unusual to see students in the classroom who are more accustomed to taking notes on their laptops rather than by hand because they can take notes at a faster speed and can more easily read and search through them later on (Kim, Turner, & Perez-Quinones, 2009). Thus, students who are more accustomed to the use of laptops for taking lecture notes may be put at a disadvantage if forced to take notes by hand. Although there are positives associated with the ability to type notes, a number of researchers have noted that a number of disadvantages exist with this method. For instance, Fink (2010) stated that the use of digital note taking reduced attentiveness when listening to a lecture. In addition, Stacy and Cain (2015) have noted that students who type their notes tend to write information verbatim instead of paraphrasing it in their own words. This would seem to explain results from studies such as one done by Muller and Oppenheimer (2014) which found that students who took notes by typing on laptops tended to have more difficulty remembering conceptual material than those who took handwritten notes while performing equally well when asked for factual information. Stacy and Cain's (2015) hypothesis would also support the findings of Piolat et al (2005) who found that those who typed notes performed worse on both conceptual and factual questions than

those who handwrote their notes. Based on these studies, it seems that handwritten notes may provide better retention of information leading to better performance. However, it is still unclear how different note taking conditions interact with different presentations of listening materials (i.e., still picture versus video) and whether changing one of these conditions on a listening assessment would lead to a markedly different outcome on test taker performance. Finally, studies on note taking that have investigated the difference between handwritten and typed note taking conditions primarily for native speakers of English. The present student attempts to expand this line of research to examine whether similar findings are present when the listening comprehension assessment is conducted in the test taker's L2.

**Methodological Framework**

Based on an overview of previous research, it is clear that further investigations need to be conducted to better understand the role that visuals and note-taking play in comprehension and test performance. While previous research has primarily focused on quantitative methods that measure overall improvements in listening assessment scores over time, the present study has sought to add to the relatively small sample of studies that have used mixed methods research designs as a means to better understand the underlying processes and factors that influence test takers as they are asked to take listening tests that utilize different input and note-taking methods. Below is an overview of the issues associated with mixed methods research designs and the role it has played thus far in the realm of L2 listening assessment research.

**Mixed methods research.** According to Leech and Onwuegbuzie (2009), the differences between the quantitative and qualitative paradigms of research resulted in a type of "paradigm war" (Gage, 1989) in which these two types of methodologies were at odds with each other until the 1960s. During this time, the mixing of these two approaches to research was introduced and

has recently come to be more common in a variety of fields such as education (Johnson & Onwuegbuzie, 2004), psychology (Waszak & Sines, 2003), and program evaluations (Greene, Caracelli, & Graham, 1989). The development of mixed methods research arose from the idea that concentrating solely on either quantitative or qualitative methodologies prevented the researcher from viewing the entire picture and that pieces of information vital to fully understanding certain phenomena were therefore missing. Campbell and Fiske (1959) introduced this idea by introducing the idea of "multiple operationalism," stating that more than one method should be used as a way to validate results because it ensures that the variance that researchers find is actually due to the trait being examined rather than the method itself. Webb, Campbell, Schwartz, and Sechrest (1966) further expanded on this idea of multiple operationalism by introducing the concept of triangulation, stating that "once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced." (p. 3) Denzin (1978) has stated that this triangulation can be conducted using either within-methods designs (multiple methods that are either all quantitative or all qualitative) or between-methods designs (multiple methods pulled from both paradigms). However, Denzin recommended the use of between-methods designs, arguing that within-methods triangulation will result in any inherent weakness of a particular paradigm manifesting itself whereas a between-methods triangulation would make it so that "the bias inherent in any particular data source, investigators, and particular method will be canceled out when used in conjunction with other data sources, investigators, and methods (p. 14).

Recent discussions of mixed methods research have added to these previous attempts to mix the two paradigms by attempting to more concisely define what it means to conduct research within this paradigm, making it possible for it develop into a more distinctive methodology

(Greene, 2008). In their analysis of this type of research, Johnson, Onwuegbuzie, and Turner (2007) analyzed a series of 19 definitions for mixed methods research they obtained from methodologists in the hopes of developing a clearly stated description of what mixed methodology actually is. Based on their analysis of the definitions received from the methodologists that they contacted, Johnson et al. (2007) developed the following general definition of mixed methods research:

> Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration (p. 123).

While this conceptualization of mixed methods research appears straightforward, further discussion has been required for fully determining the extent to which each research paradigm should be used and how they should be used in relation to each other for a researcher's methodology to truly be considered as mixed methods.

Johnson et al. (2007) addressed these concerns when they discussed their conceptualization of research methodology as a quantitative-qualitative continuum. In this description, it is evident that mixed methods research is subject to a variety of forms dependent upon the degree to which each of these paradigms is mixed with the other. On either end of the continuum is pure quantitative or pure qualitative in which only one research paradigm is used. However, in the middle of the continuum, it is clear that each of these paradigms can be mixed with the other to various degrees, thus creating "pure" mixed research in which quantitative and qualitative methods are given equal status, quantitative mixed methods in which the methods are

primarily quantitative, but qualitative data and approaches are included to add further detail to the data, and qualitative mixed methods in which the methods are primarily qualitative, but quantitative sources of data and approaches are included within the project.

In addition to identifying the ways in which methods can be mixed, discussions regarding the terminology of method sequencing have also arisen which have helped to frame and categorize mixed methods research in a more concrete manner. Teddlie and Tashakkori (2006) have contributed to this discussion by establishing and defining four different categories of mixed methods research design: concurrent, sequential, conversion, and fully integrated. As they define them, concurrent designs exist when two or more independent strands of quantitative and qualitative methods are used in conjunction with each other. In doing so, the researcher would collect data from each strand and synthesize them in order to better understand the issues being examined. In contrast, sequential designs would use quantitative and qualitative methods chronologically in which data is collected from one strand and the results of this analysis inform the research in determining new research questions and data collection for the following strand. In conversion designs, the researcher uses qualitative and quantitative strands during all stages in a study and this data is collected and analyzed in that strand before it is converted using the other strand for further analysis. Finally, fully integrated designs involve cross-talk between the two strands of analysis at every stage of research. Therefore, as the study proceeds along its path from conceptual stage to methodological and analytical stages, to inferences stages, to making meta-inferences, both quantitative and qualitative methods are being employed at each stage as a means to inform the researcher as to what questions to formulate, what analyses to conduct, and what conclusions to draw from the data. By establishing such terminology, Teddlie and Tashakkori (2006) have helped to establish what a mixed methods design truly looks like.

While these different approaches to mixing and sequencing methodologies exist, it is important to remember that not all approaches to combining quantitative and qualitative methods equate to a mixed methods framework. Brown (2014) discusses this issue, stating that it is not enough for a study to simply be outside of the "pure qualitative" or "pure quantitative" categories for it to be considered as mixed method. In his discussion, he argues that in order for a study to truly be mixed methods, it must use qualitative and quantitative methods in such a way that they complement each other. Therefore, if the researcher is using quantitative and qualitative methods concurrently or sequentially, but neither method interacts with the other in such a way so as to inform the researches interpretations of the data, then the research would be more accurately labeled as multi-method research rather than mixed method research.

Several issues arise with the development and use of mixed methods as a research paradigm. As would be expected, some paradigmatic purists have spoken out in the past. For instance, quantitative purists such as Nagel (1986) believed that social science research should be objective and that the research should remain detached and uninvolved with the objects of study. Conversely, qualitative purists such as Guba and Lincoln (1989) argue that multiple constructed realities exist and that it is impossible to make generalizations without taking into account time and context, nor do they believe it possible to be detached. While Tashakkori and Teddlie (2003) question whether such differences in philosophy can be reconciled, Johnson and Onwuegbuzie (2004) state that both traditions are valuable and that it is possible to draw from the strengths of both quantitative and qualitative positions. Regardless of the opinions that purists hold, there is no doubt that mixed methods research is becoming more and more common and has been increasingly utilized in the field of language assessment over the past decade.

**Assessment validation using mixed methods.** The new paradigm in validity mentioned earlier in this chapter, in which validity is seen as an argument based on evidence that either supports or opposes interpretations and uses of test scores, would seem to benefit from the use of mixed methods research design given the complementary strengths that the quantitative and qualitative paradigms bring (Onwuegbuzie & Johnson, 2006). Even though these two paradigms conceptualize validity differently, when they are combined, a complementary notion of validity is able to be obtained (Jang, Wagner, & Park, 2014). Indeed, this conceptualization allows for one to construct stronger inferences by embracing "argument alongside evidence, divergence alongside agreement, contextual understanding alongside causal explanations, and inclusion of the uses and action consequences of the inferences rendered" (Greene, 2011, p. 89). Thus, the inferences made on test scores in a mixed methods study go beyond merely making claims based on statistical inferences of a small sample of the overall population, but also allow for the inclusion of contextual explanations for performances, preventing situations in which validity claims may make the social and consequential responsibilities of test developers and researchers unclear (Davies, 1997).

Mixed methods as a research paradigm has recently become more common in the area of language assessment research, though it is not always explicitly named as such when used. In a review of recent assessment studies that have made use of this paradigm, Jang et al. (2014) found 35 studies in language assessment using mixed methods research for a variety of purposes. Of these purposes, the most common were triangulation (in which findings are cross-validated in order to offset biases different methods may possess), complementarity (in which results from one method are clarified, explained, and elaborated using results from another method), and

developmental (in which further exploration of data or further interpretations are guided by the preceding method).

Several studies examining the validity of various language assessments do an excellent job of illustrating how mixed methods research can be used to more effectively evaluate the validity of various types of language assessments. For instance, Anthony (2009) conducted a study in which he investigated the use of reflective, timed-essay responses that encouraged students to reflect on memories of experiences in their undergraduate studies that were most meaningful to them. In his study, he sought to investigate the construct and consequential validity of the writing prompts as well as the inference quality that they yielded. Anthony states that the mixed methods research design of his study was necessary given that neither a qualitative or quantitative paradigm was enough for fully addressing these validity issues, concluding that the use of mixed methods actually strengthened his overall validity argument. In addition, Harsch and Martin (2012) performed a mixed methods research study examining the validity of a revised rubric based on the Common European Framework proficiency scales. Their study involved a quantitative analysis of rater agreement and consistency, which was complemented by discussions with raters targeting items that had low agreement and consistency. This feedback from raters was used to inform them in ways to revise scale items that indicated poor quantitative characteristics, leading to a more valid scale.

Further studies have also been conducted by Jang (2005, 2009) that have utilized mixed methods research as a means to provide a more robust validity argument for different language assessments. In addition to extensive quantitative data in both of these studies, Jang (2005, 2009) also collected qualitative data in the form of think-aloud protocols, classroom observations, interviews, and surveys. In both cases, Jang concluded that her overall argument for the validity

of reading comprehension diagnosis within the Next Generation Test of English as a Foreign Language was strengthened. As more studies are conducted using mixed methods designs, it is becoming increasingly clear that the use of information from multiple paradigms that were previously relatively isolated from each other is a promising means for better analyzing the validity of assessment measures by providing both quantitative data analyzing test items and tasks, and qualitative information regarding the context and stakeholder views of the test itself.

**Mixed methods in second language listening assessment**. As mentioned in an earlier section of this chapter, research within the area of L2 listening assessment has primarily focused on how visuals affect L2 learners' performance on different listening tasks. Many of these studies have focused on making direct comparisons between scores across various conditions that altered the presentation of visuals in some obvious way, with most primarily comparing either a blank screen or a still picture to a version of the listening enhanced by video (Coniam, 2001; Suvorov, 2009, 2015; Wagner, 2010b). However, while such quantitative methods have been used frequently in examining the effects that video has on listening comprehension, few researchers have made use of qualitative methods (Purdy, 2010). More recently, there have been calls made for the use of more qualitative methods in L2 listening assessment research, with individuals proposing that think-aloud protocols (Wagner, 2007), retrospective verbal reports (Gruba, 2006), and interviews (Ockey, 2007) be used. However, even though there has been a wider call for the use of qualitative methods in L2 listening assessment, a use of qualitative methods in isolation in order to examine L2 listening performance is problematic due to the multidimensional nature of the act of listening (Bodie, Janusik, & Valikoski, 2008). Based on this fact, if studies in L2 listening assessment are to make use of qualitative methods, it would seem best to use a mixed methods approach to do so in order to better capture the different

aspects of the process that are occurring within the test taker.

Although a mixed methods approach would be useful in examining listening, little research in L2 listening assessment has actually adopted this methodological design. In fact, of the 35 studies investigated by Jang et al. (2014) in their review of the use of mixed methods in language assessment, they found that only one of the studies they collected (Lee & Winke, 2013) actually used a mixed methods design within the context of an L2 listening assessment. Instead, the majority of research in language assessment that has used mixed methodology has focused on its use as a tool in rating scale development. Thus, while calls for incorporating more qualitative methods have been made for L2 listening assessment, it appears that as of yet few have actually attempted to act on them. Without incorporating such methods in research targeting such a multifaceted skill, the knowledge obtained from studies will remain limited. Further studies need to examine the perceptions of students taking the tests in terms of how they perceive the videos they are presented with and their views on notetaking methods to make for more solid interpretations of data attempting to explain their performance and extrapolating this performance to the TLU domain. The present study attempts to use a quantitative-dominant mixed methods design in order to do this.

**Research Questions**

The present study addresses the gaps in the literature mentioned above by investigating the roles that visuals and note-taking medium play in listening comprehension. Specifically, the impact of these conditions on overall test scores, item performance, and performance on listening comprehension subskills are investigated. Additionally, the study seeks to contribute to the area of L2 listening assessment by using a mixed methods research design to answer the call for more research using such methodology in order to better understand student perceptions of test

conditions. Therefore, in order to accomplish these goals, answers for the following research questions within the construct and TLU domain of L2 academic listening comprehension are sought:

1. To what degree do participants' performances on listening tests vary when presented with listening input (i.e., video-based versus audio-only material) and note-taking (i.e., handwritten versus typed) conditions? To what degree do these conditions interact with each other?
2. How do item characteristics differ between the video-based and audio-only conditions? How do they differ between handwritten and typed note-taking conditions?
3. What is the extent to which visual support and note-taking conditions influence examinees' abilities to answer items testing them on different listening comprehension skills?
4. What perceptions and opinions do examinees have of the different conditions to which they are exposed and how do these perceptions and opinions and explanations shed light on the results from the previous questions?

Based on the research briefly reviewed above, it is believed that both visual and note-taking conditions will significantly influence both overall and item-centric test performance. In particular, based on other studies in which visuals played an important role in the overall listening comprehension (Sueyoshi & Hardison, 2005; Wagner, 2008), it is believed that the influence of visuals play a positive role in performance and that this role will be associated with visuals significantly influencing certain listening skills. Similar to the role that visuals play, it is also hypothesized that note-taking conditions will play a significant role, with typed note-taking

providing a positive significant effect on performance. In relation to item difficulty, it is believed that both of these conditions will lead to potentially lower difficulty scores for certain items.

Finally, based on previous research from Cubilo and Winke (2013) and Suvorov (2015), it is believed that test taker comments will reveal (among other things) that the visuals, while distracting, were very helpful for focusing on key details. This will be helpful in interpreting performance on different listening subskills as well as overall test performance.

# CHAPTER 3
# METHODOLOGY

This chapter describes the methodology used to collect data for this study. The chapter begins by providing a description of the participants of the study. Following this description, the chapter provides a detailed discussion of the materials and instruments that were used for data collection. Finally, the chapter provides a description of the research design and explains the procedures followed for test administration, survey collection, and data collection.

**Participants**

A total of 200 adult learners of English as a second language participated in this study. All students were studying within a university or community college setting in the United States, and all voluntarily participated in the study. Participants were either fully-matriculated students taking English for academic purposes classes in addition to the classes required by their major or they were in gateway programs that allowed them to complete intensive study of English for academic purposes prior to applying and matriculating in a four-year university. Table 3.1 provides a breakdown of the demographics of the participants. The examinees came from 22 different first language backgrounds including Chinese (40.5%), Korean (20%), Japanese (14.5%), Spanish (5.5%), Arabic (4.5%), Vietnamese (2.5%), Turkish (2.5%), Cantonese (1.5%), German (1.5%), and various other first language backgrounds represented by one or two participants as detailed in Table 3.1. Among the 200 participants, 53.5% were female and 46.5% were male with the majority of participants being undergraduate students (66.5%). Participants represented a range of ages (minimum = 18, maximum = 43) with the average age of the participants being around 21 years old ($M$ = 21.5). Participants were recruited from several

Table 3.1

*Participant Demographics*

| **L1** | **N** | **Percentage** |
|---|---|---|
| Chinese | 81 | 40.50% |
| Korean | 40 | 20% |
| Japanese | 29 | 14.50% |
| Spanish | 11 | 5.50% |
| Arabic | 9 | 4.50% |
| Turkish | 5 | 2.50% |
| Vietnamese | 5 | 2.50% |
| Cantonese | 3 | 1.50% |
| German | 3 | 1.50% |
| Hindi | 2 | 1.00% |
| Bulgarian | 1 | 0.50% |
| Farsi | 1 | 0.50% |
| Icelandic | 1 | 0.50% |
| Italian | 1 | 0.50% |
| Malayalam | 1 | 0.50% |
| Norwegian | 1 | 0.50% |
| Persian | 1 | 0.50% |
| Portuguese | 1 | 0.50% |
| Surigaonon | 1 | 0.50% |
| Thai | 1 | 0.50% |
| Urdu | 1 | 0.50% |
| Yapese | 1 | 0.50% |
| **Gender** | | |
| Male | 93 | 46.50% |
| Female | 107 | 53.50% |
| **Academic Status** | | |
| Undergraduate | 133 | 66.50% |
| Graduate | 13 | 6.50% |
| Other | 54 | 27% |

different proficiency levels (e.g., beginner, intermediate, and advanced) as determined by their respective programs.

## Materials and Instruments

**Academic listening test.** Two parallel forms of the academic listening test used in this study were developed in accordance with a series of specifications that were developed based on the model of test blueprint design put forth by Bachman and Palmer (2010). The fundamental information from this blueprint is provided in Table 3.2.

Lecture topics were chosen so that each test would include lectures that would represent topics from the humanities, the natural sciences, and the social sciences that would potentially be found in an introductory-level university course. Topics were selected and adapted from the University of Hawaii English placement test as well as from presentations on TED.com and from TOEFL test preparation materials. These topics and lectures were then adapted to incorporate additional content so that each lecture was of approximately the same length and the script was edited to ensure that the language was representative of an introductory-level university course so as to avoid issues with unnecessary use of jargon that it was not necessary for test takers to understand (See Appendix A for an example). A summary of the topics, the length of the lecture, and word counts of each lecture are found in Table 3.3.

Upon completion of developing scripts for the test, video recording of the lectures proceeded. Each lecture was accompanied by a series of eight to thirteen slides that provided a mixture of content visuals that were meant to facilitate the lecture (See Appendix B for an example). Slides were limited to the display of pictures and/or key words or phrases. Full sentences and paragraphs were not used and written words on slides were kept to a minimum. When recording, the same lecturer was used for all six videos in order to ensure that the dialect of American English used throughout the exam was consistent as well as the body language and overall rate of delivery. The lecturer also wore a lapel microphone in order to ensure that the

Table 3.2

*Test Blueprint for the Test of Listening Comprehension*

| Component | Description |
|---|---|
| **Purpose** | To obtain a measure of L2 learners' academic listening comprehension when presented with video-mediated lectures that contain primarily content visuals and when required to type or handwrite notes. |
| **Construct to be assessed** | The academic listening construct, defined as the ability to process and understand spoken material and content visuals from academic lectures representative of academic topics and vocabulary that require students to utilize skills in identifying the gist of a topic, recalling specific details from the lecture, making inferences about the lecture material, and understanding the attitude of the speaker toward the topic being discussed while utilizing effective note-taking strategies. |
| **Setting** | A computer lab equipped with the Internet and headphones |
| **Time allotment** | Approximately 45 minutes per test form |
| **Instructions** | You are about to be presented with a series of academic lectures on several different topics. Before listening, please adjust the volume on your headset.<br><br>The lectures in this listening test are meant to test you on your ability to understand spoken academic English. You will be presented with three lectures. Each lecture will be approximately 10 minutes long and will cover a topic taken from an introductory university lecture course. You will only hear each lecture once. You may not open the questions while listening to the lecture.<br><br>After watching the lecture, you will be presented with ten multiple-choice questions on the computer screen, for which you will choose the best answer. You will be given five minutes for each set of questions. Do not move on to the next page until you have completely answered the questions because you will not be able to go back.<br><br>You may take notes while listening to the lecture, but you will be instructed as to whether you may handwrite or type your notes on the computer screen. You may use these notes to help you answer the questions. |

(continued)

| Component | Description |
|---|---|
| **Instructions (cont.)** | The test lasts 45 minutes. Please click on the link to start the video lecture. |
| **Characteristics of input and expected response** | Each test form consists of three lectures. Lectures will either be videos or strictly audio-only dependent upon the condition that the test taker has been randomly assigned. A total of 30 multiple-choice questions will be on each test form. Questions will be delivered through Google Forms. Lectures for each form will represent topics related to the humanities, social sciences, and natural sciences. Videos will consist of content visuals (e.g., charts, pictures, words) and gestures meant to highlight spoken information. Each lecture is followed by a set of ten questions consisting of a stem with four possible options. Examinees are then expected to select the best answer from the four possibilities. |
| **Recording method** | Answers to the questions are recorded into google forms. These answers are then stored in a spreadsheet where they are converted into the simple letter option corresponding to the response (i.e., A, B, C, or D). These are then converted to scores by marking the correct answers as 1 and the incorrect answers as 0. |

Table 3.3

*Length and Run Time of Listening Topics*

| Topic | Word Count | Run Time (mm:ss) |
|---|---|---|
| Form A | | |
| *Language Policy* | 1171 | 08:08 |
| *Vaccine Development* | 1439 | 09:54 |
| *Dadaism* | 1343 | 09:23 |
| Form B | | |
| *Drake Equation* | 1153 | 09:01 |
| *Choice* | 1389 | 08:52 |
| *Morality* | 1406 | 10:41 |

sound quality was clear and consistent throughout the different videos. As she was lecturing, the lecturer was instructed to point to specific aspects of pictures to illustrate her points as she delivered the material. Examples of these images included pictures representing examples of art

a)


b)

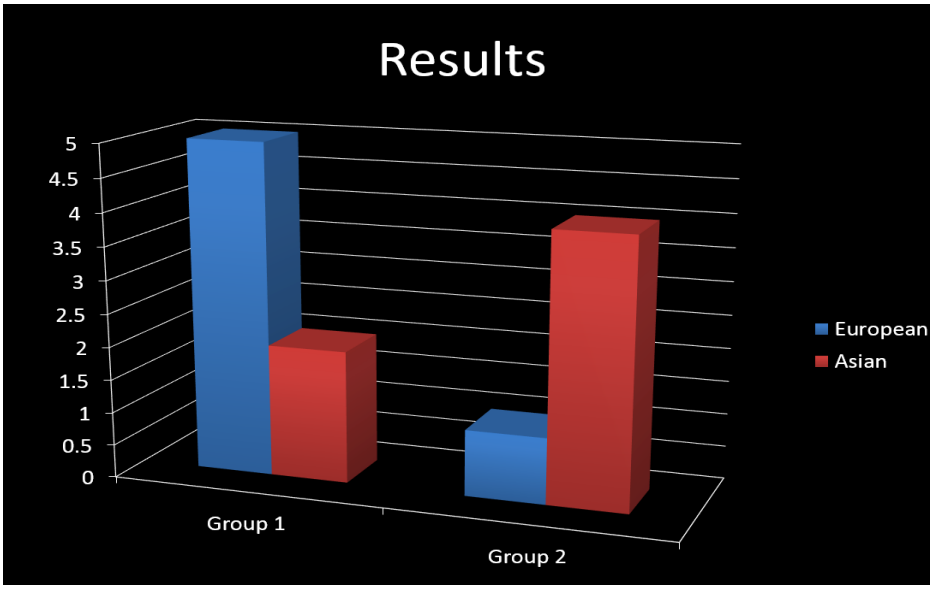



*Figure 3.1.* Sample lecture slides portraying content visuals from (a) an art history lecture and (b) a lecture on the study of choice.

from certain artistic movements (Dadaism, see Figure 3.1a) and bar graphs displaying research results (Choice, see Figure 3.1b) among others.

Following recording, three lectures were spliced together in a movie editor with each lecture separated from the other by five minutes of silence in which participants would answer

questions. This allowed the video to self-time the tasks by requiring test takers to move on to the next lecture when time was up for a particular section of the exam since it would automatically play the next lecture following the end of the five-minute response period. Two videos were created for each test form, one that consisted of video and one that consisted of only audio (which had been created by stripping video content from the original video and only using the audio file). Once these videos were completed, they were uploaded to *Youtube.com* for use on the test.

Ten multiple-choice items were developed to be paired with each lecture as specified in Table 3.4 based on item specification templates provided by Davidson (2001). Each of these items consisted of a stem with four possible options. Only one option was correct for each question and there was no partial scoring. The multiple-choice items were written to test a number of skills and, therefore, elicited answers targeting the test takers' abilities to identify the main idea of the listening passage, to identify supporting details of the passage, and to make inferences about the author's purposes for expressing certain ideas within the lecture. Question stems for each targeted skill were written based on question stems written for the same comprehension skills from other well-established tests such as the *TOEFL iBT* and the *IELTS*. A breakdown of how the question types were distributed between the different lecture topics is found in Table 3.5.

While an equal representation of each question type would have been ideal for the purposes of the data analysis procedures described below, this would not have been entirely representative of the actual TLU domain. Tests and assignments associated with lectures for introductory courses tend to be more heavily focused on recall of relevant details, with higher-order skills related to making inferences and understanding the speaker's attitude having a lower

Table 3.4

*Item Specifications*

| Component | Description |
|---|---|
| **General Description** | Examinees will be asked to answer sets of 10 four-option multiple-choice questions for each lecture they listen to. By answering these questions, examinees will be able to demonstrate their ability to listen to and comprehend material from an academic lecture setting. |
| **Question Types and Prompt Attributes** | Each item will test the construct of academic listening comprehension within the TLU domain of an academic lecture. Each item will consist of a question stem followed by four possible answers to the question stem with only one option being the correct answer. There will be four types of comprehension skills targeted for each lecture with each skill type being used at different frequencies to represent the relative frequency at which each skill may be used in an academic lecture for an introductory course. The comprehension skill types used are:<br>a. *Understanding the Gist*: This skill is targeted with questions asking the overall and very general idea of the lecture. These questions should not focus on any specific element of the lecture. Rather, they should target global understanding.<br>b. *Identifying Details*: This skill is targeted with questions asking the examinee to recall specific elements of the lecture. These questions could target specific concepts, definitions, or events described by the lecturer but should not require the examinee to do any more than recall information that the lecturer presented.<br>c. *Making Inferences*: This skill is targeted with questions asking examinees to make connections between different details described by the lecturer. In order for examinees to answer these questions, they will have to use their understanding of the details presented by the lecturer to draw connections between concepts and to come to novel conclusions.<br>d. *Identifying Speaker Attitudes*: This skill is targeted with questions asking examinees to |

(Continued)

| Component | Description |
|---|---|
| **Question Types and Prompt Attributes (cont.)** | draw conclusions about the attitude (e.g., skeptical, neutral, positive) the lecturer has towards the topic they are presenting. In order to answer these questions correctly, examinees will have to make conclusions based on the connotations of the words used by the lecturer as well as the prosodic elements and phrases that the speaker uses. |
| **Sample Items** | |
| *SI1: Understanding the Gist* | What is the main purpose of the lecture?<br>a. To describe the origins of vaccinations and the problems vaccine supporters experience today<br>b. To discuss what vaccines are and how they work.<br>c. To explain the history behind vaccination and how this has caused problems associated with vaccinations today<br>d. Explain the biological processes used to make vaccines and why people object to these processes. |
| *SI2: Identifying Details* | What conclusion does the speaker come to in relation to Americans' view of choice?<br>a. That more choice does not necessarily mean that people will be happier.<br>b. More choice is an important part of the American Dream and should not be reduced.<br>c. That, for Americans, limitless choice is necessary for greater happiness.<br>d. That the American idea of choice needs to be gradually introduced into other cultures. |
| *SI3: Making Inferences* | What is implied in the lecture about the philosophy of the Dada movement?<br>a. It was not taken seriously by other artists.<br>b. It varied from one country to another.<br>c. It challenged people's concept of what art is.<br>d. It was based on a realistic style of art. |

(Continued)

| Component | Description |
|---|---|
| **Sample Items (cont.)** | |
| *SI4: Identifying Speaker Attitudes* | What word would best describe the speakers' opinion towards the possibility of extraterrestrial life existing?<br>a. Critical<br>b. Neutral<br>c. Hopeful<br>d. Uncertain |

Table 3.5
*Question Type Number by Test Form and Topic and Question Stem Examples*

| | Form A | | | Form B | | |
|---|---|---|---|---|---|---|
| **Question Type** | **Drake Equation** | **Choice** | **Morality** | **Dadaism** | **Language Policy** | **Vaccination** |
| Gist | 1 | 1 | 1 | 1 | 1 | 1 |
| Detail | 7 | 6 | 6 | 8 | 7 | 6 |
| Attitudes | 1 | 1 | 0 | 0 | 1 | 1 |
| Inference | 1 | 2 | 3 | 1 | 1 | 2 |

prevalence (Alderson, 1990). This is most likely due to the pressure they put on processing speeds of examinees (who are already under greater pressure from having to process information in their L2), the time limit constraints they are under, and the fact that they have not had previous exposure to the material. Additionally, while research has shown that focusing on higher-order skills in tests is beneficial (Jensen, McDaniel, Woodard, & Kummer, 2014), these benefits only present themselves when students know ahead of time that such skills will be assessed and have time to study for tests testing these skills, showing that having an equal focus on such skills in testing situations such as the one in this study would not be appropriate since test takers do not have time to study the material. Finally, analyzing other large-scale listening comprehension tests exhibited similar characteristics supported the decision in determining relative frequencies for test items. However, even though this method of item development led to an unequal distribution in question-type frequencies, the analyses conducted for question 3 (described

below) are believed to exhibit results that are trustworthy, as explained in the analysis provided in the next chapter.

Once all questions were created, tests were assembled using *Google Forms*. Web links to test videos were provided to students via the first page of instructions for the exam. Due to technical limitations, lecture videos and audio had to be displayed using the *YouTube* page while simultaneously projecting the *Google Form* on the other side of the screen. Questions were presented to participants in sets, with all questions for a particular lecture being present on a single page. Therefore, the test was developed to present test takers with three ten-item sets of questions that were found after each lecture. The questions in a set were not available to the participants until the lecture for a given set had finished.

**Post-test questionnaire.** A post-test questionnaire (Appendix C) was created using *Google Forms* to collect demographic information and to ask learners about their opinions in relation to the note-taking and video conditions on the exam. The questionnaire asked participants for their age, gender, years of English study, native language, student status, and test preparation experience. In addition, the survey contained several open-ended questions asking them to discuss their typing ability, their preferences for note-taking medium, their preferences for listening medium, what they remembered focusing on in the lectures, and whether they were familiar with any of the topics.

## Procedures

**Piloting the academic listening test.** The listening test was first piloted with 20 volunteers with varying language proficiency levels. During the pilot, several issues with test administration were observed, leading to revision of procedures. For example, it was found that it was easy for test takers to accidentally skip ahead to the question set prior to the lecture

finishing. Future administrations limited this possibility as much as possible by instructing test takers to place the mouse behind the computer monitor until it was time to answer questions so that they would not be tempted to click on certain links before they were supposed to do so. In addition, it was determined that there should be a limit of 10 participants per administration. This was due to Internet speed limitations in some computer labs and the difficulty of monitoring larger groups of participants.

Piloting also provided data about the multiple-choice items and parallel test forms. Using the data collected from the participants, reliability analysis of the two test forms, item facility, and distractor analysis was conducted. Reliability analysis yielded a relatively high internal consistency for observed scores on both test form A (Cronbach's alpha = 0.811) and test form B (Cronbach's alpha = 0.803). Item facility and distractor analysis revealed that items were mostly behaving in the way they were intended. However, three items had either distractors that were never chosen (Item 4 from Vaccines and Item 10 from Dadaism) or distractors that proved to prevent the correct answer from being chosen at all (Item 3 from Choice). As a result, minor revisions were made to these items, and the use of the test was made operational.

**Test administration.** Participants were asked to sign up for and attend testing sessions in computer lab spaces that were reserved for the test. Groups of ten participants were scheduled for each testing date. Upon arrival to the computer lab, the consent form for the study was explained to the volunteers, who were asked to sign it if they agreed to continue participating. The format of the test was then explained to the group in relation to how the listening passages would be presented on the screen, note-taking restrictions for each form of the test, how questions would be displayed on the screen, and how they would be able to assign their answers to each of the multiple-choice questions. A sample item was given at the beginning of the exam prior to any

listening passage to ensure that all participants understood how to input their answers to questions. Once the test administrator had finished providing instructions to the participants, any remaining questions were answered and the participants were instructed to test their audio using a sample *YouTube* video already loaded on the computer. When participants had set their headphone volume, they were instructed to begin the first set of listening passages. Each listening passage was roughly 10 minutes in length and was followed by 5 minutes to answer 10 questions. Participants were able to take notes while listening (typed or handwritten, depending on the condition they were place in) and could refer to these notes while answering the questions for that listening. While listening, participants were presented with the video display on one half of the screen. The other half of the screen was then used for note-taking in a word processing document when the participants were required to type their notes (see Figure 3.2 for an example of what the screen looked like). When a lecture finished, students were instructed on the screen to click on the test tab in the Internet browser, which opened the *Google Form* containing the multiple-choice questions. When they completed the 10 questions, they clicked the next button on the screen, which brought up a page instructing them to click the tab back to the lecture recording where they would wait for the next lecture to start. The screen set-up was the same for all conditions in terms of the Internet browser window size. However, when participants were asked to take notes by hand, there was no word processing document on the screen for them to type into as in Figure 3.2. Instead, this space was simply an empty space with a plain blue desktop background.

Upon completion of the first three lectures, students were brought to a page telling them to stop and inform the test administrator that they had finished the first test. Upon notification, the test administrator gave the participants a short break while setting up the second half of the

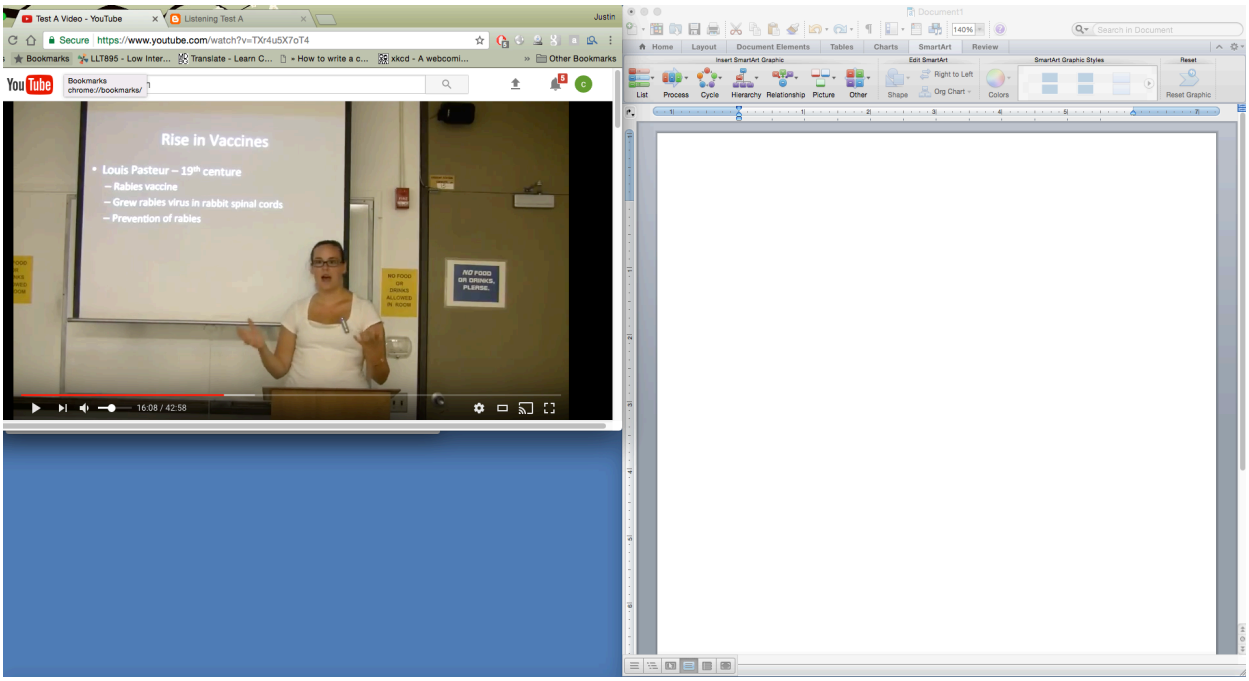



*Figure 3.2*. Example of screen set-up for examinee in the video and typed note-taking conditions.

test. Once all participants returned, the test administrator notified the participants of changes to note-taking and video conditions they would encounter in the second half of the test, answered any questions, and then started the exam, which ran in the same manner as the first half. When completed, participants pressed the submit button on the screen in order to record their results and were then instructed to wait until time was up. Following submission of the test, students were then asked to fill in the post-test questionnaire after which they were finished with the study.

**Study design.** In order to examine the effects of note-taking medium and audio-visual input on listening comprehension scores, the study follows a crossed split-plot design in which the medium of note-taking and the presence or absence of visual input with the listening passage served as the independent variables. Each participant was asked to take a test in which they were required to both handwrite and type notes and were exposed to a set of listening passages that

Table 3.6.

*Possible Experimental Conditions to Which Participants Could be Assigned.*

| | Test 1 | | | Test 2 | | |
|---|---|---|---|---|---|---|
| **Condition** | **Form** | **Input** | **Notes** | **Form** | **Input** | **Notes** |
| 1 | Form A | Video | Typed | Form B | Audio | Handwritten |
| 2 | Form A | Audio | Handwritten | Form B | Video | Typed |
| 3 | Form A | Video | Handwritten | Form B | Audio | Typed |
| 4 | Form A | Audio | Typed | Form B | Video | Handwritten |
| 5 | Form B | Audio | Handwritten | Form A | Video | Typed |
| 6 | Form B | Video | Typed | Form A | Audio | Handwritten |
| 7 | Form B | Audio | Typed | Form A | Video | Handwritten |
| 8 | Form B | Video | Handwritten | Form A | Audio | Typed |

either had visual input or did not. In total, there were eight different conditions in which participants were placed, which are summarized in Table 3.6.

**Data Analysis**

For the quantitative portion of the analysis in this study several different methods were used to analyze the data. For research question 1, a 2x2 between groups ANOVA was conducted using IBM SPSS in which the scores were compared in relation to the two independent variables of visual input and note-taking. However, in order to control for any effects the test forms may have on the comparison conducted in the ANOVA (and to essentially equate the two forms), the procedure illustrated in Figure 3.3 was used. In this procedure, half of the scores were randomly selected from each block by first grouping scores according to examination conditions for each form. This was then followed by randomly selecting 25 scores from each of the "audio-only" blocks. The remaining 25 scores for the audio-only selection were not used. Rather, these remaining participants had their "video" scores used in the analysis. Thus, for each form there were 50 audio-only and 50 video test scores represented. Additionally, because these selections were based on the conditions outlined in the boxes in Figure 3.3, note-taking conditions were

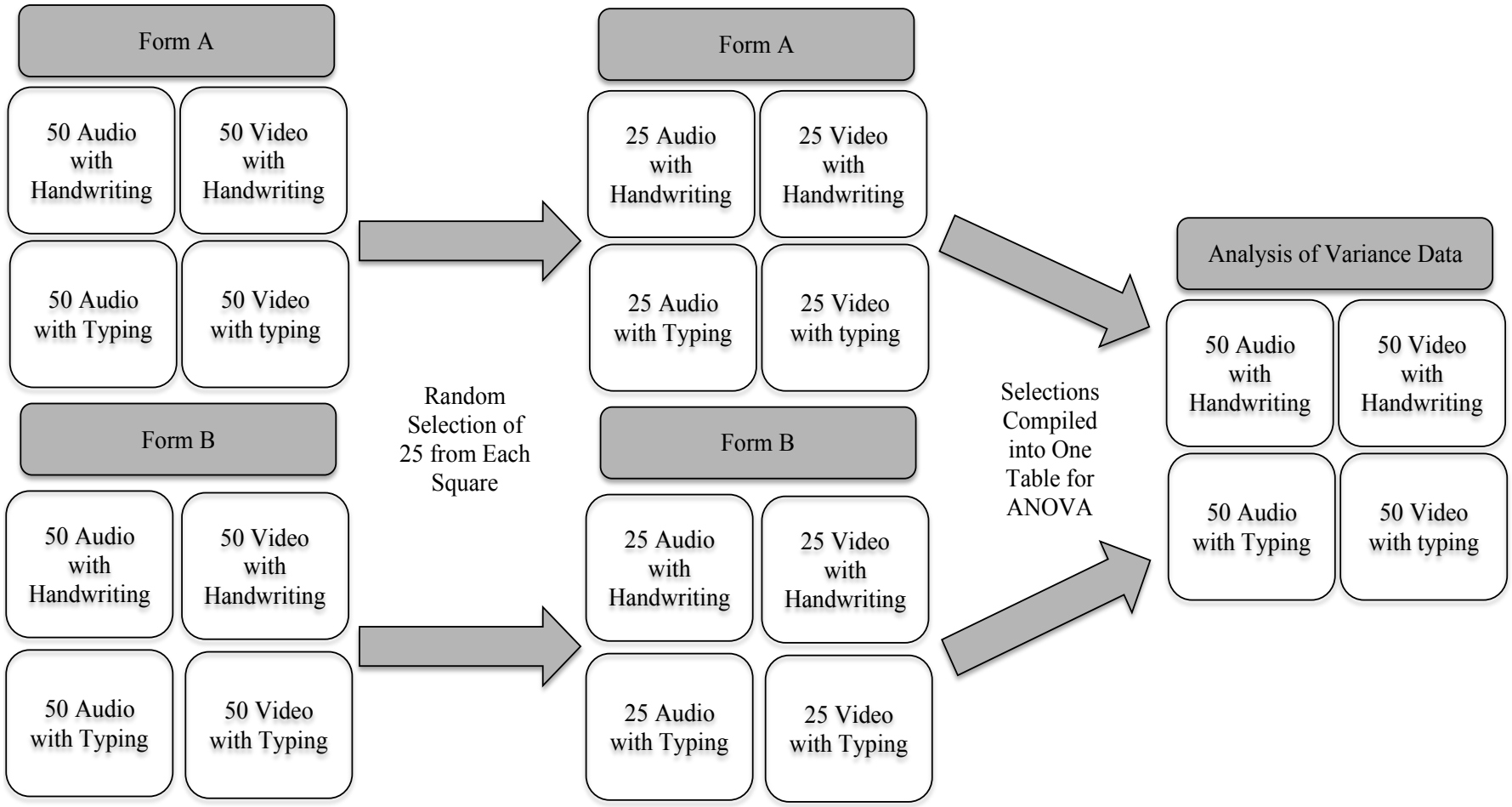



*Figure 3.3.* Data selection procedure for ANOVA analysis

similarly represented for each form. This resulted in a total of 50 of the 200 total participants being placed randomly in each condition, which allowed for full analysis of the data. As with many previous studies, this analysis focused on the differences in scores based on the effects of the listening conditions (i.e., video versus audio-only) and note-taking condition (handwritten versus typed) and the interactions that existed between the two conditions and focused on composite score differences to get a general sense of the impact each condition had on test scores. While this particular analysis may or may not show overall effects that different administration conditions have on performance, it does not show the micro-level effects seen on different items or test subscales, which is provided by the next two research questions.

Whereas research question 1 was focused on the more general effects of the conditions on test scores, questions 2 and 3 were focused on the specific effects related to the impact that visual and note-taking conditions had on item performance as well as comprehension subskill performance. In order to examine the differential functioning of the items between the different conditions targeted by research question 2, a many-facet Rasch measurement (MFRM) was conducted using the computer program FACETS, version 3.71/4 (Linacre, 2014). MFRM is a probabilistic model that enables one to plot items and examinees along a similar scale for comparison using an interval scale. The advantage of this method is that it ensures an equal distance between any set of data points represents an equal difference in person ability or item difficulty, facilitating interpretation that other classical test statistics do not necessarily provide (Bond & Fox, 2007). In order to do this, MFRM uses a logit interval scale to compare person ability and the model fit of an item. In essence, Rasch modeling allows one to answer the question of whether an examinee at a particular ability level on a particular item in a particular

condition is likely to succeed. In order to run data analysis using Rasch analysis, data were entered into a specification file, which was then run in FACETS, producing a variable map that allows the viewer to directly compare test facets (Eckes, 2009). Therefore, this allowed the direct comparison of the effects that audio-visual format and note-taking format on the items. For the data in this study, a four-facet model was run with the facets being examinees, items, note-taking format, and audio-visual format. The model for MFRM or dichotomous items is expressed as:

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

where $P_{nij}$ is equal to the probablility of examinee $n$ with ability $B_n$ succeeding on item $i$ with difficulty level $D_i$ in condition $j$ with difficulty level $C_j$. In the model for this study, the audio-only and handwritten note conditions were anchored at zero logits. The reason for this was that these are the traditional test formats and it would allow the new test formats (video listening and typed note-taking) to be compared more easily on the Wright map produced by FACETS.

Finally, in order to examine the relative impact that note-taking and audio-video conditions had on question type, a path analysis was run using MPlus, version 7.3 (Muthen & Muthen, 2014). Subscores were obtained for each question type under examination and were then submitted to an initial model in which all item types were connected to the independent variables of video and typing by a path coefficient. Three separate models were run in which forms A and B were examined separately followed by a model consisting of pooled results from both forms. Based on the initial model results, the model was revised and run again with the same three sets of data. By doing this, it was possible to see how the listening subskills were being influenced by each of the experimental conditions.

The qualitative aspect of this study relied on responses to open-ended survey questions that were analyzed as a means to shed light on how participants' perceptions of different

conditions and how their preferences for certain conditions may have related to performance on the test. Responses for survey items related to preference for audio-video and note-taking formats were coded and mapped to participants' scores to provide possible explanations for potentially better performance in one condition over another.

Once survey data was collected, inductive analysis of the responses was conducted to determine the different codes that would be used to classify responses into different categories. For each question, responses were analyzed to determine the preferences expressed and the common reasons provided by participants for a given preference. Based on this list of reasons, a group of codes and thematic categories was developed based on these reasons that could then be applied to each response. For this analysis, codes are defined as the preference categories expressed by the participants (e.g., when asked whether they preferred audio or video, one preference code would be "audio" and one would be "video") while thematic categories are the categories associated with the reasons for the different preference codes (e.g., if they preferred video, the reason provided for this preference was categorized under a certain thematic category). The process of developing these codes and themes involved first determining which preference code a response would be assigned to. This proved to be rather straightforward since all but 9 responses from participants explicitly stated the preference the questions sought. Once codes were assigned to the responses, a second analysis of responses was performed to develop the list of thematic categories accompanying each code for classifying reasons provided by participants. The reasons were analyzed to determine thematic categories and each list was then further examined to condense those thematic categories that overlapped. As an illustration, if the question asked whether the participant preferred audio-only listening passages or video-mediated listening passages, the response as a unit was assigned a preference code first (i.e., audio, video,

both, or off-topic). Once the preference was classified, themes were developed for the specific preference based on the secondary analysis specifically examining the reasons provided for the preference the participants expressed, and the response was placed into the appropriate thematic category. Each of these responses was treated as a single unit, but each unit could potentially have multiple thematic categories assigned to it if the reasons encompassed more than one thematic category (though this was not very common). Thematic category assignments were then tallied and representative responses from each of them were used to provide discussion of the perceptions that participants had towards each testing condition.

In order to ensure reliability of codes and themes applied to responses, the researcher asked an additional coder to code 10 percent of the data using the developed list of codes and thematic categories to calculate inter-coder reliability. Asking a second coder to code 10 percent of the data seems to be common practice (Brown, 2001b; Lee & Winke, 2013). Once the second coder had completed the task, both coders discussed disagreements and changed codes where agreements were ultimately reached. After this discussion, intercoder reliability was then calculated in NVivo 10 (QSR International, 2014) in order to obtain Cohen's kappa values for each of the responses scored by both coders. These values were then averaged together to obtain the overall reliability ($\kappa$ = .89) which was only slightly higher than the initial reliability prior to discussion ($\kappa$ = .87), signaling that overall reliability was rather consistent between coders even prior to discussion. Based on Plonsky and Derrick's (2016) meta-analysis of reliability indicators in applied linguistics, this kappa value is acceptable and comparable to what others have found using similar procedures. Additionally, both of these values fall squarely within the recommended range of 0.85-0.90 provided by qualitative researchers such as Saladaña (2009). Therefore, the coding results were deemed to be acceptable overall. These qualitative data were

used to explain unanticipated results in the quantitative data and to consider the participants' perceptions of the different conditions and how these perceptions provide a more robust explanation for test performance results. In particular, explanations of why they preferred certain conditions or how they felt certain conditions affected their performance and/or behavior were useful for fully understanding the effects that the changes in such conditions might have on the overall validity of the test.

# CHAPTER 4
# RESULTS

This chapter provides results as they relate to each of the research questions outlined in the methodological framework chapter. The chapter opens by first discussing the overall descriptive statistics and classical test statistics to provide an analysis of the appropriateness of the test item performances. Following this overview, the chapter then proceeds to provide an overview of statistical analyses as they relate to each of the five research questions. All research questions were answered by primary analysis of the test data across the various conditions that participants were exposed to. Additionally, coded data from the post-test questionnaire was analyzed in connection test performances to better understand how preference relates to performance in the different conditions. Additional qualitative analysis related to open-ended explanations and perceptions described by participants in the survey will be provided in chapter 5 when offering further explanation of the study's results.

**Descriptive and Classical Item Statistics**

Descriptive and classical item statistics were obtained for test takers and items for the two forms of the test as a whole and for the different conditions. Descriptive statistics are displayed in Table 4.1 for the two test forms. Examinee scores for form A of the exam ranged from 1 to 30 and from 0 to 30 for form B. The mean score for form A was 15.22 ($SD$ = 6.22) with a median of 15.00, and the mean score for form B was 15.29 ($SD$ = 6.28) with a median of 14.00. The data for both forms were slightly positively skewed with form A exhibiting a skewness value of 0.19 and form B exhibiting a skewness of 0.12. This indicates that more examinees were clustering to the left of the mean when compared to an ideal normal distribution for both test forms, suggesting that the tests was slightly difficult for test takers. Kurtosis values for both forms were

Table 4.1

*Descriptive Statistics of Test Forms A and B*

| | **Form A** | **Form B** |
|---|---|---|
| Mean | 15.22 | 15.29 |
| *N* | 200 | 200 |
| Median | 15 | 14 |
| Mode | 17 | 14 |
| Range | 29 | 30 |
| *SD* | 6.22 | 6.28 |
| Kurtosis | -0.23 | -0.42 |
| *SEK* | 0.34 | 0.34 |
| Skewness | 0.19 | 0.12 |
| *SES* | 0.17 | 0.17 |

negative, with form A exhibiting a value of -0.226 and form B exhibiting a value of -0.423. Thus, the distribution for both forms was leptokurtic, indicating a slightly higher peak around the mean than what would be found with an ideal normal distribution. Overall, as is seen in Figure 4.1 and 4.2, the score distributions for both forms are relatively normal overall.

Item statistics were then calculated for each form, with items being pooled together for all conditions (audio/video, handwritten/typed). These item statistics included both item facility calculations and item discrimination indices. When calculating item facility, the facility value is the result of all examinees' correct responses to a single item over total number of responses. A higher value indicates an easier value, while a lower value indicates a more difficult item. Ideal values falling between 0.30 and 0.70. Item discrimination indices were obtained using point biserial correlations using the equation:

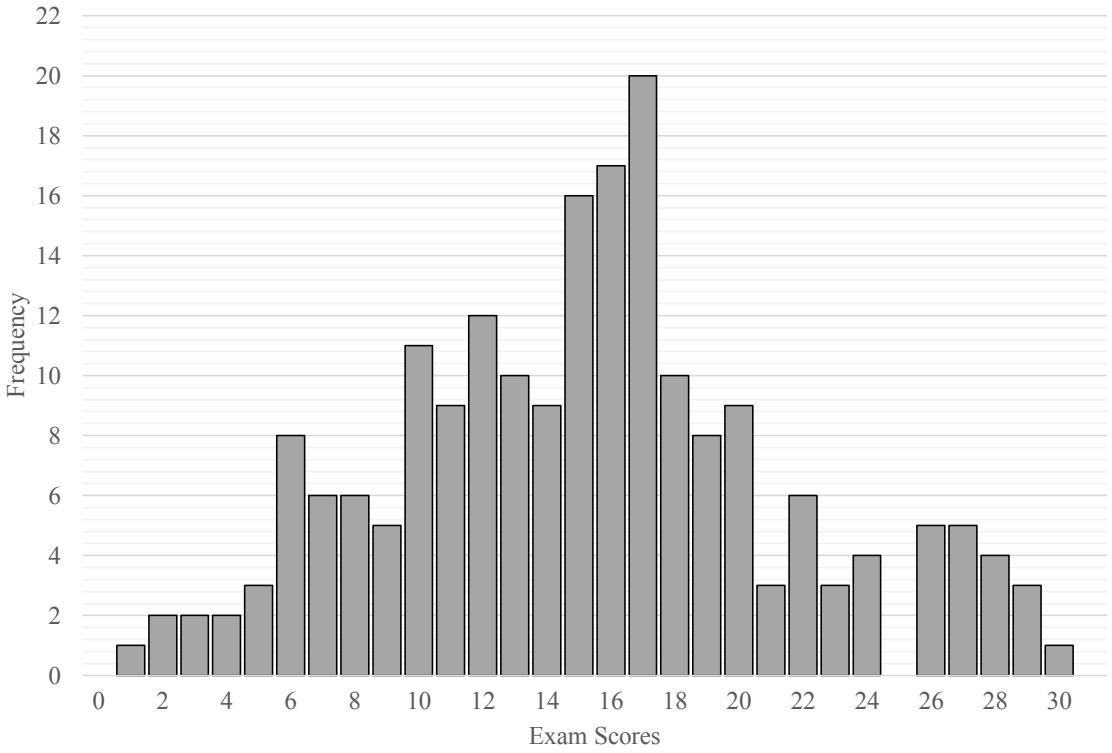



*Figure 4.1.* Score distribution for form A of listening test.

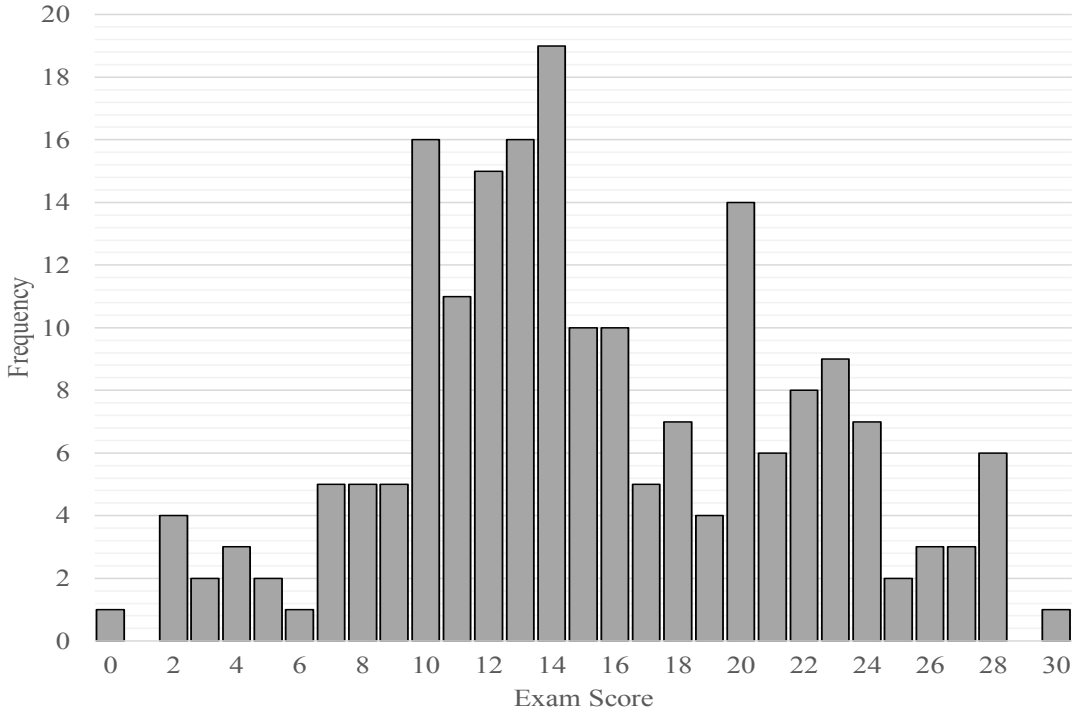



*Figure 4.2.* Score distribution for form B of listening test.

$r_{pbi} = ((M_p - M_q)/S_t)*(\sqrt{pq})$

$r_{pbi}$ = point biserial correlation

$M_p$ = mean test score for those who answered the item correctly

$M_q$ = mean test score for those who answered the item incorrectly

$S_t$ = standard deviation of all test scores

$p$ = proportion of examinees who answered the item correctly

$q$ = proportion of examinees who answered the item incorrectly (Eckes, 2009)

The summarized results for both of these analyses by test form can be seen in Table 4.2 (The full results are can be found in Appendix D). Item facilities for form A ranged from 0.305 to 0.775 ($M$ = 0.507, $SD$ = 0.113) while item facilities for form B ranged from 0.23 to 0.72 ($M$ = 0.510, $SD$ = 0.136). The most difficult question was item 28 on form B, which asked examinees to compare the philosophers Plato and Immanuel Kant's views on morality. The easiest question was item 4 on form A, which asked examinees to recall the approximate percentage of how many people in India speak Hindi.

Table 4.2.

*Summary of Classical Item Statistics by Test Form*

| **Test Form** | **Average *IF*(*SD*)** | **Average $r_{pbi}$(*SD*)** |
|---|---|---|
| Form A | 0.51 (0.11) | 0.42 (0.06) |
| Form B | 0.51 (0.14) | 0.43 (0.09) |

The point biserial correlation coefficient is used to investigate the degree to which a nominal and interval scale are related (Brown, 2001a). In this study, the purpose of using this is to investigate the degree to which dichotomously scored multiple choice item, which function as nominal scales where 1 indicates a correct answer and 0 indicates an incorrect answer, is related to overall scores on the test form in order to better estimate the discrimination of the item (i.e., the ability of the item to differentiate between high and lower performing examinees). The

higher the relationship between the item and the overall score, the higher the $r_{pbi}$ coefficient and the better the overall discrimination of the item. Point biserial correlation values for items on both test forms showed a range of 0.25 to 0.55 for form A and a range of 0.21 – 0.57 for form B. While Hatch and Lazaraton (1991) suggest a cutoff value of 0.20 or above to indicate good item discrimination, Brown (2005) states that a value of 0.40 or above is ideal to ensure proper discrimination. Therefore, the latter value was determined to be a guide to indicate items with excellent discrimination with the former indicating acceptable discrimination. Since all items fell at or above 0.20 with most falling above the 0.40 mark, it was determined that items on the two forms (ignoring different delivery conditions) were sufficiently discriminating between the more and less advanced listeners in the participant group.

Classical item statistics were also examined for each of the conditions, which can be found in Table 4.3 below. Forms A and B showed some similarity between conditions for item functioning. However, it is interesting to note that there were some differences in item facility and discrimination across the conditions. In particular, while Form A remained rather constant in regards to item facility, there is clearly some variation in item discrimination across the different

Table 4.3.

*Summary of Classical Item Statistics by Condition*

| | Form A | | Form B | |
|---|---|---|---|---|
| **Condition** | **Average *IF* (*SD*)** | **Average $r_{pbi}$ (*SD*)** | **Average *IF* (*SD*)** | **Average $r_{pbi}$ (*SD*)** |
| **Input** | | | | |
| *Video* | 0.50 (0.11) | 0.40 (0.07) | 0.56 (0.15) | 0.42 (0.11) |
| *Audio* | 0.51 (0.13) | 0.45 (0.10) | 0.46 (0.13) | 0.43 (0.09) |
| **Note-taking** | | | | |
| *Handwriting* | 0.51 (0.13) | 0.36 (0.09) | 0.48 (0.13) | 0.44 (0.12) |
| *Typed* | 0.51 (0.10) | 0.48 (0.07) | 0.54 (0.14) | 0.42 (0.07) |

conditions. The opposite can be said for form B, where the discrimination was rather constant, but facility was a bit more varied across conditions. This could unfortunately not be examined further since running an ANOVA on these values would violate the assumption of independence of observations, but could a possible future point of examination.

Reliability analyses using Cronbach's alpha were also conducted on the test for the forms as a whole and for each of the different conditions that each form was used under (Table 4.4). Reliability values ranged from 0.764 when handwritten notes were required for form A to 0.89 when typing was required on form A, with other values falling at or above 0.80. This indicates that forms and conditions were all internally consistent overall.

Table 4.4

*Cronbach's Alpha Values for Test Forms A and B*

| | **Form A** | **Form B** |
|---|---|---|
| Input | | |
| *Video* | 0.82 | 0.84 |
| *Audio* | 0.86 | 0.85 |
| Note-taking | | |
| *Handwritten* | 0.76 | 0.86 |
| *Typed* | 0.89 | 0.84 |
| Total | 0.84 | 0.85 |

**Research Question 1**

Question 1 asked how overall performance on listening tests varied between video-mediated listening passages and audio-only listening passages as well as between handwritten and typed note-taking conditions. It then continued along this line of inquiry to ask how these factors interact with each other. In order to answer these questions, participants were separated into four groups as described in Chapter 3 and a two-way analysis of variance (ANOVA) was conducted. The results of which can be seen in Table 4.5.

Table 4.5

*Analysis of Variance Results*

| **Source** | ***df*** | ***SS*** | ***MS*** | ***F*** | ***p*** | $\eta^2$ | **Power** |
|---|---|---|---|---|---|---|---|
| Input | 1 | 68.45 | 68.45 | 1.79 | 0.18 | 0.01 | 0.27 |
| Note-taking | 1 | 73.21 | 73.21 | 1.91 | 0.17 | 0.01 | 0.28 |
| Input*Note-taking | 1 | 28.13 | 28.13 | 0.75 | 0.39 | 0.004 | .14 |
| Error | 196 | 7503.78 | 38.29 | | | | |
| Total | 199 | 7673.56 | | | | | |

The results of the two-way ANOVA showed that while the mean for scores in the video-mediated listening ($M$ = 15.75, $SD$ = 6.3) and typed note-taking ($M$ = 15.77, $SD$ = 5.99) conditions were slightly higher than the means in the audio-only listening ($M$ = 14.58, $SD$ = 6.08) and handwritten note-taking ($M$ = 14.56, $SD$ = 6.39) conditions, the difference was not statistically significant either for the main effect of listening input condition ($F$(1,196) = 1.788, $p$ = .183), or the main effect of note-taking condition ($F$(1,196) = 1.912, $p$ = .168). Results were also not significant for the interaction between the two conditions ($F$(1,196) = 0.735, $p$ = .392). The effect size obtained from this test also indicates that the difference between each of the means was quite small with the highest partial eta squared value being 0.01 for the main effect of note-taking. Therefore, the results show that neither visual input, nor note-taking medium, nor the two working in tandem had a statistically significant effect on L2 test-takers' overall listening test performance. However, it should be noted that the overall power of the results was rather small, indicating that further examination with a larger sample size is necessary to conclusively make any statements regarding the actual effect the different conditions have on test-takers' performance. While these results indicate little effect on the overall test scores on listening tests for participants, it does not reveal the smaller scale effects that these different conditions had on scores. In order to further investigate this and more fully consider the impacts that visual and

note-taking conditions had on individuals' listening comprehension, further analysis was conducted across conditions on item performance and the listening comprehension subskills.

**Research Question 2**

Question 2 sought to get a more in-depth look of learner performance across conditions by investigating the item characteristics based on the condition under which they were presented to the examinee. In order to investigate this question, a many facet Rasch measurement (MFRM) analysis was conducted on the data. The variable map for the output is presented in Figure 4.3.

The first column of the variable map is the measure column which shows the equal-interval log odds scale ranging from -3 at the bottom of the figure to +3 at the top of the figure. The next column displays examinee ability as they relate to overall performance on the exam. The next two columns show input difficulty (audio-only versus video-based) for listening passages and note-taking difficulty (handwritten versus typed). In these columns, the traditional testing formats of handwritten and audio-only are anchored at zero while the more experimental formats of typed and video-based are allowed to float freely to provide a clearer relative difficulty for the conditions. Finally, the last column consists of item difficulty estimates. Each of these columns is plotted on the logit scale at the left of the figure. In relation to the logit values, an examinee that is plotted higher up on the logit scale is more able while an item plotted higher on the logit scale is considered more difficult than those below it. Linacre (2014) explains the relationship between these two facets by stating that if an examinee and item stand at the same logit value, the examinee will have a 50% likelihood of answering the item correctly. In addition, if the item value is 1.1 logits less difficult than an examinee's ability, then the likelihood of answering the item correctly for that person increases to 75%. Based on these assumptions, a higher placement on a variable map such as the one seen in Figure. 4.3 can be

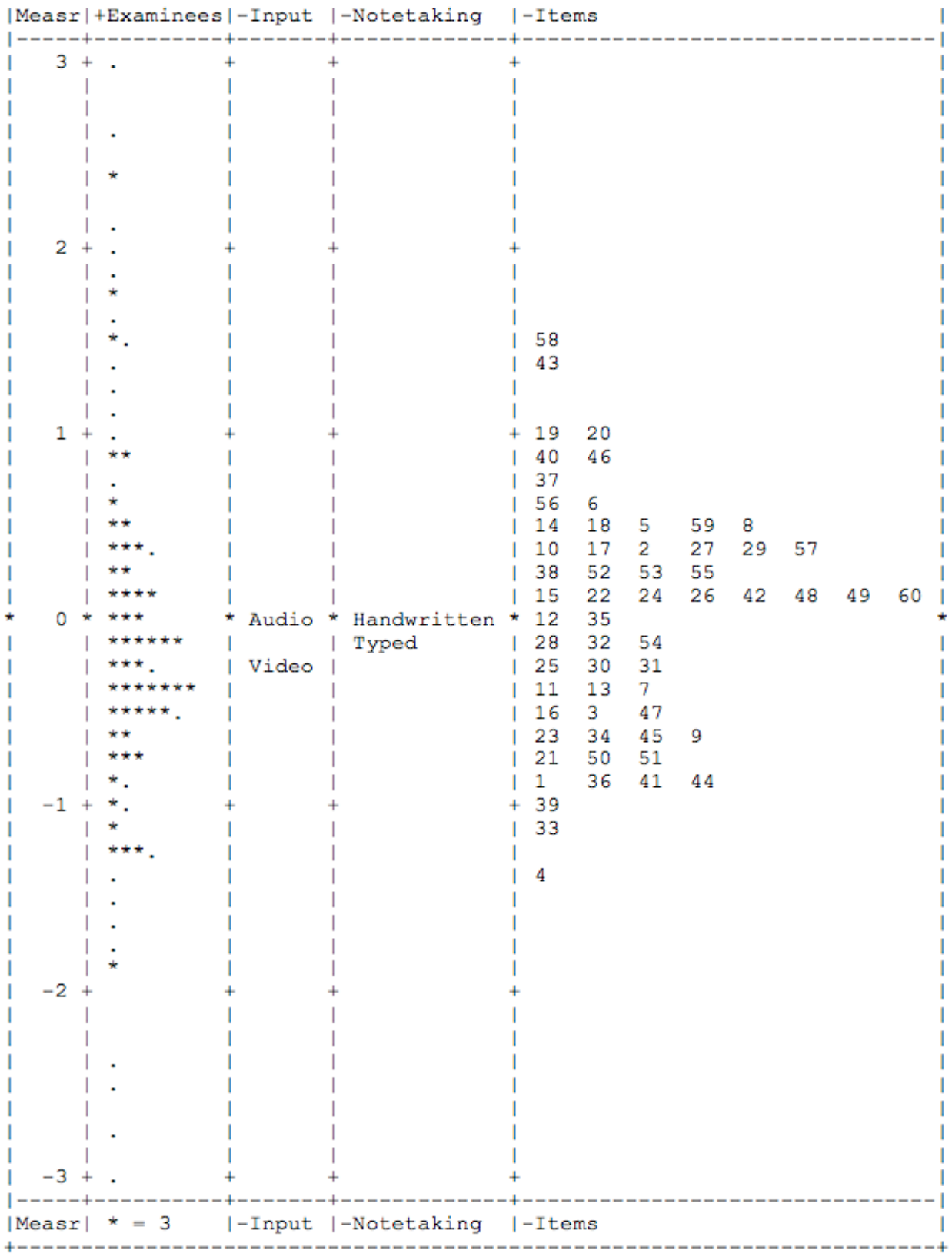



*Figure 4.3*. Variable map obtained from the many-facet Rasch analysis comparing items, input condition, note-taking condition, and examinees.

interpreted as corresponding to a more advanced proficiency level for the individual with the opposite being true for those positioned lower on the map.

**Item and examinee characteristics.** Logit values for all examinees ranged from –2.99 to +2.95, with a mean logit value of –0.11. The separation index for the examinees was 3.12, indicating that examinees were divided three to four statistically distinct groups with an examinee reliability output for the model equal to 0.91. The reliability value obtained from these models is analogous to reliability values of Cronbach's alpha (Bond & Fox, 2007). The fixed chi-square value for examinees in this model was equal to 1514.6 ($df = 199$, $p<.01$) showing that examinees in the model were statistically different in relation to their listening proficiency (the ability measure). The root mean square error (*RMSE*) for the examinees was found to be 0.33. Since this value is synonymous with standard error, a lower value is often sought and preferred (Brown, Trace, Janssen, & Kozhevnikova, 2016). While the value is low here, it does still indicate that there is some other noise within the data. Item difficulty was also assessed using similar measures obtained from the model. Items ranged in logit values from –1.42 to +1.51 with a mean logit value of 0. The separation index for items in this model was 3.79 showing that items were divided into three to four statistically distinct groups with a 0.94 reliability. The fixed chi-square for this model was equal to 830.1 ($df = 59$, $p<.01$) indicating significant differences in item difficulties. The *RMSE* for the items model was found to be even lower than that of the one obtained from examinees at a value of 0.16. Based on the separation index results, it would appear that the test is providing appropriate division of test takers into different proficiency levels. In general, the higher the separation index, the better, as this indicates that the test is differentiating test takers into enough different ability levels to provide meaningful interpretations. In addition, the mean logits for the examinees and items were quite similar to

each other in magnitude. This indicates that the items were overall well matched to the examinees and that they were able to complete the test.

Item model fit was also examined to ensure no serious problems were present within the items. When using Rasch analysis, item fit is examined through the use of infit and outfit statistics, with infit generally being the statistic that researchers focus on. Bond and Fox (2007) state that an item with an infit close to 1 is ideal as this indicates that the observed data fits the overall model and have proposed an appropriate range of 0.75 to 1.30 to indicate good fit. Others such as Wright and Linacre (1994) have actually suggested a more conservative range for dichotomous items that ranges from 0.80 to 1.20, which is what is used in this study. Items with an infit value below this range indicate that there is model overfit due to a lack of variation while a value over this range indicates that there is model underfit indicating that there is too much variability. Based on these criteria, item fit was found to be satisfactory for 59 of the 60 items between the two test forms. Question 20 on form B was found to be just slightly over the upper limit of the ideal range with an infit value of 1.25. This was considered negligible and therefore it was concluded that all items were targeting a similar listening construct. Appendices E and F provide complete lists of item and examinee logit values as well as item infit statistics.

**Item and condition comparisons.** Table 4.6 presents mean logit values, standard error, and confidence intervals for each of the different conditions examinees were exposed to in the test. Appendices G and H display the complete measurement report for both condition types and the Logit values for each item by condition.

Mean logits for both the audio-only and the handwritten conditions were anchored at 0 with each having a standard error of 0.03. Mean logits for the video and typed conditions

Table 4.6

*Summary of Logit Comparisons*

| Condition | Average Logit | SE | CI (.95) |
|---|---|---|---|
| Input | | | |
| *Audio* | 0.00 | 0.03 | (-0.06, +0.06) |
| *Video* | -0.24 | 0.03 | (-0.30, -0.18) |
| Note-taking | | | |
| *Handwritten* | 0.00 | 0.03 | (-0.06, +0.06) |
| *Typed* | -0.13 | 0.03 | (-0.19, -0.07) |

were –0.24 and –0.13 respectively, indicating that items answered in the video-based condition were slightly easier than those answered in the typed condition and that both were easier than those answered in either the audio-only or handwritten condition. Standard errors were used to calculate confidence intervals. The confidence intervals show no overlap between the different conditions, indicating that there may be some meaningful difference between the different conditions in terms of difficulty. However, Ockey, Papageorgiou, and French (2016) state that for differences to be meaningful, a logit difference of more than 0.5 should be present, suggesting that the differences in item performance for this study are not actually significant, which would be in agreement with findings from the two-way ANOVA described above.

In addition to looking at relative logit values between conditions, a bias analysis was run to investigate the differential item functioning within each condition. While no items were flagged as significantly departing from expected responses, several did appear to approach significance and are listed in Table 4.7 along with their bias size, probability, and the condition their bias size is in connection to. A *p*-value of 0.090 was determined as a cut-off for bias analysis in this case given the lack of significant bias estimates, but the desire to investigate possible sources of significant future bias.

Table 4.7

*Results with Greatest Significance from Bias Analysis*

| Item | Bias Size | *p* | Bias and Condition |
|---|---|---|---|
| 2A | -0.44 | 0.063 | Participants performed worse than expected in audio condition |
| 2B | 0.44 | 0.059 | Participants performed better than expected in video condition |
| 4A | -0.42 | 0.085 | Participants performed worse than expected in video condition |

Items 2 and 4 on form A and item 2 on form B showed some level of bias for the experimental input condition. No items displayed a bias of $p \leq .090$ in relation to the note-taking conditions. Items 2 on form A and item 2 on form B showed better than expected performance on the video condition while item 4 on form A showed worse than expected performance on the video condition. All three items were in relation to listening passages related to India's three-language policy (form A) and Fermi's Paradox and the Drake Equation (form B). Item 2 on form B presented test takers with a stem asking "What is Fermi's Paradox?" while item 4 on form A presented them with the stem asking "When India became independent in 1947, what language policy was planned?" This was surprising given the item facility and discrimination indices for these items, which indicate a moderate level of item facility for both items and quite satisfactory discrimination indices. Additionally, item 4 from form A performed less well than expected in the video condition. This is to be expected because the item does have a lower IF index (though still satisfactory), but is still surprising given the content visuals that were provided in the video condition. Reasons for these potential biases are discussed further in the next chapter.

**Research Question 3**

Question 3 asked about the extent to which visual input conditions and note-taking conditions accounted for the variance in performance on items meant to target different listening

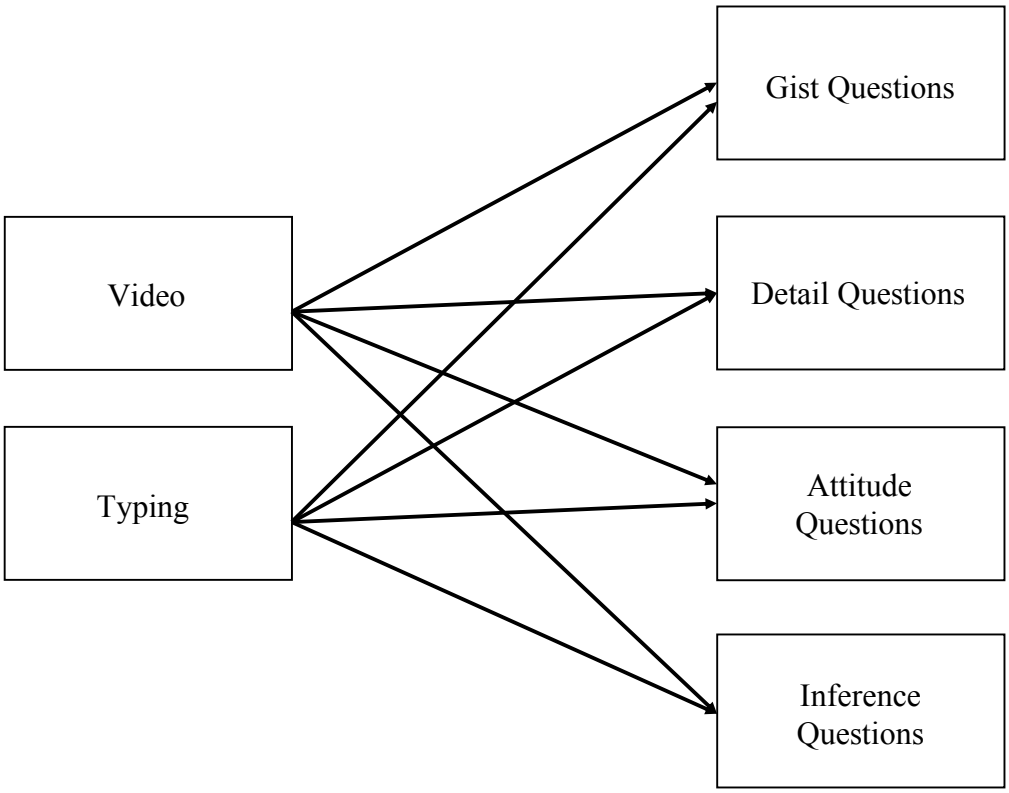



*Figure 4.4*. Initial proposed path model between condition type and question type.

skills. The questions in these tests targeted four main skills: (a) identifying the gist, (b) recalling main details, (c) identifying speaker attitudes, and (d) making inferences and connections. Composite scores were made for items in each of these categories, and these scores as well as the independent variables of input and note-taking condition were then subjected to a path analysis. The proposed path structure is provided in Figure 4.4.

**Initial model analysis.** In path analysis and structural equation modeling, variables are labeled as either endogenous or exogenous. Exogenous variables are independent variables that do not experience influence from any other variable found within the model. They are strictly accounting for the variance on the other variables within the model, which are endogenous. In this model, the independent variables of input and note-taking condition are present as two correlated exogenous variables. Since they are nominal variables, they are essentially treated as

Table 4.8

*Standardized Path Coefficients for Initial Model*

| | Video --> | | | | Typing --> | | | |
|---|---|---|---|---|---|---|---|---|
| | Gist | Detail | Attitude | Inference | Gist | Detail | Attitude | Inference |
| Form A | –0.050 | 0.023 | 0.114 | –0.030 | 0.025 | 0.012 | –0.010 | 0.015 |
| Form B | 0.146* | 0.265* | 0.021 | 0.169* | 0.070 | 0.093 | 0.064 | 0.143* |
| Form A + B | 0.057 | 0.121* | 0.065 | 0.081 | 0.041 | 0.045 | 0.025 | 0.081 |

*indicates significant path at p<.05 level

dummy-coded variables within the model (Byrne, 2012). They in turn have paths leading from themselves to the different endogenous variables, in this case the different question types. This same model was run for three different data groups: (a) form A alone, (b) form B alone, and (c) form A and B question types pooled together. Because the model used in the first portion of this question are more exploratory in nature, they are presented as "just identified models" meaning that all possible paths from exogenous to endogenous variables are being tested. Therefore, while normal path analysis or structural equation modeling would make use of fit statistics because they use overfitted models, this particular analysis will not use them because fit statistics for just identified models are not conclusive (generally delivering perfect results).

Table 4.8 displays the path coefficients between the condition variables and the question type subscores for each form of the test with each significant path marked by an asterisk. Examining the path coefficients for form A of the test, it is clear that none of the path coefficients from the conditional variables to the listening subskills is significant, indicating that neither of these conditions is significantly accounting for the variance in any of the subskills. What is possible to see from these coefficients though is that the role of video appears to play at least a slightly greater role than what typing plays in listening comprehension skill subscores. In contrast, the path analysis run on test form B displays several significant path coefficients, particularly in relation to the video condition. Upon examining these coefficients, it is clear that

video is having a significant positive effect on skills related to identifying the gist and details of the lecture as well as in making inferences. Additionally, typing was found to have a positive significant influence on inference type questions, but no other question types.

From Table 4.8, it is clear that the results for form B are quite different than they were in form A. In form B, path coefficients between input and gist and detail questions are significant while path coefficients between these two question types and note-taking condition are not significant. Since dummy coding assigned video as "1" and audio as "0", the positive coefficient on the paths associated with the input box indicate that as input variable increased to one (or went from video to audio), composite scores were found to be positively impacted. Thus, video is found to have a significantly positive effect on the scores for these question types. Inference based questions also found significant path coefficients for both test conditions, showing that increased scores were associated with typing notes and viewing video-based listening passages. Attitude questions did not exhibit any significant path coefficients and in all significant pathways, coefficients indicate a greater impact of video condition relative to note-taking condition on listening subskill performance.

Data from listening subscores on both tests was also compiled, providing results seen in the last row of Table 4.8. This model indicated that the variance of gist, inferential, and attitude questions was not significantly accounted for by either the input or the note-taking condition within the test. However, one significant path coefficient was found between detail recall questions and input type, indicating significant impacts of video-mediated listening passages on increasing scores on these questions and showing that input accounts for a significant part of the variance on detail related questions.

**Revised model analysis.** Based on the data from the three just-identified models, the initial model was revised so that detail question composite scores were used as an intermediary variable that was directly affected by input and notetaking, leaving the question categories of gist questions, attitude questions, and inference questions to be indirectly affected through details questions. This was decided based on the results of the model with the data combined from both test forms. It was clear based on this model that there was a clearer more direct effect on detail questions from the independent variables since the paths from the independent variables to detail-type questions were most commonly significant (especially in the case of video-based listening passages to detail questions). Since detail questions were more likely to be directly impacted by visual and note-taking conditions and the different listening comprehension subskills should be related in some manner, it was hypothesized that the detail questions would be directly affected by the independent variables and that the detail-type questions may serve as an intermediary variable that directly loads onto the other comprehension subskills. Thus, the model was revised based on these hypotheses, with the new proposed model provided in Figure 4.5. In this figure, it is seen that video and typed note-taking are hypothesized to directly predict detail comprehension. This effect then goes on to indirectly effect the other comprehension skills. Stated non-graphically, one could describe this situation by saying that video and/or typed notes either positively or negatively predict performance on detail-type questions. In turn, since detail questions directly predict performance on other comprehension skill types, it can be reasoned that predictions of more positive outcomes on detail questions due to the different condition will lead indirectly to predictions of more positive outcomes on other comprehension skills.

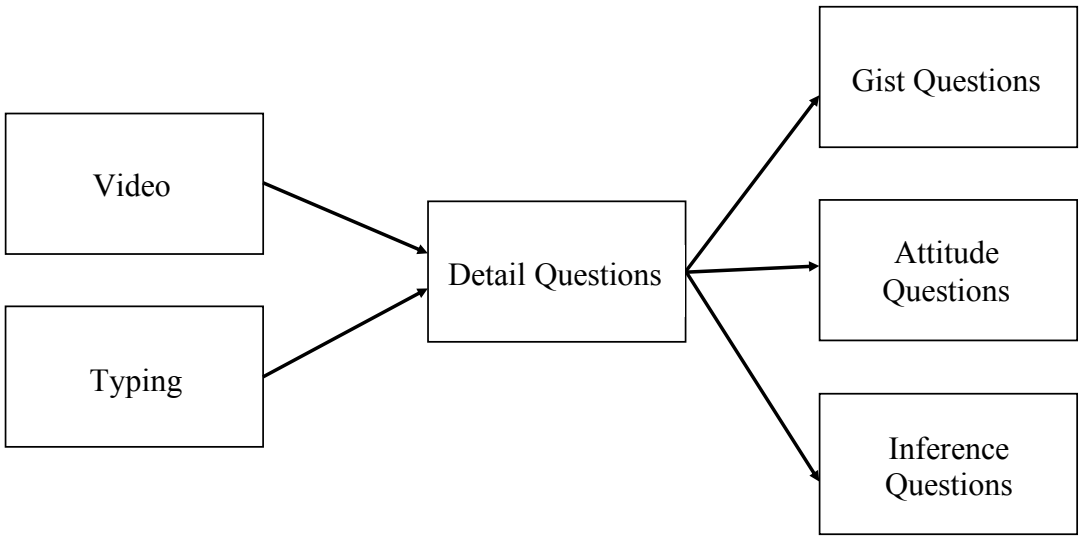



*Figure 4.5.* The revised path model with detail-type listening questions as an intermediary variable.

As with the original model, forms A and B were tested separately and then combined for a third path analysis as a composite test. As this model no longer tests all possible paths and restricts several paths from input and note-taking conditions to certain question types, it is no longer a just-identified model. Rather, it is now an overidentified model meaning that it has fewer parameters than observations (Kline, 2011; 2012). As such, fit statistics must be provided to assess the quality of the model overall. Fit statistics for each of these models is found in Table 4.9.

As can be seen, in all cases the chi-square value is not significant, indicating that the specified model does not significantly differ from the observed values. Therefore, this is a good indication of model fit, especially since larger sample sizes can easily force the chi-square test to output a significant value (Byrne, 2012). Additionally, Byrne (2012) states that it is possible to compare mean and covariate structures across models representing different sets of data by comparing the chi-squre values and degrees of freedom. In doing this, it was found that all

Table 4.9

*Chi-Square and Fit Statistics for the Revised Path Model.*

| | $\chi^2$ | *df* | *p* | *CFI* | *RMSEA* | *SRMR* |
|---|---|---|---|---|---|---|
| Form A | 3.825 | 6 | 0.70 | 1.00 | 0.00 | 0.03 |
| Form B | 3.450 | 6 | 0.75 | 1.00 | 0.00 | 0.02 |
| Form A + B | 3.534 | 6 | 0.74 | 1.00 | 0.00 | 0.02 |

versions of the revised model were not significantly different from each other ($p > .50$), thus leading to the conclusion that the models are essentially the same.

In addition to chi-square values, three other fit indices are also examined. These are the root mean square error of approximation (*RMSEA*), the standardized root mean square residual (*SRMR*) and comparative fit index (*CFI*). While the *CFI* is scaled as a goodness of fit index, the *RMSEA* and *SRMR* are scaled as a badness of fit index, making it so that a higher *CFI* is desired and a lower *RMSEA* and *SRMR* are desired (Kline, 2011). In general, for a model to be considered to have good fit based on these indices, a *CFI* value close to 0.95 is sought (Hu & Bentler, 1999), an *SRMR* value of less than or equal to .08 (Kline, 2011), and an *RMSEA* value less than .05 (Browne & Cudeck, 1993) are desired. Based on these indices, it is clear that the new, revised model fits the data adequately and the parameters can be examined more closely.

Figures 4.6, 4.7, and 4.8 display relevant path coefficients, disturbances, and correlations for variables in the revised model. The arrows and coefficients in the path analyses represent regression coefficients that show the predictive power of one variable on another. The circles that point to each of the comprehension subskill boxes represent the disturbances (variances) of each of the measures, which fall on a scale of 0 to 1 where a lower value indicates that the measure is accounting for more of the predictability and a higher value indicates that some other possible factor may be influencing predictability. Finally, a correlation arrow is seen between the

video and typed boxes, which is standard practice for the independent variables. Since these variables are dummy coded as 0's and 1's, there is a correlation of 0 for this particular arrow.

Examining the path coefficients in the new model, it is possible to see some changes. Figures 4.6, 4.7, and 4.8 display path coefficients for form A, form B, and the combined forms respectively. While the effects of input and note-taking conditions on detail questions are similar to what they were in the previous models, it is easy to notice that gist-, attitude-, and inference-type question all are significantly related to detail questions in some way, suggesting that the relationships between question types are somewhat similar across test forms, but that the content differences between the lectures about which the examinees are being asked may not be affected by input in all situations. When looking at path coefficients, it is still clear that input provides a greater impact overall. When compiled, the results are similar in that they indicate significant paths between other question types and the detail question types and that detail question types do indeed seem to be functioning as an intermediate between input and the other question types rather consistently.

Altogether, the results provided by these models is slightly strange given the differences between the path coefficients and their significance between forms A and B, which could be a result of several factors such as some lecture topics being more interesting than others or the types of visuals used for certain lectures. However, even given the odd nature of the results, it is still relatively clear that input does account for variance within at least some of the items for the two tests and that this is particularly clear when considering the impact that the different conditions that test takers were exposed to are investigated in relation to detail question types. In addition, as was mentioned in Chapter 3, there was an unequal distribution of items across subskills in an effort to more accurately reflect the academic listening construct. While there was

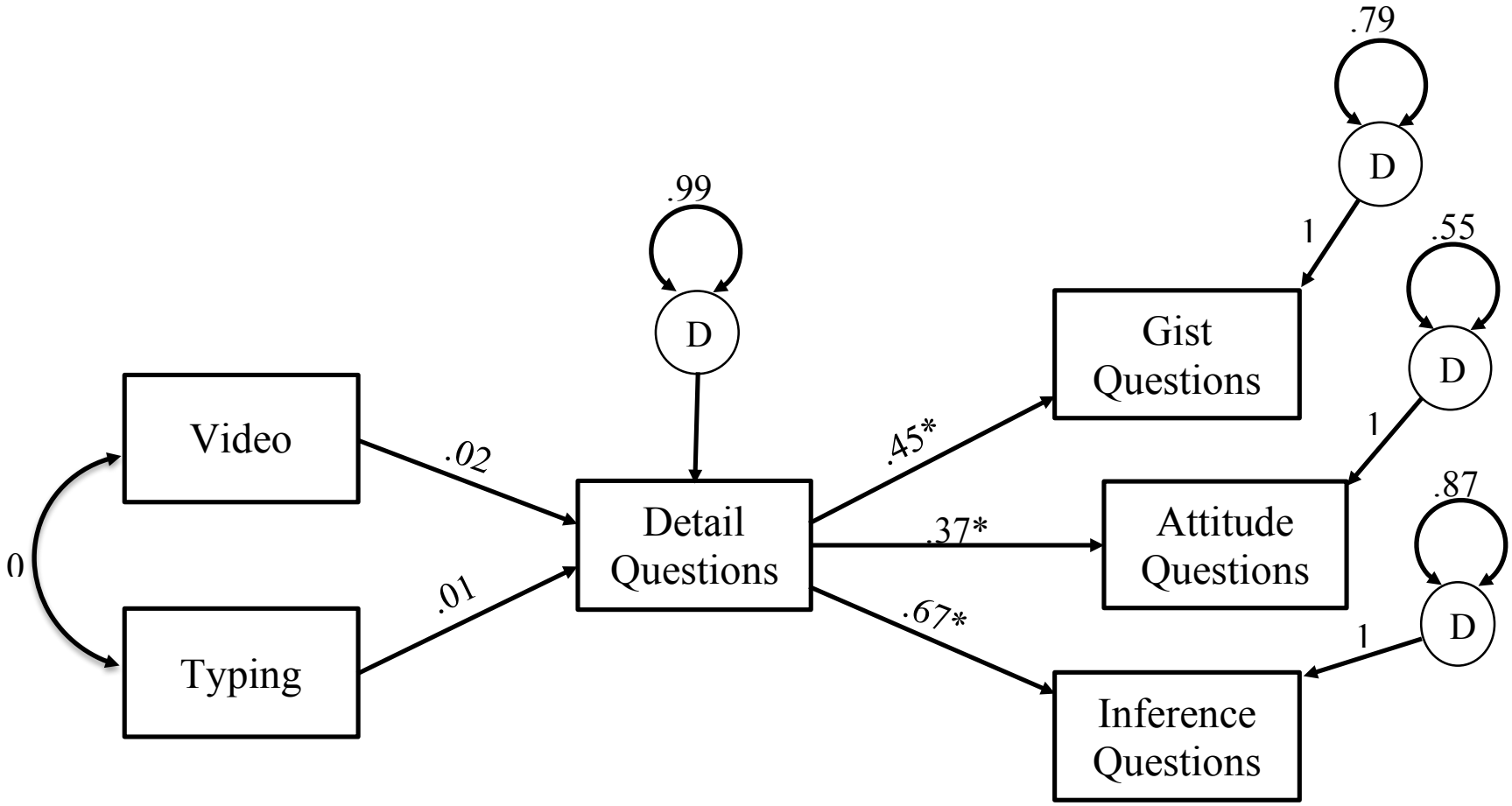



*Figure 4.6.* Revised model with path coefficients for form A.
* = a significant path coefficient at $p<.05$

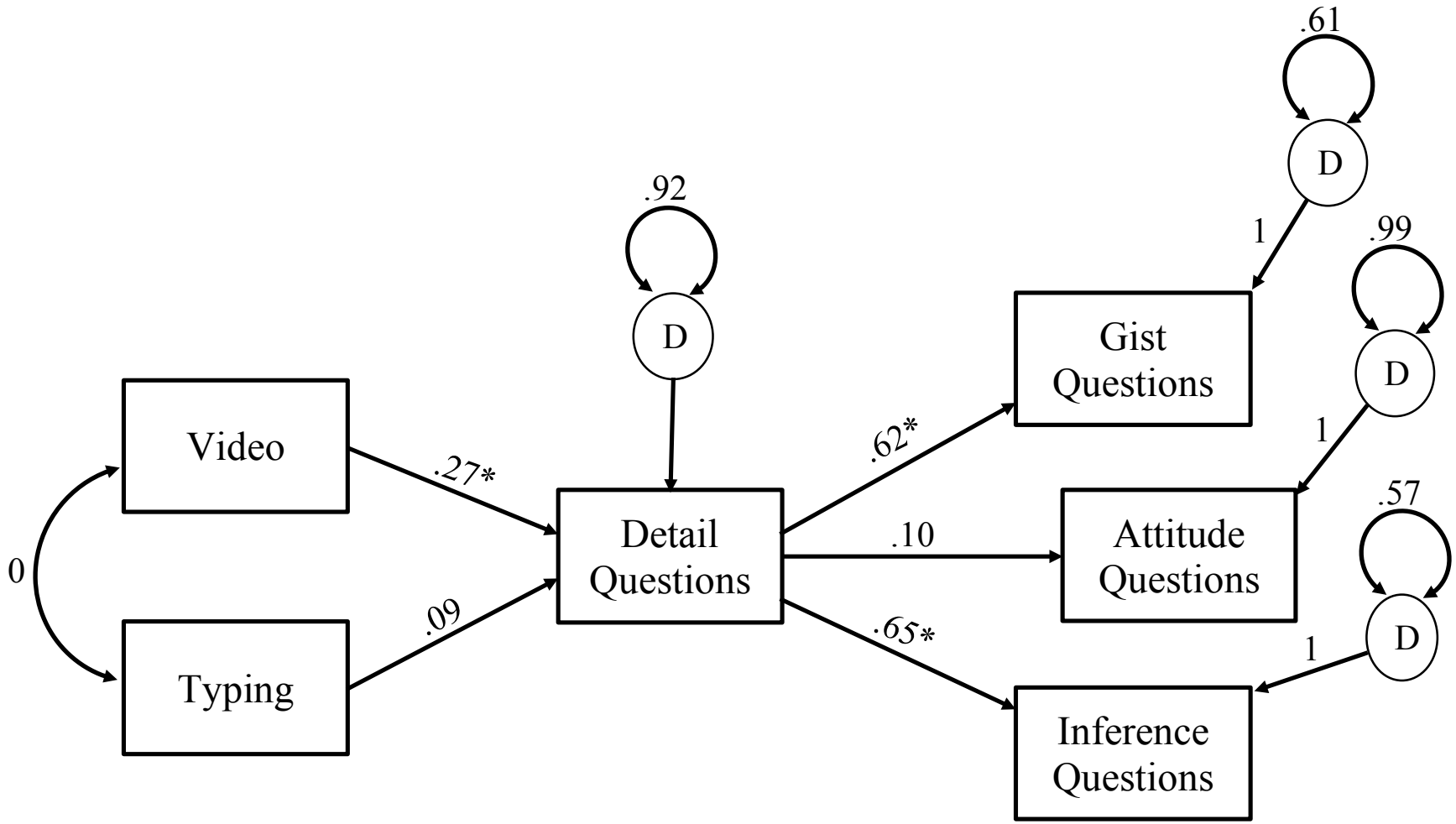



*Figure 4.7*. Revised model with path coefficients for form B.
* = a significant path coefficient at $p<.05$

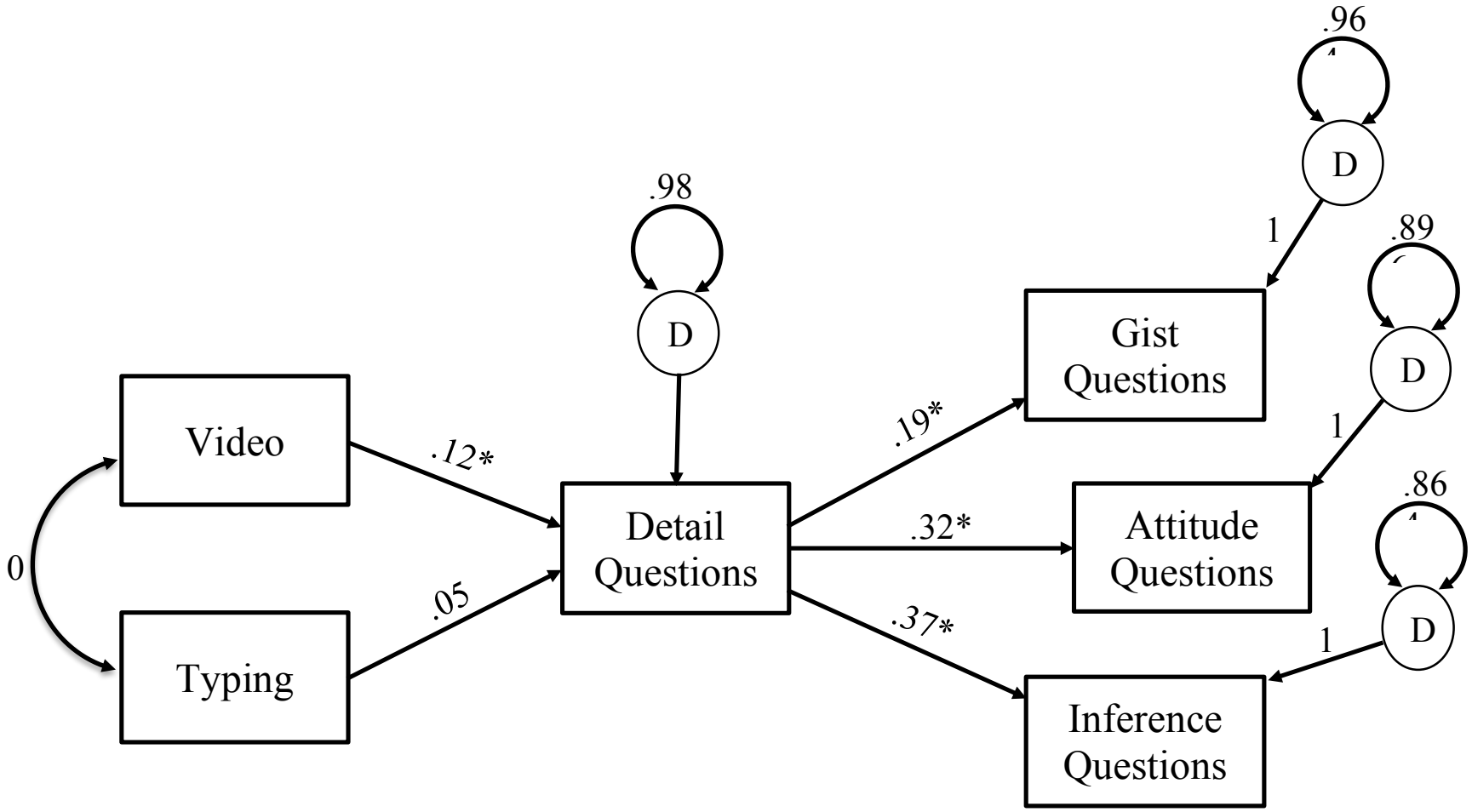



*Figure 4.8.* Revised model with path coefficients for forms A and B.
* = a significant path coefficient at $p<.05$

a chance that having a greater number of detail-type items could have an influence on the results, the analysis here can lead one to confidently claim that this was not the situation for this set of data. If a greater number of detail-set question was leading to higher loadings, it should have been the same across all models tested. However, as was seen in the results, detail-type questions had low and insignificant loadings for both models tested with form A. If item numbers were an issue here, similar results should have been seen across all models on that particular path coefficient. Thus, the results were determined to be unaffected by the number of items, though there is the possibility that some items with lower numbers may have displayed higher loadings with greater numbers (particularly the attitudes-type items), which is worth examining in the future.

In an attempt to gain a clearer understanding of these quantitative results, qualitative analysis was done in order to answer research questions 4 to investigate if learner perceptions of the different tasks and conditions could explain the results obtained in the first three analyses.

**Research Question 4**

The final research question investigated the test taker perceptions related to each condition in relation to the preferences they expressed and what they found themselves focusing on the most in the video conditions as well as any challenges they experienced. Of the eight questions found on the post-test questionnaire, five were coded and assigned to thematic categories, as the other three were used to check for previous exposure to passage content, to collect information for a later examination of the data, and one question unexpectedly did not provide much in the way of additional information that the other questions did not provide. Responses to the question that did not provide much additional information were collapsed with responses to the preceding question since these questions essentially complemented each other

and served to provide clarification in some circumstances. These responses were then coded and assigned to appropriate thematic categories based on the procedure outlined in Chapter 3 to better explain results found regarding the previous research questions. A table of key words and phrases and their associated codes and themes can be found in Appendix I.

Table 4.10

*Input Preferences*

| **Input Preference** | ***N*** | **Percentage** |
|---|---|---|
| Video | 174 | 87% |
| Audio | 20 | 10% |
| Both | 4 | 2% |
| Other/No Response | 2 | 1% |

**Lecture style preference.** Participants were asked whether they preferred the audio-only or the audio-video lecture style. Raw numbers for their stated preferences are presented in Table 4.10. The vast majority of test takers stated that they preferred the video format ($N = 174$) while 20 others stated that they preferred audio, two said they had no preference, and one person failed to provide a relevant answer to the question (which appears to be due to her low proficiency level, as she received scores of 0 and 2 on the two test forms). Upon examining the reasons that participants provided for their preferences, several themes emerged related to each of the three main preferences and they are outlined in Table 4.11.

By far the most common of the themes was the opinion that the video aided in comprehension of new and unfamiliar terms. This theme was seen in 107 of the 174 comments stating a preference for video-mediate lectures. Responses representative of this theme expressed by participants consist of the following[2]:

[2] All responses are provided verbatim from participants and are not changed except where indicated by brackets for the purposes of clarity.

Table 4.11

*Themes Associated with Responses Based on Preference*

| **Preference** | **Theme** | ***N*** |
|---|---|---|
| Video | Video Aided in Comprehension | 107 |
| | Video Provided Enhanced Focus | 57 |
| | Video Created Greater Authenticity | 6 |
| | Easier to Get Back on Track when Lost in Material | 3 |
| | Other/Off-Topic | 8 |
| | No Response | 2 |
| Audio | Easier to focus on listening/taking notes | 15 |
| | Difficulty due to note-taking type (hand vs type) | 3 |
| | No Response | 2 |

*Audio with picture is better because some topics are unfamiliar subject and it was hard to memorize or write but if the lecture provides picture I could write to prepare to study.* (Participant 26)

*I prefer the audio with video because it helps me understand more being about to see pictures and diagrams.* (Participant 108)

*I prefer the audio with picture because the picture help to understand the content of speaking* (Participant 58)

*I prefer the audio with picture because I could see some notes, key words about the lectures* (Participant 198)

As can be seen from these responses, comprehension was aided in several ways. For instance, while Participant 26 mentioned in a more general sense that it was easier to understand the lectures with video-mediated passages for unfamiliar material, Participants 108, 58, and 198 stated that it was easier to understand due to specific aspects of the video, such as the pictures and diagrams or key words on the screen. Responses in this category all mentioned the videos in this manner.

An additional 57 comments relate to the video providing enhanced focus while listening, with several participants stating that this helped them and made them prefer the video presentation method. The following are several representative responses stating this from the survey:

> *The audio with picture. I was more focused and I concentrated only what was told.* (Participant 4)
>
> *Video, it provided helpful slides for the guiding my focus and I could see the teacher.* (Participant 167)
>
> *I like the audio with picture* [video condition] *because then you can concentrate only on listening*. (Participant 88)
>
> *I prefer the audio with picture. Because first of all, the picture make me pay attention. Second, it is easier for me to understand the lectures with picture and information showing.* (Participant 107)

By examining these responses, it is possible to see that responses belonging to this category exhibited several common characteristics. Responses explicitly referenced enhanced focus in some manner and made it clear that it was a positive effect. Phrases such as "make me pay attention," "more focused," or "guiding my focus" are seen in responses made by Participants 4, 167, and 107. Participant 88's response did not quite fit these characteristics. However, it was still placed within this category due to its mention of concentration, which, when taken in context, led the researcher (and the second coder) to interpret the respondent as meaning that less time was spent processing more difficult or unfamiliar words because of the visuals, so more content of the listening was able to be focused on. It should also be noted that Participant 107's

response was coded twice due to the second portion which mentioned that it was easier to comprehend lectures, thus, this was counted as part of the previous theme as well.

The remaining reasons for preferring video were less frequent but could be classified into themes such as video making the listening more real or authentic ($N = 6$) or making it possible to get back on track if they found themselves lost at any given moment ($N = 3$). Any remaining reasons were either classified as "other" ($N = 8$) since they did not fit cleanly into a major theme, or were nonexistent ($N = 2$). The following comments express ideas related to the themes of authenticity. Any response that explicitly mentioned that the test taker felt that it was real or realistic or that it was easier to imagine the classroom was placed in this category:

> *I prefer the video because I can imagine myself better to be in a classroom.* (Participant 11)
>
> *Video. It was more realistic and helpful. Audio-only is more harder to understand* (Participant 175)
>
> *Video, because it felt more real. More easy to understand.* (Participant 16)

These comments are particularly interesting because they indicate that examinees could see that the video was more realistic, which would seem to have a number of implications for how the academic listening construct is defined within a test of listening comprehension and serves to provide face validity for the video-mediated format of the exam.

The following represent reasons related to the theme of using video to get back on track when comprehension efforts derailed:

> *I prefer the audio with video because I could see some key words which the lecturer were talking about, so it sometimes helped me when I can't understand.* (Participant 17)

*Taking notes sometimes made me get lost. The words and pictures on slides helped me keep up.* (Participant 151)

*Audio with video. It gives more material to catch up if I miss anything.* (Participant 119)

These three participants each clearly related to getting lost in the listening input and the fact that they found it easier to get caught up when their comprehension skills were not able to keep pace with the listening material.

In addition to themes arising for those who preferred the video input method, several themes arose for those who preferred the audio-only input method as well. The primary theme expressed by 15 of the 20 participants stating that they preferred audio was that they found it easier to focus on listening and/or taking notes. Examples of statements related to this them are as follows:

*I like only audio. Video sometimes takes my attention away…* (Participant 199)

*Audio-only, it's easier to focus on listening and making notes* (Participant 154)

*I liked the audio-only lecture. this is because I cant concentrate on both of taking notes and watching a video.* (Participant 3)

*I prefer audio lecture because I can be more focused. Somehow, the slides with only letters are distracting. If the slides were showing only pictures, it would be helpful.* (Participant 5)

Responses that fit within this theme are well-represented by the comments from these four participants. For responses to fall in this theme, participants had to explicitly state that audio-only lectures made it easier for them to focus on the lecture/taking notes as is seen in Participant 5 and 154's responses, or they must have stated that video prevented them from doing these activities due to distraction, as in Participant 199's and 3's responses. Normally, focus on taking

notes and focus on listening would have been separated, but for this sample of participants, the two ideas were far too connected to make it worthwhile to do so.

The remaining five participants either provided reasons that were related to the type of note-taking required of the participants and how it caused some difficulty (N = 3), or did not provide any reason at all (N = 2). Two respondents stated that they had no preference for either conditions, providing the following statements:

> *Both, because I'm an auditory learner. It's enough so long as the information is given clearly and I can take notes on it.* (Participant 45)
>
> *For audio lenguage I prefer Indian lecture and audio and video I prefer* [lecture] *about choices.* (Participant 9)

Here, the two participants stated that it didn't matter to them (as is the case for Participant 45) or that they preferred visuals for certain lecture topics, but not others (as is the case for Participant 9). Participant 9's statement in this preference group raises an interesting issue showing that some participants may have found visuals more useful for certain lecture topics, or even that certain lecture topics may be more interesting with visuals (or in general) than others. These potential issues have several implications for test development and future research that will be discussed later on in Chapter 5. Overall, the responses indicated a preference for video-mediated listening material, with most of the reasons indicating that it helped with focus and comprehension. Several additional comments that did not cleanly classify into one of the major categories will be discussed in more detail later. Taken together, the comments found here related to preference seem to indicate some reasons for why the Rasch analysis discussed earlier may have indicated that video-based listening passages were slightly easier. Based on their responses, students were able to use these passages to get back on track when lost and to more

easily focus on and make sense of material due to the content provided in visuals. This is very useful for further discussing implications for test development in the following chapter.

**Helpful and distracting characteristics of video lectures.** In addition to asking examinees about their lecture input preferences, they were also asked specifically if the video lecture caused difficulty in focusing on note-taking and if they felt that visual elements necessarily facilitated the recall of information from the lecture. An overall summary of the results and the themes related to these results is found in Table 4.12.

Table 4.12

*Responses to Whether Video Caused Distraction and Associated Themes*

| **Opinion** | **Theme** | ***N*** |
|---|---|---|
| Video did not distract | | 59 |
| | Visual aids were helpful | 57 |
| | Made lectures authentic | 3 |
| Video distracted | | 49 |
| | Focused more on visuals than listening | 33 |
| | Unable to divide attention | 10 |
| | Video lectures were faster | 5 |
| | Context videos distracting | 1 |
| Sometimes | | 4 |
| Off-Topic | | 104 |

**NOTE: in tables 4.12-4.14, some responses had to be counted twice due to fitting in multiple categories, so total may be greater than 200 in some cases.*

Overall, while the majority of examinees found the video to be their preferred method of listening passage delivery, and all of these examinees felt that the video aided in comprehension to some degree, a small majority of examinees found that the video made it easier to concentrate, though fewer found that it aided concentration than preferred video as the input method. For both those who said it aided and those who said it distracted from note-taking and attention, the

primary theme for both was often related to the visual aids themselves in the video. Of the 59 examinees who felt video was helpful for concentration, 56 mentioned visual aids as a factor. The following are representative statements from those who stated that video aided in focus and note-taking:

> *No. It helped; it provided me the high-level points to summarize.* (Participant 36)
>
> *No, I didn't lose concentration because visual aids were very helpful to write down information.* (Participant 81)
>
> *Not really. It helped me to focus on the important parts because of slides and gestures.* (Participant 4)
>
> *Video was helpful for me because only audio, I did not know what information I had to take notes. ppt was helpful for this.* (Participant 13)

Statements placed in this category were those that explicitly stated that the visuals aids were helpful in some way. Each of the statements that fell within this theme mentioned some variant of "it helped" or "it was helpful," as can be seen in each of the statements above. In addition, many of these statements pointed to specific aspects of the visuals that the participant found helpful. For instance, Participant 36 found the visuals helpful for identifying the main points for summarizing, Participants 81 and 13 found the video helpful for its ability to call attention to what should be written down, and Participant 4 found it helpful for pointing at important parts. These are the reasons seen consistently in responses falling under this theme, indicating that visuals were most helpful for summarizing information and drawing focus to the important points that should be written down in the notes.

The theme of authenticity also appeared again in responses to this question, with two participants stating the following:

*It is a common way in all class in school, so it was natural.* (Participant 126)

*No, it felt almost as same as real lecture.* (Participant 190)

Here, just as in the previous section, responses referred to the connection that the participants could see between the visuals and what they would encounter in a real classroom. These comments are particularly helpful in establishing the face validity of the test and also in giving credibility towards the construct definition provided in this study.

The remaining respondents who stated that the video did not distract from note-taking or concentration either provided an off-topic response to the question by stating that note-taking medium detracted from concentration or did not reply. Similar responses were found in the group stating that video did distract from note-taking, and those responses were likewise removed from theme analysis.

For those who found that video was primarily a distraction from note-taking and comprehension, the main theme found in 33 of the responses was that video caused a greater focus on visuals than on what the speaker was saying. This is seen in the following examples:

*It was mainly fine, but sometimes I got distracted by reading all the text on the slides so I wasn't paying attention to what she was saying anymore.* (Participant 132)

*Yes…because I am trying to copy the words from ppt, and forgot what teacher said..* (Participant 34)

*Yes. It was harder to take notes and extract information that I think is useful when the video was playing because I was essentially copying the notes from the screen.* (Participant 54)

From these representative examples, it is possible to see that the main criteria for falling under this theme was to state the nature of a distraction that related in some way to note-taking. The

distractions provided by participants were essentially related to the idea that they were more focused on writing words down from the slides than on listening to the lecturer, which is seen in all three response provided here. This is an important point when considering the construct representation of the test because it is a possible indicator that students do not have appropriate note-taking strategies and are not adequately monitoring and adjusting their comprehension skillset for the situation, signaling that the absence of the visuals from a test of academic listening comprehension such as the one in this study may result in under-representation of the construct being targeted, which would serve to put students at a disadvantage once they are put in this environment and unable to attend effectively to all stimuli. These implications are discussed further in the following chapter.

Of the remaining responses, the second major theme from this group of individuals was related to the limited cognitive resources the respondents had to take everything into consideration making it so that they could not divide their attention effectively ($N = 8$). The following are representative responses for this theme:

> *Sometimes I look at the video more and not take notes. A little difficult to concentrate because of this.* (Participant 114)
>
> *It difficult to take notes when video was playing. I cannot focus on what her talking about.* (Participant 90)
>
> *I think it was a bit difficult to take notes in the first test with the movie. Since it was so much that happen at the same time.* (Participant 29)

As can be seen from these example responses, participants would have their statements fall in this category if they mentioned difficulty in going back and forth between the different tasks they

were expected to perform (i.e., note-taking, watching the video, and comprehending the aural input).

Another interesting theme within this group that connects with the previous theme related to the talking speed of the lecturer in the video condition compared to the audio condition. This was mentioned by five different participants who represented examinees who had taken both form A and form B in video format:

> *When I took notes, the video was so fast I can't do both things so I just take notes keyword.* (Participant 200)
>
> *Difficult. Speaking more fast then listening with no picture.* (Participant 71)
>
> *Yes. To take notes is hard with video. She speak to fast. Audio is more easy.* (Participant 83)

These responses show that several of the participants made an observation related to the speed of the video-mediated listening passages in some way. This was regardless of the test form that the participant was taking at the time, indicating that some underlying factor was leading them to potentially perceive video-mediated passages as being faster than the audio-only passages, which was not the case. This could be related to issues related to the split-attention effect described in Chapter 2 (Horz & Schnotz, 2010; Mayer, 2005).

Finally, one participant mentioned that the context visuals found in the video were distracting, thus falling into her own theme:

> *Some texts on the wall, 'do not drink…' that distract me from the lecture make video difficult to concentrate.* (Participant 119)

This statement was somewhat surprising. When developing the videos for the lecture different lectures, care was taken to prevent context visuals from being too prominent in the video so that

focus would be on content visuals provided. However, this comment indicated that even some of the smaller, stationary context visuals may distract individuals while trying to comprehend a listening passage. This comment does show agreement with Suvorov's (2009) findings that context visuals can be potentially distracting and shows that anything in the classroom environment during a lecture has the potential to lead to some form of interaction between itself and the listener.

In addition to asking participants about the effect that video-based listening passages had on their concentration, they were also asked to reflect on whether video aided in their ability to recall important information from the lectures, the results of which are seen in Table 4.13. The

Table 4.13

*Responses and Themes Associated with Whether Video Aided Recall While Answering Items*

| **Opinion** | **Theme** | ***N*** |
|---|---|---|
| Aided Recall | | 158 |
| | Slide images activated image memory | 79 |
| | Allowed visualization of lecture during recall | 10 |
| | No reason given | 69 |
| Did not Aid Recall | | 34 |
| | Lack of concentration hindered later recall | 25 |
| | Video helped concentration but not recall | 1 |
| | Listening material was too difficult | 1 |
| Not Sure | | 4 |
| | Only some pictures were helpful | 4 |

majority of examinees ($N = 158$) stated that they believed their recall of information was aided through the visuals provided in the video-based listening passages with the major theme ($N = 79$) being related to image-based memory being activated due to pictures and words on slides:

> *Yes, sometime I didn't have time to write down the information showing on the presentation, I could still remember it. I think it is kind of image memory.* (Participant 46)

*Yes, because it had pictures. One of the audio lectures had an equation, which was pretty hard to write down only relying on the audio.* (Participant 175)

*The pictures did help. I could remember the pictures more and this helped me anser the questions.* (Participant 193)

*Yes, because I got visual facts so it helps me to remember more than just audio lecture.* (Participant 14)

*Yes, because I remember some information with my ears and hands and eyes and brain. I think it is easier remember information of the video lecture. We have images and text that we can remember.* (Participant 156)

Responses were assigned to this theme if they referred to the images (either visual or textual) from the slides making it easier to recall information. For instance, participants 46, 193, and 156 each mentioned that they were able to recall the different images they were presented with later on when answer questions, making it easier for them to respond. In addition, participants 175 and 14 each stated that images made it easier than the audio-only lecture to remember, offering a direct comparison between the two. In addition, participant 175 provided a specific example of how these visuals aided recall, stating that certain things were more difficult to remember by just hearing them and that certain ideas or details would actually benefit from being presented with a visual (in this case the equation from the Drake Equation lecture). These were common ideas expressed by all participant responses falling under this theme.

The other primary theme was much less common, consisting of ten responses that stated that video allowed examinees to visualize the lecture while answering questions:

*Yes, I could imagine easier.* (Participant 170)

*Yes because you can visualize the video lecture.* (Participant 117)

*Yes, I could remember the pictures so it was easy to rebuilt what the lecture was like.* (Participant 169)

Thus, statements related to imagining the lecture or images while answering questions were placed in this category. Key words that helped to indicate a statement should go in this theme were "imagine" (which was the most common), "visualize," or "rebuild."

Many fewer examinees said that the video did not aid in recall of information. However, those in this thematic category gave reasons that fell into several different themes. The most common theme ($N = 25$) related to concentration once again, with many stating that lack of concentration due to video hindered later recall of what the speaker said:

*I think it was easier for me to watch no video one because I can concentrate on only the audio.* (Participant 98)

*Not really. Sometimes information on screen makes me remember only the things on screen and hard to focus on what the speaker is saying.* (Participant 40)

*I only remember what shows on ppt. only a little from what teacher said. And it is hard to make connection to the sentence form ppt.* (Participant 77)

*It was easier to listen no video because I can listen to the lecture and not lose focus.* (Participant 10)

Here, statements focus on some aspect of concentration or focus. For instance, Participants 98 and 10 state that lack of video was better for recall because it was easier to listen to and focus on the lecture without something else requiring part of their attention. This was the primary reason provided by participants falling under this category. Other reasons that related, but were much less common are seen in Participants 40 and 77. Participant 40 was placed in this theme because her response states that video led to concentration only on the visuals, with no attention paid to

what the lecturer was saying. Thus, this indicates that she was not able to use these visuals to make connections to the aural input and use that to her advantage when answering questions because her concentration was not divided in an effective manner. Participant 77 provides another example of a common response for this theme in his statement that he concentrated only on the PowerPoint slides (much as Participant 40 did). Thus, because of this lack of focus on the aural input, connections could not be made that could have potentially aided in recall.

Additional themes mentioned by one or two other respondents related to issues of concentration during the listening being helped by video, but not recall, or the listening material being too difficult:

> *First I concentrated, so I can remember about that, but I cannot final question* (Participant 108)

Here, the participant states that they were able to concentrate on the video, but that during the questions they could not remember what they had concentrated on.

> *No, there are many information in the video. It was really difficult.* (Participant 129)
>
> *Not really. Because I cannot understand the terminology so I cannot keep tracking in memorization.* (Participant 106)

These two participants fall in this category since they commented simply on the difficulty related to the amount of material presented (Participant 129) or on unfamiliarity with the academic register that was being used (Participant 106).

Finally, one small group of examinees preferred not take a side on this question, stating that the video may have helped recall, but they were not sure. One participant from Japan expresses the ideas of these individuals well by writing:

*Maybe, I am not sure. I think some pictures helpful, but words weren't easy to remember.* (Participant 29)

This indicated that perhaps some of the content visuals provided by the lectures were not as useful as others. For instance, it seems here that this participant found the pictures and charts much more helpful in recalling information than the words. This may indicate that different types of content visuals serve different functions. Some participants earlier stated that the main points helped them to get back on track when they were lost in the listening. Perhaps the primary purpose of these textual visuals is to help aid in summarizing and catching up to the input. In contrast, picture-based visuals may be better at both helping the listener to understand concepts being described and may serve as recall devices that aid individuals in answer questions upon completion of the listening passage.

**Participant focus within video lectures.** One last question participants were asked regarding the use of video in the listening passages was what aspect of the video they found themselves focusing most of their attention on. Table 4.14 shows a summary of the numbers and themes. Several themes arose from the responses obtained, with by far the most common theme being that the slides were the main focus in the video. In fact, of the 200 survey responses,

Table 4.14

*Themes Related to Focus While Watching the Video-Based Lecture*

| **Theme** | ***N*** |
|---|---|
| PowerPoint slides were primary focal point | 147 |
| Instructor's gestures and speech | 27 |
| The subject matter of the lecture | 1 |
| Listening for key words, NOT visuals | 2 |
| Did not look at video | 2 |
| Off-topic/Unclear/No Response | 35 |

147 stated that the PowerPoint slides were their primary focal point. The following responses illustrate this point:

> *When I was listening the video with pictures, I did not focus on her statements, instead, I focused on typing the slides.* (Participant 73)
>
> *I find myself paying most attention towards the information that were written at the screen* (Participant 109)
>
> *Mostly the screen because there are key points on the screen, so it is helpful to understand what they talking about the time.* (Participant 149)
>
> *Picture on the screen because it had power point behind her so even if I couldn't catch up with some words, I could see it from the power point. Other wise, I wouldn't be able to answer the question of the lecture.* (Participant 100)

Responses were determined to be representative of this theme when certain key words related to the slides were seen. These key words are all seen in the example responses seen here and are "slides," "information/key points on the screen," "pictures on the screen," or "PowerPoint/power point/ppt." These words indicated a primary focus on the slides projected to the left of the speaker.

In addition to themes related to the slides presented in the video, a smaller subset of individuals ($N = 27$) commented on the instructor's gestures and speech being their primary focal point. The following statements illustrate this:

> *The content, I liked the way the lecturer explained beyond the slides of the presentation. Also she uses her hands for emphasizing, it was good but sometimes distracting. The slides were thorough.* (Participant 65)

*When she speech more strong accent something, and information presented on the screen. I can more attention the video lecture.* (Participant 181)

*Her gestures and voice is really good. I understand very well.* (Participant 175)

*I payed attention to the teacher and her gestures, the powerpoint and of course the content.* (Participant 119)

These statements display the representative key words that led to them being classified in this theme. Words such as those related to focus on certain body parts of the lecturer, such as in Participant 65's response, where the participant made reference to her hands or face/mouth movements led to this theme classification. In addition, reference to the speech of the lecturer led to classification in this theme. For instance, Participant 181 refers to the accent of the lecturer's speech (i.e., the use of her voice to emphasize certain points), and Participant 175 also made reference to her voice. Participant 119, whose response was classified multiple times due to his reference to the PowerPoints and content in addition to the lecturer, mentioned her gestures specifically. Thus, responses making reference to the speaker's gestures, voice, or motions of specific body parts (most commonly her hands) were placed under this theme.

Beyond these themes came a number of other statements that were not necessarily repeated by other participants or were off-topic or unclear. One related to focusing more on a lecture of a particular topic because it was interesting:

*I was paying the most attention towards the lecture about alien life because it was more interesting compared to the others.* (Participant 36)

This participant indicates that interest in content of the lecture may have also had an impact on performance. This could have implications for test development, though it may be quite difficult to control for given the wide range of interests test takers have.

Two others said they focused on listening for certain key words rather than on any certain visual component, thus indicating that they were not primarily focused on visuals:

*I focused on the questions in lectures, such as HOW, WHY, WHEN* (Participant 142)

*I cared about words like but, however…* (Participant 177)

Another two participants stated that they did not look at the video:

*I tried not to look at it because it was distracting.* (Participant 80)

*I avoided looking at it.* (Participant 174)

These two responses were particularly interesting. First of all, by stating this, these students indicate that they have developed an alternative strategy for listening in an academic listening environment. While this does not seem particularly effective for listening and taking notes in a real lecture, it may be quite efficient within the confines of a listening comprehension test. This also may indicate that some individuals understand that they are likely to have their attentional resources strained with the addition of video, so they choose to shut it out to focus only on the language that they are presented with. While this may not have many implications for test development since it is not realistic to force individuals to watch a video while listening (unless certain questions correspond specifically to information found in the visuals), it does give some further information about examinee behavior during the test.

Beyond these responses, a remaining 35 participants either provided unclear (and therefore unclassifiable) answers, provided off-topic answers, or provided no answers at all, thus they were not analyzed any further for this study.

**Note-taking preferences.** The final open-ended question, other than one asking participants if they were familiar with content from any of the lectures, asked examinees their

preferences regarding note-taking method. The numbers of participants preferring each method and the themes that arose from their reasons are presented in Table 4.15.

Most participants ($N$ = 143) preferred handwriting their notes to typing while only 52 test takers preferred typing. Of the remaining five participants, one said that he had no preference, and four did not answer the question. For those who answered that they preferred taking notes by hand, three main themes arose. One major theme was that respondents felt that their speed in note-taking was significantly increased when writing them out by hand due to lower typing ability (N = 52). The following statements are representative of this:

> *Handwriting because I don't like typing and I feel like the I write faster than typing. I have more training in school in writing down notes than typing.* (Participant 68)

Table 4.15

*Note-Taking Preferences and Themes Classifying Participant Reasons*

| **Preference** | **Theme** | ***N*** |
|---|---|---|
| Handwriting | | 143 |
| | Note-taking speed was increased | 52 |
| | Aided in memory of lecture material | 27 |
| | Provides a better platform for taking notes | 37 |
| | Greater comfort | 3 |
| | Typing is noisy | 3 |
| | Classroom policies | 2 |
| Typing | | 52 |
| | Greater speed and facility | 41 |
| | Easier to read later | 6 |
| | Greater comfort | 1 |
| | No Reason | 4 |
| Both | | 1 |
| | Dependent upon context | 1 |
| Off-Topic/No Response | | 4 |

*Handwriting. It is sometimes hard to catch up the lecture speed but I'm not a goo[d] typer* (Participant 126)

*Yes, I prefer handwriting because it is easier than typing because I'm too slow to type English.* (Participant 114)

*I think I am more used to handwriting than typing.* (Participant 161)

As can be seen here, each respondent makes reference to either their typing or handwriting ability in some way. Participants 68 and 126, and 114 each state that it was easier for them to keep up with handwriting because they are faster at handwriting or slower at typing. In the case of Participant 126, she stated that she found it easier to catch up by writing because she did not have the necessary typing skills, indicating that writing was faster for her. Finally Participant 161 was placed within this theme category due to her stated that she was more used to handwriting than typing, indicating that she could do so faster. These four responses are very typical of this theme, with all others serving as some variant to these with the same ideas expressed.

In addition to comments on speed and typing ability, many examinees also mentioned that they felt that handwriting aided in their memory of the lecture material (N = 27) and that they, therefore, preferred that method. Statements representative of this them are the following:

*I like to write down the things the teacher says that are not written on the power points. I feel I then will remember the topics easier.* (Participant 63)

*Handwriting, I can easily remember lots of informations, it is hard tasks though.* (Participant 144)

*I prefer handwriting, because it may help me remember better. But in the other hand, typing is more quickly than handwriting.* (Participant 37)

*Handwriting because it helps me focus more and remember better.* (Participant 188)

In order for the response to be classified under this theme, it had to mention memory being aided in some way. This is seen in the responses provided here, which are representative of the other responses found in this theme. For instance, Participants 63 stated that they could "remember the topics easier" or "easily remember" the information while Participants 37 and 188 stated that they could "remember better" by handwriting. Statements in this category either had these exact phrases or a very close variant of them. These comments were quite insightful in that they provided some participant confirmation of studies on handwritten note-taking that were reviewed in Chapter 2, but they were also surprising given the slight (though insignificant) advantage seen in typed note-taking seen from the analysis of research question 2 above, indicating the complexities of the issue at hand.

The last major theme expressed by many of the participants who preferred handwriting was that handwriting provided a better platform for putting information into notes (N = 37). Statements such as the following demonstrate the types of note-taking activities students prefer handwriting notes for:

> *By handwriting, because I have more freedom to draw and organize my notes compare to typing. I can only type vertically or horizontally when I use computer. Also, I can take notes faster when I handwrite.* (Participant 173)
>
> *By handwriting. I like to use Korean and English and draw some picture in my paper.* (Participant 113)
>
> *I think handwriting is better for me. I can organize important parts.* (Participant 66)
>
> *I prefer handwriting, I am not good at typing and in case of the typing I cannot use a[rr]ows or circle so handwriting is better than typing.* (Participant 62)

Any response that fell in this category made direct reference to handwriting allowing them to manipulate information in some way. These references related to being able to draw arrows, used their native language, and use clearer organization. The comments related to the use of the L1 were quite telling in relation to the usefulness of handwriting as opposed to typing. It is obvious from these comments that translanguaging may play a key role in the preference of handwriting over typing, since the typing condition in this experiment did not provide any method for switching between languages. It may be worth examining how preferences are affected by the ability to switch between language typefaces in the future to see if this ability would have any significant effect.

Beyond these themes are several others that are not necessarily representative of the vast majority of responses, but are still important to consider moving forward. These themes have to do with comfort, noise, and classroom practices. For instance, in relation to comfort, one student said:

> *I feel more comfortable by handwriting. It is easier to control than typing, but typing is faster than handwriting.* (Participant 145)

If a response was placed under this theme, it specifically mentioned the word "comfort" or some variant of it. Two other respondents provided some variation of this response. Noise was also an unexpected theme mentioned by three different students. One of them said the following in relation to noise with the other two being quite similar:

> *I am not good at typing and if everyone types, there might be more noise. Noise distraction from others may be another factor.* (Participant 198)

Finally two other participants stated their preference in terms of two different somewhat external issues:

*Handwriting. Some of teacher hate typing, so I have never used computer in the classroom.* (Participant 22)

*By handwriting. Because I don't like to bring my heavy computer to the class.* (Participant 84)

In these students' cases, it is clear that they have either been conditioned by certain teachers to cope without typing notes, or they prefer the convenience of not having to carry extra weight around with them. Thus, to them, handwritten notes are more convenient and easier to handle for the sake of consistency since it would be easier to handwrite notes for all teachers than to switch between both methods depending on the preferences or certain teachers.

For those who expressed typing as their preference, two primary themes arose across respondents. The most popular theme, expressed by 41 participants to some extent, was related to speed and facility. The following are characteristic responses displaying this theme:

*Typing is much faster than handwriting. So I can catch the main topic.* (Participant 171)

*By typing. Because I need to erase the miss words by hands in handwriting and it is slow. I might miss to listen the following lecture in erasing them. In addition, for typing it is easier to organize the structures.* (Participant 133)

*Typing is easier for returning to the adding of information I learned later on in the lecture.* (Participant 96)

*Typing. It makes it easier for me to focus on the lecture or screen.* (Participant 130)

As seen above with the responses that preferred handwriting due to speed, responses for a typing preference due to speed of note-taking consist of the same ideas and key words. For instance, Participants 171, 96, and 130 state that typing is "much faster" or "easier" for different reasons related to ease of adding information or focusing on the screen. Additionally, as is the case for

Participant 133, some participants commented that typing made it possible to avoid having to go back to erase content making note-taking faster since it is easier to select and press the delete key.

Another less common theme related to typing was mentioned by 6 participants was related to clarity of notes. Two statements representative of this theme are the following:

> *I prefer typing them because I can read it clearer due to my bad hand writing and I don't have to worry about losing it.* (Participant 38)
>
> *Typing, because I'm a faster at typing than I am at handwriting, and the notes look neater.* (Participant 30)

Here it is possible to see that responses falling in this category referred to poor handwriting or "neater" notes due to the typeface. Thus it is clear that some students recognize that their handwriting may serve as something that may inhibit their performance later one, which they may not be able to address while trying to understand that listening passage and take notes.

Finally, one student mentioned that both methods were good for note-taking, but that it was dependent upon the context in which notes were to be taken. He stated the following:

> *It depends on the class. If the class is in small class, I would hand write, however if the class is in lecture style, I would prefer to take notes by typing. Because in small class, it is easy to interrupt class to ask the questions or speak out the opinion which means I think that the class goes slower than lecture style, so I am able to take notes well.* (Participant 122)

This response proved interesting because of the way the examinee essentially compartmentalized their preferences into different situations, indicating that even though someone may have specific

preferences, these preferences are not necessarily absolute in nature, thus suggesting that there may be a need for greater adaptability of testing conditions in the future.

**Summary of research question 4 findings.** Overall, survey data reported that the majority of participants preferred the video-mediated listening passages, finding them helpful for recall of lecture details and saying that they helped in some way. This was the case even though many of them felt that concentration was hindered in some way due to the presence of the video due to the fact that they could be more easily distracted by what was happening on the screen. Some of this can be explained by what the respondents said they focused on during the video lecture. For instance, while some respondents indicated that they used the visuals to get back on track and to help them understand key points, many stated that they simply only focused on writing down what was on the slide or simply found it too difficult to attend to multiple input sources at once. Because of the mixed preferences and affects that these responses provided, they help to explain the lack of overall effect on the test scores seen in the previous analyses by showing that there are a range of effects (both positive and negative) that visuals can have on listening comprehension. Yet another interesting finding from these responses indicates that the types of visuals and the content of the lectures may also have some impact on overall effectiveness of visuals and test scores. It was seen from responses that some visuals (mainly pictures) were much easier to recall while some topics were viewed as being more interesting and were therefore easier to remember. These issues are brought up in more detail in the following chapter in relation to their impacts on test development.

Additionally, it was found that for note-taking, the majority of participants preferred to take notes by hand rather than by typing due to its being easier to switch languages and organize information. However, a reasonable number of individuals still stated the opposite for typing,

saying that they preferred typing for many of the same reasons that those who preferred handwriting said that they preferred that method. These mixed preferences indicate that typing may have had less of an effect in previous analyses due to participants' lack of experience with typing notes on a computer, which may have minimized any effect the typing had. In addition, responses related to translanguaging while handwriting notes indicate that students are more accustomed to switching between English and their L1than they are to being restricted to one language, which they may be forced to do if their L1 uses a writing system incompatible with the keyboard used in the language of the test. This has implications for the development of tests that may want to incorporate typed note-taking in their administrations in the future. Taken together, the results from survey questions related to both the video and typed note-taking conditions yield information that provide further explanation for the results obtained from earlier analyses seen in this chapter. These results are discussed in tandem with the results from the other three research questions in the following chapter.

# CHAPTER 5
# DISCUSSION AND CONCLUSION

This chapter discusses the findings described in the previous chapter and draws conclusions from the data in order to put forth an argument urging others to consider redefining the listening construct as it currently stands within tests of listening comprehension. The chapter first provides possible explanations for the effects (or lack thereof) that input and note-taking conditions had on test taker performance. Then the discussion turns to the implications that the results from this study have for how the listening construct is defined, making the argument for the need to redefine it. Finally, the chapter ends with a discussion of the limitations of the study and provides possible future lines of inquiry that can lead to further understanding of the best methods for testing listening comprehension.

**The Role of Visual Input in Listening Comprehension**

While it was hypothesized that test takers' scores on the listening assessment conducted in this study would be significantly different between video-mediated and audio-only input conditions, this was, overall, not the case. Statistically, scores on exams remained the same, agreeing with studies conducted by Gruba (1993), Baltova (1994), and Londe (2009). At the same time, these results conflict with those finding that video-mediated listening passages have the potential to either help or hurt performance on comprehension assessments (Coniam, 2001; Suvorov, 2009; Brett, 1997; Chung, 1994; Wagner, 2010a, 2010b). The difference in results obtained by all of these studies may be related to several different factors. While differences in findings may be related to issues in study design that were described in Chapter 2, it is also possible that the tests may have been testing different arrays of listening skills. The findings of the path analysis in this study showed that visual input may not have had a significant effect on

question types focusing on skills associated with making inferences or understanding the gist of a listening passage; however, they did indicate that there may be a significant, positive effect on questions associated with identifying details within the listening passage when video-mediated listening passages are used. This indicates that some of the effect may be lost when all question types are simply pooled together, which is something that should be considered when defining the target use domain and construct of academic listening. This should definitely be kept in mind since some studies have found item bias towards certain visual conditions (Batty, 2014) and this study found several items that had potential for expressing similar bias.

The Rasch analysis also showed, somewhat contradictorily to the path analysis, that the overall effect of video-based listening passages on items was to make the items slightly more difficult relative to audio-only listening passages. However, this difference was less than one logit, so, while it indicates a small effect of video on item performance, it cannot necessarily be considered as a significant factor. Even though this effect appears small, the question still remains as to how the presence of video provides better overall performance on detail questions while still making the items slightly more difficult overall. Upon examining participants' responses to the post-test survey, it became clearer how this could be the case. Based on the qualitative data collected for research question 4, participants' opinions seemed to reflect these findings. While the vast majority felt that the presence of content visuals aided in comprehension because of emphasis added through the lecturer's voice or gestures and because of the PowerPoint, some of these individuals also went on to say that it did make it more difficult to concentrate. For instance, participant 5 stated the that the content of the PowerPoint helped to emphasize key points and facilitated comprehension, he also stated that the lecturer herself was somewhat distracting because he felt like he had to look at her, making it more difficult to listen

for the necessary details from the lecture. Additional comments provided insight into what the examinees were focusing on in the video, with participant 101 (quoted in Chapter 5) stating that she was actually distracted by context visuals that were not related in any way to the content of the lecture. These findings appear to support previous research in the role of visuals in listening comprehension (Ginther, 2002; Wagner, 2010b) while also indicating that context visuals do play a role in video by having the potential to distract from the task at hand.

In addition to the role played by the content and context visuals found in the listening passages, it was also clear from the qualitative analysis that there were issues related to the cognitive burden placed on the participants in the video condition. In line with what was observed by Chung (1994) and Cubilo and Winke (2013), several participants reported that the video passage was faster than the audio passage, regardless of which form they had taken (i.e., some taking form B with video said it was faster than the form A audio-only passages and some taking form A with video said it was faster than the form B audio-only passages). This would appear to indicate that certain participants are falling prey to a split attention effect observed by Wagner (2008) and described by Sweller and Chandler (1994). Under this theory, when cognitive load is excessively burdened due to the need to integrate a number of different sources of information into one cohesive whole, this essentially results in breakdowns in processing of information. In this case, the video would seem to be having a distracting effect on certain participants that is causing them to perceive time in the video as advancing more rapidly than it is due to the inability to effectively integrate and process all relevant information. Thus, the presence of video creates a situation in which test takers need to not only attend to different modalities (i.e., visual and audio), but they also need to attend to different task components (i.e., comprehension and note-taking). For those who may not have the necessary level of automaticity

in their language use, this may prove a formidable task, resulting in perceptions of audio being faster in video conditions, being distracted by irrelevant visuals on the screen, or even focusing too much on relevant elements without attending to the auditory stimuli necessary for comprehending the lecture and answering questions about it.

Beyond the potential processing strains participants may have been referring to, it is clear that the videos were still helpful several ways. In particular, participants mentioned that the slides provided on the screen as the lecturer was speaking were helpful for determining key words and ideas while also allowing them to get back on track if they found themselves lost over the course of the listening. In addition, several participants mentioned that the gestures and accented speech of the lecturer helped to clue them in to important points, and that her lip movements helped them to better discern some words. Comments such as these indicate that while listening, people use these details either consciously or subconsciously to aid in their comprehension, agreeing with findings by Sueyoshi and Hardison (2005). In addition, these comments also show that L2 listeners use these signals to some extent, though the extent to which they are used is still under investigation and has been the subject of research involving eye tracking (Suvorov, 2013; Wagner, 2007). Furthermore, when asked if video helped recall later on, participants pointed to many of these elements positively in affirming that they had aided in recall of information. While video doesn't appear to have affected inferential questions or gist questions directly, it does appear to have an indirect effect on them through aiding in the ability of examinees to recall details. Based on examinees' accounts, it sounds like much of the content visuals found in the slide were directly related to recall of specific details within the lectures that they listened to; thus it would make sense that these visuals would lead to positive effects on performances on these questions. Likewise, it is not possible to make adequate inferences

without fully understanding the main details of a passage, so it would appear that the primary role of visual information in a listening passage is to aid learners in understanding necessary details so that they can go on to better determine the appropriate inferences that they need to make and have a better sense of the overall purpose of a particular passage.

Taken together, the results of this study indicate that visual input plays a complex role in listening comprehension. It seems that while visual input is viewed predominantly as an ally among test takers, it still poses some challenges even for those holding positive view. Thus, the inclusion of visuals presents itself as a double-edged sword that has the potential to both aid in comprehension and distract, requiring examinees to adequately process both aspects in order to perform well. Based on the findings of this study, since there are no major overall differences between test scores of audio-only and video-based listening passages, it may be that the interplay between helping and distracting forces of the video are counteract each other enough to make video-based listening passages similar in effect to the challenges faced by audio-only listening passages.

**The Impact of Note-Taking Medium on Listening Comprehension**

The hypothesis that note-taking would have significant impacts based on the medium that examinees were required to use and that typing would be the preferred note-taking medium was shown to be incorrect. The ANOVA showed that there was neither a significant main effect of note-taking conditions on test scores, nor was there any significant interaction between these conditions and audio-visual conditions. Additionally, whereas the Rasch analysis showed that there was at least a slight difference in item difficulty created by input conditions, note-taking conditions showed a miniscule effect on item difficult and none of the bias values flagged in chapter 5 were related to note-taking conditions, showing that individual items did not seem to

be affected by these conditions. This result was surprising given the prevalence of laptop use in schools and lecture halls in many universities. It is not uncommon to enter a lecture hall and see at least half of the students present typing notes on a computer, and, therefore, it was believed that providing a medium that students might be more accustomed to using in the classroom would provide some gains in test performance.

In addition to the finding that students did not seem to perform significantly better in one note-taking medium over the other, it was discovered that the majority of students actually preferred handwriting to typing in this study for various reasons. The reasons provided indicated that students found it easier to manipulate and organize information by handwriting, write notes in their native languages, and remember information from the lecture. However, since there were no significant effects of note-taking medium overall, these reasons may just indicate that the preferred note-taking medium adds a level of comfort to the test taker as they attempt to navigate the requirements of the listening task. Indeed, this would be supported by the qualitative data to a degree since some of the participants actually commented on being more comfortable when either typing or handwriting.

The results for this portion of the study are also surprising for another reason. Previous research has shown extensively that the medium through which students take notes has significant impacts on later recall and performance. Several studies reviewed earlier in Chapter 2 found significant differences in performance measures dependent upon whether students typed or handwrote their notes. Mueller and Oppenheimer's (2014) study showed that performance on conceptual questions was worse when notes were typed, stating that typed note-taking resulted in shallower processing due to typing verbatim what the lecturer says rather than focusing on key details. However, based on the findings in the current study, it may be possible that there is a far

more complex interaction of variables at play. For instance, while technology may result in shallower processing, it is possible that distractions play an equal role in poorer performance (Fink, 2010; Stacy & Cain, 2015). A number of potential explanations can be found for the fact that there were no significant differences found between note-taking requirements in this study.

One potential explanation for the lack of significant differences in light of previous research is that the differences are not as applicable for short-term storage and access. In a testing situation, learners are expected to listen to a passage and then immediately answer questions related to the passage. This would lead to less decay since the examinees are expected to retain and use this information over a short period of time. In addition, generally the purpose of taking notes is to be able to review them later so that one can better internalize the information and potentially transfer it to long-term memory for later use on a test through active learning (Voss et al., 2011). Test takers in situations such as the one presented in this study do not allow for such active review of notes, so differences between conditions may not be readily observed, especially because the entire act of listening and taking notes and answering questions took place over the course of 15 minutes for a single lecture.

Another possible explanation for the lack in significant differences between typing and handwriting may be related to the actual quality of the notes. Regardless of whether examinees felt more comfort in one condition over the other and regardless of whether handwriting allows for deeper processing of material presented to listeners, the ability to correctly recall information and answer questions will still depend on the quality of the notes, as has been demonstrated by several studies (Chaudron, Loschky, & Cook, 1994; Song, 2011). If students are not sufficiently trained in how to take effective notes and, therefore, simply write down everything that they hear instead of key points, they may not actually experience the benefits of the potentially deeper

processing allowed by handwriting notes. If enough of the participants lacked such knowledge in this study, the potential benefits of one method over the other would have been rendered unobservable.

Finally, one other potential reason for the relative equality in performance between handwritten and typing conditions could have been related to the actual condition of typing. As was mentioned above, Fink (2010) and Stacy and Cain (2015) mentioned that typing notes may be less efficient due to distractions afforded to students on their laptops. Test conditions restrict how participants could actually use computers while listening to the lecture, making it so that possible distractions on the computer that were not related to the actual video were removed. It could be that one of the influences responsible for poorer performance in other studies has been neutralized within the test environment. Therefore, this may be an additional factor contributing to the absence of a significant difference between typed and handwritten note-taking conditions. Without the internet available to provide distraction as it would be on one's laptop within a lecture hall, the participants would have nowhere else to look except for either the lecture on the computer screen the scenery provided by the computer lab.

The fact that participants preferred handwriting to typing was quite unexpected; however, comments made by participants in the survey served to clarify this preference to a certain extent. Many students, as mentioned earlier, found that handwriting was faster and allowed more freedom overall than typing and claimed that it aided in memory (whether or not this was actually the case given that it did not seem to have much of an impact on scores overall). Therefore, this preference seemed to arise from an issue of comfort and convenience. However, the question still stands as to how participants who were predominantly in their early- to late-20s would not be comfortable with typing. One possible explanation for this can be found within the

survey. As Fink (2010) demonstrated in his publication, some instructors choose to ban the use of laptops in the classroom in order to create an environment in which students are more likely to pay attention to what the lecturer is presenting. One comment from the surveys expressed this idea, stating that their teacher had also banned the use of laptops, thus making it necessary to simply grow accustomed to writing out notes by hand and forego the use of a laptop. Without practice typing notes, students would assuredly not gain strategies for efficiently taking and organizing notes using a computational user interface.

Reasons for this preference may fall in two other possible categories. One of these may have something to do with the rise of handheld devices. As handheld devices have become more prevalent, the use of physical keyboards may be less common among younger generations who are more accustomed to communication through text and chat. Therefore, proper typing form that would make typing faster than handwriting may not be known. This would at least partially explain why many thought that handwriting was faster. The other reason (and probably the one that holds much more sway on speed and efficiency) is the fact that these examinees are operating in their L2. Several test takers complained that they were not able to use their first language while typing, making it more difficult to take notes on the computer. Therefore, typing may have been viewed more favorably by participants if they had not artificially been restricted to using their L2 while listening, as this most likely hindered them. Perhaps future administrations with typed note-taking conditions could allow for switching between languages since this is a rather standard feature of many word processors now.

Overall, the impact that note-taking condition had on actual performance measures was negligible. However, what can be seen from the data is that it is clear that participants do express certain preferences for how they take notes based on their classroom experiences. Additionally,

while these preferences do not necessarily translate into changes in performance on listening measures, they do contribute to the overall levels of comfort experienced by test-takers within a testing situation. Such issues may be important to consider for the way in which the TLU domain and listening construct are defined, a topic turned to in the next section.

**Implications for Defining the TLU Domain and Listening Construct**

As defined earlier, the TLU domain refers to contexts outside of the test in which individuals are required to perform certain tasks (Bachman & Palmer, 2010). Ideally, when writing tests, the TLU task should be generalizable beyond the test and represent performance within the real-life domain. In the case of this study, the TLU domain would be a lecture hall or classroom in which students are presented with lecture materials and expected to comprehend them to an extent that would demonstrate that they have the listening comprehension ability to perform well in an academic context. A number of factors come into play in this context that the students must master. For instance, they must be able to demonstrate their ability in utilizing different listening subskills, such as identifying the gist, or main idea, of a listening passage, identifying important details, and making inferences. These skills have been well documented by studies and are relatively common among different taxonomies (Anderson, Krathwohl, &Bloom, 2001; Bejar et al., 2001; Lund, 2008). The TLU domain is not only represented by the necessary skills that students are expected to show a certain level of ability in, but it is also represented by the contextual elements associated with the TLU task. In the case of academic listening, this involves the contextual visuals associated with a lecture hall, such as the lecturer and the slides used, and the ability use these visuals as a source of listening support.

Finally, another aspect contributing to performance within this domain relates to the students' note-taking ability. Not only can this be related to the need of students to navigate

instructor limitations that may create a situation in which students may not be allowed to use electronic devices in the classroom, thus forcing them to be reasonably acquainted with using a different modality for note-taking, but it can also be related to the need of students to demonstrate sufficient note-taking skills, signaling that the learner can focus on key details of the lecture and ignore points that may be tangential or even simplistic in nature. Taken together, these skills should represent the TLU domain for an individual performing an academic listening task in a real-world domain. As such, when developing a test, the tasks should mirror these features in such a way so as to produce measures that can be representative of performance within the real-world domain. As a result, it is clear that the ideas of the TLU domain and construct validity are closely related to each other. The present study has produced a series of findings based on tasks meant to mirror the real world domain in what ever way possible that have several implications for the way in which the listening construct is defined.

One of the main effects examined in this study was the impact that visual input had on score outcome. While there were no significant differences in performance between input conditions, several findings did present themselves and suggest that visuals may have a role in how the listening construct is defined. For instance, the differences in path analyses between forms A and B were somewhat surprising. If detail questions were found to be directly and significantly influenced by input condition on one form, it would stand to reason that they would be equally affected on the other form. However, this was not the case. This may have been due in part to the actual subject matter between the two tests. Upon examining the two forms, it is possible that the topics found in form A may have been more abstract in nature than those in form B. For instance, while the social science topic in form A was related to the development of a three-language policy in India, the social science topic in form B was about the concept of

choice. What was different about these two was that the choice topic discussed more studies comparing differing opinions between countries of what choice means to them while the language policy topic discussed more in the way of potential risks and benefits of such a policy. This led to a topic on form B that was more amenable to including graphs and figures as opposed to the form A topic, which was better represented by key terms. Some participant comments in the post-test survey pointed to this as a possible explanation for differences since some test takers said that that more pictures would have increased memory retention since they did not generally remember the words.

This finding has implications for the way in which the listening construct is defined. While previous studies have primarily looked at the effect that video-based input has on overall scores for a listening test, few, if any, have examined the effects it has on different subskill types. The fact that this study found that visuals significantly contribute to detail comprehension in some situations signals that the impact of visuals may be more refined that previously envisioned. If the listening construct is going to be defined in such a way as to make it so that listening passages are only associated with still pictures of the lecturer or of an object that the lecturer is talking about, then it is necessary to limit this type of test only to subjects that do not see any underlying effects from video-based input. Rather, what seems the safer option is to redefine listening tasks by including the video-based component. In this case, those subjects that do not experience effects from video will produce similar output regardless, while those topics that lead to benefits from its inclusion are allowed to produce these benefits. To remove video-based listening from topics that are affected by it only serves to threaten the construct validity of the test through construct under-representation. That is, if a known factor in listening comprehension is removed, the results of the test will be less representative of the construct

being investigated since a part of it will be missing. As a result, any interpretation and use for the score will come into question, leading to a lack of confidence in the validity of the test. As it stands now based on the results of this study, many listening tests seem to be under-representing the construct through a lack of adequate visuals, calling into question the validity of their scores since it is not possible to fully know at this time what topics are more likely to be influenced by video-mediated input.

Video-based input also led participants to state a number of other opinions related to the helpfulness of the visuals and the authenticity of the lecture. Even though the slides were constructed to actually contain less information than what one would generally see on slides in a lecture in the real-world domain, participants overwhelmingly stated that they were helpful. In addition, several participants were also clear in stating that they felt that the lectures were more realistic with the video and that it reminded them of attending a real lecture. These findings have several implications. The first of these is that it is clear from the comments, once again, that visuals do play a role in comprehension. Participants made note of both the information on the slides and the lecturer's gestures, body language, and lip movements. This indicates, as previous studies have also done (Sueyoshi & Hardison, 2005), that visual and aural processing are interconnected and that divorcing the two would lead, once again to construct underrepresentation within a listening task. Additionally, comments made by participants related to the visuals providing greater authenticity not only help to provide confirmation that the TLU domain is being adequately represented, but also help to establish face validity of the test.

One other result related to the presence of video in the listening passages that has implications for the construct validity of listening tests was the fact that many participants found elements of the video distracting. Discussions related to the role of video in listening assessments

found in Chapter 2 mentioned that some have stated that listening is solely an auditory event and that inclusion of visual input unfairly disadvantages those who are not able to utilize the visual input in an effective and non-distracting way. However, the results from this study could be used to argue the opposite: not including visuals on a listening test, particularly video-based visuals, unfairly disadvantages examinees who cannot use them effectively because their absence does not truly reflect their listening ability within a lecture hall. Students must be able to navigate multiple modalities while listening to a lecture in the classroom. They cannot simply decide to turn visuals off in real life and only listen to the sounds the lecturer is making unless they close their eyes or record it and listen to it later, which is hardly a realistic way to learn. Therefore, if the test does not represent this aspect of listening, it does not truly indicate their ability to perform listening tasks in the lecture hall. This in turn could lead to inaccurate decisions that may negatively impact students' future performance.

Thus, in order to ensure appropriate construct representation within listening exams, it would be best to ensure that visuals are present where one would be accustomed to encountering them at the risk of inaccurately determining that a student is ready for lecture-based courses when they may not have developed the ability to process input in such a way as to attend to aural input and the visual components that support it while looking past the rest. Test developers need not provide excessive visuals at the learner to ensure that they are able to do so (in fact, doing so would serve to produce potentially unrealistic errors in performance due to split attention affects that even L1 learners may face), but they should include enough to avoid construct misrepresentation.

Finally, while results related to video-based input seem to provide the greatest number of implications related to the listening construct, note-taking conditions also had an implication

worth examining. While the findings for his study indicated that there were no significant differences between scores across note-taking conditions, participants were quite opinionated regarding their actual preferences for note-taking. On the one hand, many participants preferred to take notes by hand, making it so that the current means by which most listening tests require learners to take notes is fair enough for most of the population. On the other hand, however, there were still those who were more accustomed to typing their notes and who stated that they preferred to take notes in this manner because they were better and faster at doing so. This could call into question the way in which the TLU domain is currently conceptualized within academic listening tasks. Even though there were no differences in performance, the fact that many participants stated that one method led to greater comfort than another suggests the need to address this issue. Tests already present themselves as a stress-inducing event. To deprive examinees the comfort of doing something in a way they would do it in the TLU domain poses potential challenges to the current definition of the TLU domain and, by extension, the construct validity. It may be worth considering whether future test development will take this issue into account and how it can be addressed, though such initiatives may be difficult since they could potentially require that students have constant access to the ability to switch between language typefaces quickly during the test, which may be a difficult feat to accomplish given the number of L1 backgrounds test takers come from.

Although many of the results from this study did not indicate significant overall effects of either input or note-taking conditions, the qualitative data helped to shed light on the views that participants had regarding the way in which these conditions were representing actual classroom activity and how visuals and note-taking conditions were being used by students in their efforts to listen to and understand the lectures and answer questions. Hopefully future test development

initiatives will take these issues into account and consider redefining the current conceptualization of listening by moving beyond still pictures or blank screens and moving towards video-based listening passages that take into account the preferences of test takers in relation to their note-taking practices.

**Limitations**

There are several limitations in the current study. The first of these is related to issues related to the overall design. Due to the crossed nature of the design and the need for two different test forms, direct comparisons in performance between different conditions was not completely possible. This led to the need to randomly arrange participants into one of four groups in order to answer the first research question. While this randomization most likely resulted in accurate findings, there is still the potential that findings may have been inaccurate since not all of the participants were being compared in each condition and its related effect on the overall test score. Additionally, while it is possible to gather evidence to determine whether two different forms are parallel, it is not always possible to know with absolute certainty that they are parallel. For the tests in this study, the two test forms provided data that appeared to be similar in many respects, but due to both a lack of resources and a limited pool of participants, it was not possible to entirely assess the degree to which topics were equivalent, nor was it possible to pilot the exam past the first round of 20 students. This did end up appearing to work out fine in the end, but future use of the test would need to be preceded by further piloting to ensure topic choice is appropriate and items are truly equivalent overall.

A second limitation to this study was in relation to the distribution of questions for the subskills in each form. While each form did have a number of items representative of each subskill, they were not as even as would be ideal. This was due to certain topics being more

applicable toward some listening subskills than other topics. Although this is not a major cause for concern for the results in the present study since the most uneven distribution of subskill types was seen in detail type questions (which accounted for the largest portion of question types), it would be ideal to ensure that equal numbers of questions represent each subskill if test forms should be equivalent.

Another limitation seen in this study is in relation to the video quality. While the videos were clear, the size could have made it potentially more difficult to see some of the content images that were provided on the lecture slides, which could have muted some of the effects on test score outcomes. This was the result of somewhat limited resources for videotaping the lectures and, of course, the screen size of the computers that were in the computer lab. Participants did not seem to comment on this as being an issue for them, so it may not have had the impact that it potentially could have had. Future administrations would most likely want to determine a better method for video delivery and recording.

Finally, one last limitation was related to the collection of qualitative data for the purposes of the mixed methods design in this study. Because participants were all students enrolled either in a full course load in either an English language program or in both English language classes and university classes, their time was severely limited for participating in this study. The survey method for collecting qualitative data was selected because it allowed for a more rapid collection of data that did not take much more additional time from the participants than what was already being asked. While the open-ended survey responses were informative and did provide some helpful explanations of the results, other formats of qualitative data collection would have provided much richer data that may have helped uncover better understanding of the processes going on in the examinees' minds while listening and taking

notes. Studies such as those done by Goodwin (2017), Wagner (2007), and Suvorov (2013) have made use of some combination of semi-structured interviews, stimulated retrospective recall, and surveys for listening tests, and this has led to data that is much richer and allows for better explanation of the internal processes that occur within the test takers' mind. However, since participants were already being asked to spend two hours of their time participating in this study and due to a number of other extenuating circumstances, the use of multiple qualitative methods was not wholly possible for this study. Additionally, while participants' notes were collected for this study and could have elucidated some of the survey comments related to using multiple languages to take notes, this analysis was not performed for the present study due to the comments related to translanguaging being unexpected and the inability to fully examine content written in the participants' L1s given the large number of languages represented. However, such measures will be used in future examinations of this issue.

**Future Research**

Based on the limitations and the findings of this study, several possible avenues can be taken in future research on this topic. One major line of research that should be further pursued is the investigation of the role content visuals play in performance outcomes related to different listening subskills. Few if any studies have actually sought to examine this, and, should they play a significant role in certain subskills as they appear to do so from the results of this study, this would have major implications regarding the question of whether to incorporate video-based passages on listening assessments.

In relation to investigating the role visuals play in subskill performance, another line of future research would be to investigate the interplay between visuals and topic and whether the visuals are primarily of textual or pictorial origin. Several participants in the present study

mentioned that it was easier to recall pictures than it was to recall words from the lecture slides. Thus, it would be worth determining if the presence of pictures does provide an edge to comprehension scores. Additionally, it may also be true that certain listening passage topics are more susceptible to having visuals impact the listener's comprehension. It has already been discussed in the present study how certain topics were more readily associated with picture-based images in the lecture slides than others. Understanding the interplay between topic and visual presentation would make for a more standardized method by which to construct listening test passages that utilize video-based listening material in the future.

Another possible area of future research in this area would be related to eye tracking and the way in which individuals interact with the visual component of the listening passages and the impact this interaction has on note-taking and test score. While several studies have already been conducted in relation to video-based listening passages and what test takers focus on (Suvorov, 2013; Wagner, 2007), there has not been much in the way of investigating how handwriting notes detracts from the examinee's ability to look at the visuals on the screen and the relation of these eye movements to test scores is still not fully understood. Thus, more investigation in this area would be beneficial. In relation to note-taking in particular, this could be useful for seeing if benefits exist for typed notes over handwritten notes since the ability to keep one's eyes on the computer screen may be able to lead to individuals looking away from the visuals less.

In addition to investigating the impact typed notes may have on eye movement in relation to video-based listening passages, investigating the interactions between note-taking preferences, ability, and listening performance when typing and handwriting notes would be beneficial. Results from a study such as this would have implications for the method by which listening assessments are administered to examinees, providing greater clarity in how note-taking

preference and practices impact the score of an individual. For instance, if an individual appears to be scoring lower when handwriting than when typing when they state that their preference for taking notes in class is to type, this would point to flaws in the validity of the scores and possibly lead to the need to redefine the TLU domain and/or the construct of academic listening.

Finally, one last area of research that would be beneficial is related to the comments provided by participants in this study related to the comfort or relaxation they felt in certain conditions. Researchers such as Arnold (2000) and In'nami (2006) have investigated and found conflicting results related to the effects that test anxiety can have on listening comprehension. The comments made by participants in this study indicated that general feelings of comfort may have an impact on listening comprehension in some manner. Therefore, it may be worth more carefully considering the affective dimensions of including video-based listening passages or typed note-taking options on exams in the future, as anxiety and discomfort should be kept at a minimum in testing environments in order to reduce error and obtain valid and trustworthy results.

**Conclusion**

This study investigated the effect that input and note-taking conditions had on the English listening comprehension scores of international students in the United States. Data were obtained by having students take two different forms of a listening exam under each of the possible conditions in a crossed research design, and by asking them to complete an open-ended survey regarding their perceptions of the different conditions after completing the exam. Through the use of various analyses and the inclusion of qualitative responses, the study employed a mixed methods design in order to triangulate data and provide richer and more robust interpretations of the results. While results were overall not significant in relation to the effects that input and note-

taking conditions had on test scores, it was found that these conditions did offer something to participants in the way of comfort while taking the exam. In addition, it was found that visual input had a significant contribution on detail type questions, which indicates that it may provide an indirect effect to other listening subskills as well. These findings all provide important implications for future test development and are indicative of future lines of research that should be conducted in order to better understand these conditions so that better tests that are more representative of the construct they are trying to measure are created.

# REFERENCES


Anderson, J. C. (1990). Testing reading comprehension skills. *Reading in a Foreign Language, 6*(2), 425-438.

Anderson, L. W., & Krathwohl, D. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. New York: Longman.

Anthony, J. J. (2009). Classroom computer experiences that stick: Two lenses on reflective timed essays. *Assessing Writing, 14*, 194-205.

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

Asl, Z. A., & Kheirzadeh, S. (2016). The effect of note-taking and working memory on Iranian EFL learners' listening performance. *International Journal of Research Studies in Psychology, 5*(4), 41-51.

Ayres, P., & Cierniak, G. (2012). Split attention effect. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3172-3175). New York: Springer.

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

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford: Oxford University.

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556-559.

Baker, L., & Lombardi, B. R. (1985). Students' lecture notes and their relation to test performance. *Teaching of Psychology, 12*(1), 28-32.

Baltova, I. (1994). The impact of video on comprehension skills of core French students. *Canadian Modern Language Review, 50*(3), 507-531.

Batty, A. O. (2014). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing, 32*(1), 3-20.

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

Bloomfield, A., Wayland, S., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension* (Technical Report No. E.3.1 TTO 81434). College Park, MD: University of Maryland, Center for Advanced Study of Language.

Bodie, G. D., Janusik, L. A., & Valikoski, T.-R. (2008). Priorities of listening research: Four interrelated initiatives. A white paper sponsored by the Research Committee of the International Listening Association. Retrieved from http://www.listen.org/WhitePaper

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.

Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System, 25*(1), 39-53.

Brown, J. D. (2001a). Point-biserial correlation coefficients. *Shiken: JLT Testing & Evaluation SIG Newsletter, 5*(3), 13-17.

Brown, J. D. (2001b). *Using surveys in language programs.* Cambridge: Cambridge University.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Brown, J. D. (2014). *Mixed methods research for TESOL.* Edinburgh, UK: Edinburgh University.

Brown, J. D., Trace, J., Janssen, G., & Kozhevnikova, L. (2016). How well do cloze items work and why? In C. Gitsaki & C. Coombe (Eds.), *Current issues in language evaluation, assessment, and testing: Research and Practice* (pp. 2-39). Newcastle upon Tyne, England: Cambridge Scholars Publishing.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage Publishers.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293-332.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York: Routledge.

Chaudron, C., Loschky, L., & Cook, J. (1994). Second language listening comprehension and lecture note-taking. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 75-92). Cambridge: Cambridge University.

Chung, U. K. (1994). *The effect of audio, a single picture, multiple pictures, or video on second-language listening comprehension*. Unpublished PhD dissertation, University of Illinois at Urbana-Champaign.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cubilo, J. & Winke, P. (2013). Redefining the L2 listening construct with an integrated writing task: Considering the impact of visual-cue interpretation and note-taking. *Language Assessment Quarterly, 10*, 371-397.

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

Davidson, F., & Lynch, B. K. (2001). *Testcraft: A teacher's guide to writing and using language test specification.* New Haven, CT: Yale University Press.

Davies, A. (Ed.). (1997). Ethics in language testing. *Language Testing, 14*.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods.* New York: Praeger.

Desnoyers, L. (2011) Toward a taxonomy of visuals in science communication. *Technical Communication, 58*(2), 119-134.

Dunkel, P., & Davy, S. (1989). The heuristic of lecture notetaking: Perceptions of American and international students regarding the value and practice of notetaking. *English for Specific Purposes, 8*(1), 33-50.

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*. Frankfurt, Germany: Peter Lang.

English, S. L. (1982, May). Kinesics in academic listening. Paper presented at the 16th annual convention of Teachers of English to Speakers of Other Languages, Honolulu, HI. (ERIC Document Reproduction Service No. ED 218 976).

English, S. L. (1985). Kinesics in academic lectures. *The ESP Journal, 4*(2), 161-170.

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

Fink, J. L. (2010). Why we banned use of laptops and "scribe notes" in our classroom. *American Journal of Pharmaceutical Education, 74*(6), Article 114.

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.

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, UK: Wiley-Blackwell.

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher, 18*(7), 4-10.

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

Goodwin, S. J. (2017). *Locus of control in L2 English listening assessment*. Unpublished doctoral dissertation, Georgia State University, Atlanta, GA.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7-22.

Greene, J. C. (2011). The construct(ion) of validity as argument. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice, new directions for evaluation* (pp. 81-92). San Francisco, CA: Jossey-Bass.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal, 15*(1), 85-88.

Gruba, P. (1999). *The role of digital video media in second language listening comprehension.* Unpublished PhD dissertation, Department of Linguistics and Applied Linguistics, University of Melbourne. Retrieved December 21, 2016, from http://eprints.unimelb.edu.au/archime/00000244/

Gruba, P. (2006). Playing the videotext: A media literacy perspective on video-mediated L2 listening. *Language Learning & Technology, 10*(2), 77-92.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage Publications.

Hadar, U., Wenkert-Olenik, D., Krauss, R., & Soroker, N. (1998). Gesture and the processing of speech: Neuropsychological evidence. *Brain and Language, 62*, 107-126.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163-180.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228-250.

Hartley, J. & Davies, I. K. (1978). Note-taking: A critical review. *Innovations in Education & Training International, 15*(3), 207-224.

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

Hayati, A. M., & Jalilifar, A. R. (2009). The impact of note-taking strategies on listening comprehension of EFL learners. *Canadian English Language Teaching, 2*(1), 101-111.

Horz, H., & Schnotz, W. (2010). Cognitive load in learning with multiple representations. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 229-252). Cambridge: Cambridge University.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

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

Institute of International Education. (2016). "International Student Enrollment Trends, 1948/49-2015/16." *Open Doors Report on International Educational Exchange*. Retrieved from http://www.iie.org/opendoors

International Listening Association (1995). A ILA definition of listening. *The Listening Post, 53*, 1-5.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* Unpublished PhD dissertation. University of Illinois at Urbana-Champaign.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing, 26*(1), 31-73.

Jang, E. E., Wagner, M., & Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34*, 123-153.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test…or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307-329.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. T. (2002b). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kane, M. T. (2006) *Validation*. In R. Brennan (Ed.), *Educational Measurement*, 4th ed. (pp. 17-64), Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kellerman, S. (1992). "I see what you mean": The role of kinesic behavior in listening and implications for foreign and second language learning. *Applied Linguistics, 13*(3), 239-258.

Kim, K., Turner, S. A., & Perez-Quinones, M. A. (2009). Requirements for electronic note-taking systems: A field study of note-taking in university classrooms. *Education and Information Technologies, 14*(3), 255-283.

King, P. E., & Behnke, R. R. (1989). The effect of time-compressed speech on comprehensive, interpretive, and short-term listening. *Human Communication Research, 15*(3), 428-443.

Kintsch, W. (1998). *Comprehension.* Cambridge: Cambridge University.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111-125). New York: Guilford Press.

Ladas, H. S. (1980). Note-taking on lectures: An information-processing approach. *Educational Psychologist, 15*, 44-53.

Lado, R. (1961). *Language testing: The construction and use of language tests.* London: Longman.

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123.

Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity, 43*(2), 265-275.

Levelt, W. J. M. (1993). Language use in normal speakers and its disorders. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies* (pp. 1-15). Berlin, Germany: De Gruyter.

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635-694.

Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics, 17*(1), 41-50.

Lund, R. J. (1990). A taxonomy for teaching second language listening. *Foreign Language Annals, 23*(2), 105-115.

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

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning.* Cambridge: Cambridge University.

McCuistion, P. J. (1991). Static vs. dynamic visuals in computer-assisted instruction. *Engineering Design Graphics Journal, 55*(2), 25-33.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 9-28). Cambridge: Cambridge University.

Morrel Samuels, P., Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 615-662.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25*(6), 1159-1168.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nagel, T. (1986). *The view from nowhere.* Oxford: Oxford University.

Nagle, S. J., & Sanders, S. L. (1986). Comprehension theory and second language pedagogy. *TESOL Quarterly, 20*(1), 9-26.

NVivo qualitative data analysis software (Version 10) [Computer software]. (2014). Doncaster, Australia: QSR International Pty Ltd.

Ochs, E., & Schieffelin, B. (2009). Language acquisition and socialization: Three developmental stories and their implications. In A. Duranti (Ed.), *Linguistic anthropology: A reader* (2nd ed.) (pp. 296-328). Malden, MA: Wiley-Blackwell.

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.

Olson, K. (2003). LSAT listening assessment: Theoretical background and specifications. *Law School Admission Council (LSAC) Research Report 03-02.* Retrieved from http://www.lsac.org/lsacresources/Research/rr/pdf/RR-03-02.pdf

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13*, 48-63.

Palkovitz, R. J., & Lore, R. K. (1980). Note taking and note review: Why students fail questions based on lecture material. *Teaching of Psychology, 7(3)*, 159-161.

Pettersson, R. (2002). *Information design: An introduction.* Amsterdam: John Benjamins Publishing Company.

Peverly, S. T., Garner, J. K., & Vekaria, P. C. (2013). Both handwriting speed and selective attention are important to lecture note-taking. *Reading and Writing: An Interdisciplinary Journal, 27*, 1-30.

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology, 19*, 291-312.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538-553.

Progrosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal, 14*, 34-44.

Purdy, M. W. (2010). Qualitative research: Critical for understanding listening. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 33-45). Oxford: Wiley-Blackwell.

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

Rubin, J. (1995). The contribution of video to the development of competence in listening. In D. Mendelsohn, & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 151-165). San Diego, CA: Dominie Press.

Schnotz, W. (2005) An integrated model of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49-69). Cambridge: Cambridge University.

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.

Smoker, T. J., Murphy, C. E., & Rockwell, A. K. (2009). Comparing memory for handwriting versus typing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting – 2009, 53*, 1744-1747.

Song, M.-Y. (2012). Note-taking quality and performance on an L2 academic listening test. *Language Testing, 29*(1), 67-89.

Stacy, E. M., & Cain, J. (2015). Note-taking and handouts in the digital age. *American Journal of Pharmaceutical Education, 79*(7), 1-6.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*, 661-699.

Suvorov, R. (2008). *Context visuals in L2 listening tests: The effectiveness of photographs and video vs. audio-only format* (Unpublished master's thesis). Iowa State University, Ames, IA.

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. Unpublished doctoral dissertation, Iowa State University, Ames, IA.

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

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*(3), 185-233.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13*(1), 12-28.

Teng, H. (2011). Exploring note-taking strategies of EFL listeners. *Procedia – Social and Behavioral Sciences, 15*, 480-484.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge: Cambridge University.

Toulmin, S. E. (2003). *The uses of argument: Updated edition.* Cambridge: Cambridge University.

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

Vandergrift, L., & Goh, C. C. M. (2011). *Teaching and learning second language listening: Metacognition in action.* New York: Routledge.

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching, 43*(1), 1-37.

von Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second-language research and teaching. *Canadian Modern Language Review, 36*(2), 225-237.

Voss, J. L., Gonsalves, B. D., Federmeier, K. D., Tranel, D., & Cohen, N. J. (2011). Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. *Nature Neuroscience, 14*(1), 115-120.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University, 2*(1). Retrieved from http://journals.tc-library.org/index.php/tesol/issue/view/2

Wagner, E. (2006). *Utilizing the visual channel: An investigation of the use of video texts on tests of second language listening ability.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(1), 67-86.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly, 5*(3), 218-243.

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

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

Waszak, C., & Sines, M. (2003). Mixed methods in psychological research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 557-576). Thousand Oaks, CA: Sage Publishers.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research*. Chicago, IL: Rand McNally.

Wolvin, A., & Coakley, C. G. (1996). *Listening* (5th ed.). Dubuque, IA: Brown & Benchmark.

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

# Appendix A
# Sample Listening Transcript and Test Questions

**Dadaism and Surrealism**

Today we're going to talk about the Dada art movement. If you think Dada sounds like a word that has been made up, then you're correct. The term dada is a nonsense word; it has no meaning at all. We don't know what the origins of the term are, but we can trace the art movement back to its origins. Prior to World War I in Europe, there was a sense of rationality, of order. At this time, everything was viewed as stable and following rules, and much of this can be seen in the painting of the time. However, once World War I came and went, many Europeans found their world had been changed. Millions or men died in the war and cities were destroyed. Europeans had never seen destruction like this before and they now questioned the culture of rationality and order that they had previously been so willing to accept prior to the outbreak of war. Because of this, the time following the war became known as the age of disillusionment, or a time where people were disappointed that the world was not how they actually thought it was.

So how does all of this relate to the Dada movement in art? Well, because of the war and the way it flipped culture on its side, many artists began to push art's boundaries. They started to create paintings that were meant to display the sense of "absurdity" that had appeared. They artwork was absurd in that it was meant to shock the public by presenting the ridiculous and absurd concepts the dada artists sought to paint. So, you can see that dada artists rejected reason or rational thought. Because of the war, they no longer believed that rational thought would help to solve social problems and that a new type of thinking was required.

Let's take two artists as examples of these ideas: Marcel Duchamp and Salvador Dali. Duchamp was very much the pioneer of the Dada movement. He was originally a painter and, after the war, stopped painting and turned to making sculptures that he referred to as "ready-made." They most likely got this name because they were made from readily available materials. For instance, one such sculpture was made of a stool with a bicycle wheel attached to the top of it. At this time, many people reacted to such works of Dadaism negatively. They found the art of the movement to be distasteful and outrageous. Many of them didn't even think that they were works of art at all, that's how far they were from the previously conceived culture of rationality. You would think this would discourage Duchamp and other Dadaists, but actually this is what they wanted. With the rise of dada art, they were actually seeking to produce "non-art" or "anti-art." They actually sought to disregard all rules previously made under the culture of rationality.

In addition to these sculptures, Duchamp would also try to take established cultural standards and try to challenge them through other works. One way that he did this was to take the works of the "great" painters of the past and…edit them. One such work that some of you may be familiar with is one called L.H.O.O.Q. In this painting he actually took a reproduction, or print of Leonardo da Vinci's Mona Lisa and drew a mustache and beard on her face. In doing so, he was disrespecting a treasured painting of the past that has even to this day been held in high regard, receiving praise from many people. Such disrespect was an important characteristic which characterized much of Duchamp's work and the early Dada movement

But what about other artists? How did they represent this movement? As the Dada movement started to gain momentum and popularity, you could see influences from outside of the art world show themselves. One strong influence was Sigmund Freud, a psychiatrist from Austria who had published a series of essays on the subconscious and the interpretation of dreams prior to the start of World War I. At first thought, you might think "well what does psychiatry have to do with an art movement?" but you have to remember that the Dada movement rested in the absurd and, as it advanced, its members began to focus on the more surreal and subconscious characteristics of human nature as a means of freeing human imagination and creativity. The free association technique associated with Freud allowed artists to draw whatever came to their mind, allowing for a direct connection with their subconscious. They believed that this would help to free people from the rationalism and logicality that had trapped them and led to the war. In a way, they believed that this was more effective for promoting social change than directly attacking the social conventions of the day as Duchamps had been doing.

This brings us to another great artist of the period: Salvador Dali. Salvador Dali is actually probably the best example of this "new" Dadaism. Starting out in the cubist tradition under Pablo Picasso, Dali soon became interested in Freud's ideas which ultimately led to a change in his artistic style and ended up fully embracing surrealism in his work. Now as I said, the new Dadaism stopped attacking cultural norms and started focusing on freeing people through free association techniques, which is basically when you write or draw whatever comes to mind without stopping it from happening. Dali's paintings show this in that the subject matter is almost dream-like in appearance. For instance, one of his most famous paintings called "The Persistence of Memory" presents the viewer with a picture that is full of melting clocks representing to show that time is not a rigid concept, it is actually dependent on the person experiencing it. Looking closer at the painting, one can also see ants on a pocket watch, which Dali often used to symbolize decay in his work. In this instance, they were used to show the decay of typical conventions or ideas of what time is. The painting further shows the colors blue, yellow, and brown; colors Dali often used to represent his home country, Spain. The way that this painting was made is meant to help the viewer to step inside the dreams of an artist and, by doing so, to open up the viewer's imagination and in this way rethink the cultural and social norms of the day. While Dali did not directly challenge culture and social norms through his paintings as Duchamps did, he did paint with the purpose of encouraging individual imagination by freeing it from the chains of conservative Victorian ideas. As you might recall, I mentioned that Dali started in the Cubist tradition, which was characterized as being rational, ordered, and logical. It is interesting to note here that even though he started out in the Cubist tradition, Dali began to attack the work of Cubists such as Pablo Picasso (who he initially respected and wanted to work with in his early years). Nobody really knows for certain why he suddenly started doing this, but many have speculated some of the reasons, with many looking at the different ideologies of the two movements.

So basically we had this movement in art that sought to find ways to deal with the fallout from World War I by both challenging the cultural norms that preceded the war and that many thought were a direct cause for the conflict as well as by promoting more freedom of thought as a way of moving away from conservative thought towards more open-minded views. While this art movement essentially ended at the start of the second World War, it's influences still continue

today and provided a starting point for the abstract expressionist movement that followed, with many of the best modern artists showing Dada characteristics in their artwork. Few movements in art history can claim to have had such a far-reaching affect on art as we know it, but this is something we will go into more detail in our next lecture.

1. What is implied about the philosophy of the Dada movement?
    a. It was not taken seriously by other artists.
    b. It varied from one country to another.
    c. It challenged people's concept of what art is.*
    d. It was based on a realistic style of art.
2. Which of the following paintings would most likely be representative of the movement that took place prior to the Dada movement?
    a. Impressionist
    b. Dada
    c. Realistic*
    d. Kandinsky
3. Which of the following best explains why Freud's *Interpretation of Dreams* was used as an inspiration for the later Dada movement?
    a. Dreams were associated with the subconscious and irrational side of humans and encouraged the development of creativity.*
    b. Dreams provided a good source for displaying colors that the Dada artists liked to use.
    c. Dreams allowed the Dada artists to hide their underlying challenges to cultural norms found in their art.
    d. Dreams provided new subject matter that previous artists had never tried to making paintings of before.
4. What is the main point of the lecture?
    a. To discuss the causes and beliefs underlying the Dada movement*
    b. To describe the artwork of two famous artists in the Dada movement
    c. To compare the artwork of the Dada movement to other artistic movements.
    d. To explain how the Dada movement's artwork changed over time.
5. What is the most likely reason Salvador Dali began to criticize cubists such as Pablo Picasso?
    a. The two movements followed opposite completely opposite ideologies portrayed in their works.*
    b. Pablo Picasso originally criticized Salvador Dali's artwork.
    c. Dali and Picasso were artists competing against each other for fame.
    d. Dali only wanted to make the differences between Cubist and Dadaist works very clear.
6. Dali often used symbolism in his paintings in order to represent subconscious ideas. Which of these is NOT an example of symbolism he used in his works?
    a. Ants
    b. Melting Clocks
    c. The colors blue, brown, and yellow.
    d. Cliffs and mountains *
7. What about World War I led to the rise of the Dada movement?

a. The death and destruction from the war led them to question what they already knew.*
b. The scenes from the war resulted provided artists with new subject matter to paint.
c. The politics that led to the start of the war frustrated artists who sought for reform.
d. The artists felt that the irrational behavior that led to the war must be captured in art to prevent it from happening again.

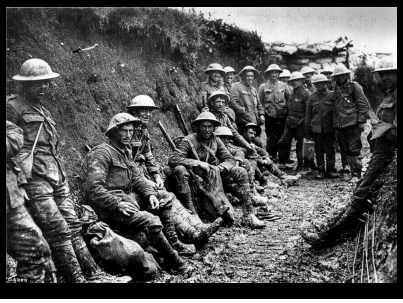
8. What is the primary reason that Dada artists sought to promote a culture of "absurdity"?
    a. Because they felt that rationality had caused World War I.*
    b. Because they wanted to push art's boundaries and challenge people.
    c. They felt that absurdity made for better and more interesting paintings.
    d. They wanted art to focus on something new that people had never seen before.
9. According to the lecture, which of the following is NOT an example of how Duchamp's work challenged cultural norms?
    a. Taking old paintings from other artists and repainting them in absurd ways.
    b. Taking readily available materials and putting them together to make up his pieces
    c. Making pieces of "anti-art" that many people felt repulsed by or hated
    d. Painting dream-like images in order create paintings with hidden meanings. *
10. Based on what is said in the lecture, the speaker could best be said to find the Dada movement to be:
    a. Admirable
    b. Distasteful
    c. Influential*
    d. Extreme

NOTE: Options were randomized when presented to examinees on test.

# Appendix B
# Sample Slides from Listening Passage

## Dadaism

- Dada = nonsense
- Arose after WWI
- Believed culture of rationalism and order had caused the war.



## Dadaism

- Pushed the boundaries of art
- Rejects rational thought
- Embraces absurdity

## Marcel Duchamp

- Pioneer of Dadaism
- Created “Ready-made” sculptures

## Reactions

- Negative
- People felt it was distasteful and outrageous

- But…

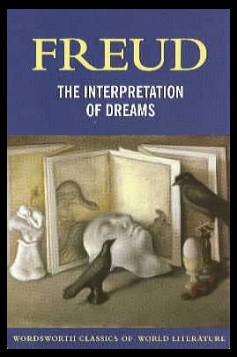
- Dadaists wanted these reactions
- Rejecting cultural norms

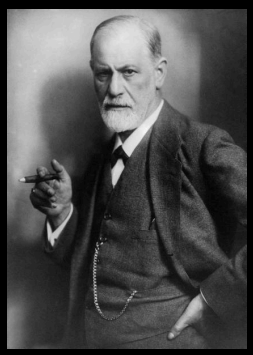
## Marcel Duchamps

- Paintings challenging cultural norms


## Influences on Dadaism

- Sigmund Freud
- Interpretation of Dreams



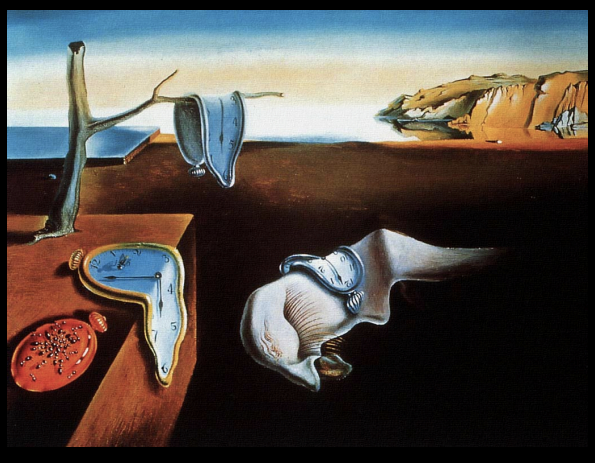



## Changes in Dadaism

- Free association in art
  - Draw whatever comes to mind
  - Free people from rationalism and logic
  - Surreal and subconcious
- Allowed for challenging culture less directly

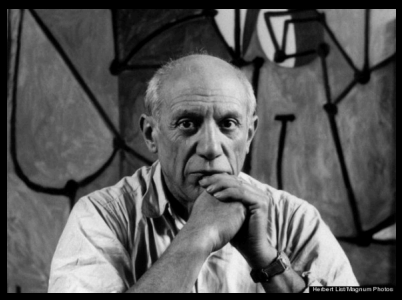
## Salvador Dali

- The Persistence of Memory



# Dali vs. Cubism

- Challenged Cubist painters such as Picasso



# Appendix C
# Post-Test Questionnaire

PLEASE FILL OUT THE FOLLOWING INFORMATION AND ANSWER THE FOLLOWING QUESTIONS AS COMPLETELY AS POSSIBLE

1. Name: a. First name:

   c. Middle initial:

   b. Last name:

2. Age:

3. Gender: ☐Male ☐Female

5. Email address:

6. Native language (first fluent language, also known as your "mother tongue"):

   a. How did you learn English?

   b. How old were you when you started learning English?

7. How would you rate yourself in your typing ability (Excellent, Good, Okay, or Poor)? Have you taken any typing classes?

8. Which of the lecture styles did you prefer, the audio or the audio with video? Why?

9. Do you think that the presence of the video aided in your comprehension of the information being delivered? Why or why not?

10. What did you find yourself paying the most attention to in the video lecture?

11. Did you find it difficult to take notes while the video was playing? If so, why? Please explain the nature of any difficulties you had.

12. Did the presence of video make it easier to remember lecture information, more difficult to remember lecture information, or have no effect on your memory?

13. How do you prefer to take notes in class? By handwriting them or by typing them? Why?
14. If you were given the choice between handwriting and typing notes while listening on a test like the TOEFL, which would you choose and why?

15. Were you familiar with any of the topics covered by the lectures on the exam? If so, which ones? If any were familiar, did this help you answer any of the questions?

# Appendix D
# Item Facility and Point Biserial Correlation Coefficients for all Items by Condition

| | All Conditions | | Audio Input | | Video Input | | Handwritten Notes | | Typed Notes | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Item** | ***IF*** | $r_{pbi}$ | ***IF*** | $r_{pbi}$ | ***IF*** | $r_{pbi}$ | ***IF*** | $r_{pbi}$ | ***IF*** | $r_{pbi}$ |
| **A1** | 0.68 | 0.47 | 0.67 | 0.44 | 0.67 | 0.49 | 0.70 | 0.47 | 0.65 | 0.47 |
| **A2** | 0.44 | 0.41 | 0.36 | 0.59 | 0.50 | 0.22 | 0.45 | 0.36 | 0.42 | 0.46 |
| **A3** | 0.62 | 0.42 | 0.64 | 0.49 | 0.59 | 0.35 | 0.65 | 0.30 | 0.59 | 0.52 |
| **A4** | 0.78 | 0.32 | 0.85 | 0.29 | 0.69 | 0.35 | 0.84 | 0.27 | 0.71 | 0.36 |
| **A5** | 0.40 | 0.47 | 0.40 | 0.56 | 0.40 | 0.40 | 0.39 | 0.52 | 0.41 | 0.45 |
| **A6** | 0.38 | 0.42 | 0.37 | 0.43 | 0.39 | 0.41 | 0.37 | 0.45 | 0.39 | 0.40 |
| **A7** | 0.59 | 0.38 | 0.61 | 0.45 | 0.56 | 0.30 | 0.56 | 0.31 | 0.62 | 0.43 |
| **A8** | 0.41 | 0.50 | 0.44 | 0.59 | 0.37 | 0.42 | 0.40 | 0.43 | 0.41 | 0.56 |
| **A9** | 0.64 | 0.41 | 0.58 | 0.40 | 0.68 | 0.42 | 0.64 | 0.35 | 0.63 | 0.46 |
| **A10** | 0.44 | 0.42 | 0.48 | 0.48 | 0.39 | 0.38 | 0.36 | 0.28 | 0.51 | 0.53 |
| **A11** | 0.59 | 0.25 | 0.60 | 0.17 | 0.57 | 0.33 | 0.56 | 0.10 | 0.62 | 0.37 |
| **A12** | 0.52 | 0.46 | 0.52 | 0.50 | 0.51 | 0.40 | 0.51 | 0.31 | 0.53 | 0.57 |
| **A13** | 0.59 | 0.44 | 0.55 | 0.50 | 0.61 | 0.37 | 0.63 | 0.31 | 0.54 | 0.55 |
| **A14** | 0.42 | 0.41 | 0.36 | 0.31 | 0.46 | 0.52 | 0.43 | 0.39 | 0.40 | 0.44 |
| **A15** | 0.47 | 0.45 | 0.55 | 0.42 | 0.38 | 0.48 | 0.54 | 0.47 | 0.40 | 0.46 |
| **A16** | 0.62 | 0.42 | 0.65 | 0.46 | 0.58 | 0.37 | 0.64 | 0.27 | 0.60 | 0.53 |
| **A17** | 0.44 | 0.51 | 0.46 | 0.53 | 0.41 | 0.48 | 0.42 | 0.40 | 0.46 | 0.59 |
| **A18** | 0.40 | 0.46 | 0.36 | 0.59 | 0.43 | 0.32 | 0.40 | 0.36 | 0.40 | 0.55 |
| **A19** | 0.32 | 0.43 | 0.26 | 0.40 | 0.37 | 0.49 | 0.26 | 0.36 | 0.37 | 0.49 |
| **A20** | 0.31 | 0.41 | 0.32 | 0.38 | 0.29 | 0.45 | 0.28 | 0.32 | 0.33 | 0.48 |
| **A21** | 0.66 | 0.37 | 0.66 | 0.37 | 0.65 | 0.37 | 0.65 | 0.39 | 0.67 | 0.36 |
| **A22** | 0.49 | 0.48 | 0.47 | 0.51 | 0.50 | 0.44 | 0.45 | 0.41 | 0.53 | 0.54 |
| **A23** | 0.63 | 0.39 | 0.63 | 0.41 | 0.62 | 0.38 | 0.63 | 0.24 | 0.63 | 0.51 |
| **A24** | 0.47 | 0.40 | 0.50 | 0.45 | 0.43 | 0.34 | 0.51 | 0.30 | 0.43 | 0.49 |
| **A25** | 0.56 | 0.44 | 0.58 | 0.55 | 0.53 | 0.34 | 0.59 | 0.38 | 0.52 | 0.50 |
| **A26** | 0.47 | 0.37 | 0.44 | 0.36 | 0.49 | 0.38 | 0.45 | 0.29 | 0.49 | 0.43 |
| **A27** | 0.42 | 0.55 | 0.41 | 0.57 | 0.42 | 0.52 | 0.38 | 0.51 | 0.46 | 0.59 |
| **A28** | 0.53 | 0.37 | 0.57 | 0.34 | 0.47 | 0.40 | 0.54 | 0.36 | 0.51 | 0.39 |
| **A29** | 0.43 | 0.49 | 0.41 | 0.48 | 0.44 | 0.48 | 0.40 | 0.48 | 0.46 | 0.50 |
| **A30** | 0.56 | 0.41 | 0.53 | 0.38 | 0.58 | 0.44 | 0.54 | 0.33 | 0.58 | 0.49 |
| **B1** | 0.57 | 0.41 | 0.52 | 0.38 | 0.61 | 0.42 | 0.52 | 0.36 | 0.61 | 0.44 |
| **B2** | 0.53 | 0.46 | 0.38 | 0.53 | 0.67 | 0.31 | 0.48 | 0.47 | 0.57 | 0.44 |
| **B3** | 0.72 | 0.49 | 0.67 | 0.52 | 0.77 | 0.44 | 0.67 | 0.53 | 0.77 | 0.44 |
| **B4** | 0.63 | 0.38 | 0.54 | 0.41 | 0.71 | 0.29 | 0.64 | 0.46 | 0.61 | 0.31 |
| **B5** | 0.51 | 0.55 | 0.40 | 0.53 | 0.62 | 0.53 | 0.46 | 0.64 | 0.56 | 0.45 |
| **B6** | 0.68 | 0.37 | 0.63 | 0.25 | 0.72 | 0.48 | 0.61 | 0.38 | 0.74 | 0.33 |
| **B7** | 0.37 | 0.37 | 0.30 | 0.36 | 0.43 | 0.34 | 0.37 | 0.39 | 0.36 | 0.36 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **B8** | 0.46 | 0.50 | 0.41 | 0.59 | 0.50 | 0.39 | 0.42 | 0.53 | 0.49 | 0.45 |
| **B9** | 0.70 | 0.38 | 0.69 | 0.34 | 0.71 | 0.43 | 0.72 | 0.43 | 0.68 | 0.34 |
| **B10** | 0.33 | 0.21 | 0.32 | 0.24 | 0.34 | 0.17 | 0.30 | 0.07 | 0.36 | 0.33 |
| **B11** | 0.69 | 0.49 | 0.62 | 0.43 | 0.76 | 0.53 | 0.66 | 0.57 | 0.72 | 0.39 |
| **B12** | 0.50 | 0.57 | 0.46 | 0.51 | 0.53 | 0.63 | 0.44 | 0.63 | 0.55 | 0.49 |
| **B13** | 0.25 | 0.31 | 0.26 | 0.39 | 0.24 | 0.24 | 0.23 | 0.28 | 0.27 | 0.33 |
| **B14** | 0.69 | 0.46 | 0.63 | 0.52 | 0.74 | 0.37 | 0.64 | 0.52 | 0.73 | 0.38 |
| **B15** | 0.64 | 0.53 | 0.57 | 0.45 | 0.71 | 0.59 | 0.55 | 0.58 | 0.73 | 0.45 |
| **B16** | 0.35 | 0.41 | 0.31 | 0.44 | 0.38 | 0.36 | 0.29 | 0.41 | 0.40 | 0.38 |
| **B17** | 0.62 | 0.45 | 0.53 | 0.37 | 0.70 | 0.50 | 0.60 | 0.38 | 0.63 | 0.53 |
| **B18** | 0.47 | 0.48 | 0.41 | 0.42 | 0.53 | 0.51 | 0.40 | 0.49 | 0.54 | 0.45 |
| **B19** | 0.49 | 0.52 | 0.44 | 0.58 | 0.53 | 0.44 | 0.46 | 0.47 | 0.51 | 0.56 |
| **B20** | 0.67 | 0.41 | 0.64 | 0.41 | 0.69 | 0.40 | 0.69 | 0.48 | 0.64 | 0.36 |
| **B21** | 0.65 | 0.46 | 0.59 | 0.47 | 0.71 | 0.43 | 0.61 | 0.47 | 0.69 | 0.45 |
| **B22** | 0.46 | 0.36 | 0.40 | 0.38 | 0.52 | 0.30 | 0.44 | 0.31 | 0.48 | 0.41 |
| **B23** | 0.45 | 0.46 | 0.45 | 0.43 | 0.44 | 0.52 | 0.46 | 0.42 | 0.43 | 0.52 |
| **B24** | 0.54 | 0.53 | 0.43 | 0.46 | 0.64 | 0.55 | 0.52 | 0.51 | 0.55 | 0.55 |
| **B25** | 0.46 | 0.45 | 0.40 | 0.40 | 0.51 | 0.49 | 0.43 | 0.45 | 0.48 | 0.45 |
| **B26** | 0.38 | 0.32 | 0.36 | 0.33 | 0.39 | 0.32 | 0.39 | 0.30 | 0.36 | 0.35 |
| **B27** | 0.44 | 0.48 | 0.41 | 0.56 | 0.47 | 0.41 | 0.45 | 0.47 | 0.43 | 0.51 |
| **B28** | 0.23 | 0.25 | 0.24 | 0.23 | 0.22 | 0.29 | 0.24 | 0.19 | 0.22 | 0.32 |
| **B29** | 0.41 | 0.45 | 0.33 | 0.31 | 0.48 | 0.55 | 0.38 | 0.44 | 0.43 | 0.46 |
| **B30** | 0.49 | 0.45 | 0.50 | 0.52 | 0.47 | 0.42 | 0.44 | 0.46 | 0.53 | 0.44 |

# Appendix E
# FACETS Examinee Measurement Report

Examinees Measurement Report (arranged by mN).

| Total Score | Total Count | Obsvd Average | Fair(M) Average | Model Measure | Model S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim. Discrm | Correlation PtMea | Correlation PtExp | Num Examinees |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 60 | 60 | 1.00 | 1.00 | ( 5.30 | 1.83) | Maximum | | | | | .00 | .00 | 124 124 |
| 57 | 60 | .95 | .96 | 2.95 | .60 | 1.03 | .2 | .86 | .0 | .99 | .13 | .15 | 198 198 |
| 56 | 60 | .93 | .94 | 2.62 | .52 | .86 | -.1 | .51 | -.9 | 1.11 | .44 | .15 | 93 93 |
| 55 | 60 | .92 | .93 | 2.39 | .48 | 1.09 | .3 | 1.21 | .5 | .93 | .03 | .18 | 183 183 |
| 55 | 60 | .92 | .93 | 2.38 | .48 | .92 | -.1 | .66 | -.6 | 1.08 | .36 | .17 | 151 151 |
| 55 | 60 | .92 | .93 | 2.37 | .47 | .86 | -.2 | .53 | -1.0 | 1.12 | .45 | .17 | 99 99 |
| 54 | 60 | .90 | .91 | 2.17 | .44 | 1.01 | .1 | .82 | -.3 | 1.02 | .22 | .18 | 120 120 |
| 53 | 60 | .88 | .90 | 2.00 | .41 | 1.00 | .0 | .92 | .0 | 1.01 | .22 | .21 | 200 200 |
| 53 | 60 | .88 | .90 | 1.99 | .41 | 1.05 | .2 | 1.02 | .1 | .97 | .13 | .20 | 170 170 |
| 52 | 60 | .87 | .88 | 1.83 | .39 | 1.08 | .3 | 1.08 | .3 | .94 | .10 | .21 | 160 160 |
| 51 | 60 | .85 | .87 | 1.69 | .37 | 1.00 | .0 | 1.11 | .4 | .97 | .19 | .23 | 23 23 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 51 | 60 | .85 | .87 | 1.69 | .37 | .86 | -.5 | .67 | -1.0 | 1.16 | .46 | .22 | 42 | 42 |
| | 51 | 60 | .85 | .87 | 1.69 | .37 | 1.09 | .4 | 1.13 | .5 | .91 | .07 | .22 | 132 | 132 |
| | 51 | 60 | .85 | .87 | 1.68 | .37 | .88 | -.4 | .69 | -1.0 | 1.14 | .44 | .21 | 95 | 95 |
| | 51 | 60 | .85 | .87 | 1.68 | .37 | .96 | -.1 | .87 | -.3 | 1.05 | .30 | .21 | 98 | 98 |
| | 50 | 60 | .83 | .85 | 1.56 | .36 | 1.12 | .6 | 1.45 | 1.4 | .81 | -.04 | .23 | 127 | 127 |
| | 50 | 60 | .83 | .85 | 1.55 | .36 | .94 | -.2 | .96 | .0 | 1.05 | .30 | .22 | 94 | 94 |
| | 50 | 60 | .83 | .85 | 1.55 | .36 | .99 | .0 | .94 | -.1 | 1.02 | .25 | .22 | 96 | 96 |
| | 49 | 60 | .82 | .84 | 1.44 | .34 | 1.22 | 1.0 | 1.58 | 1.8 | .67 | -.18 | .23 | 134 | 134 |
| | 48 | 60 | .80 | .82 | 1.32 | .33 | .99 | .0 | 1.00 | .0 | 1.01 | .24 | .24 | 92 | 92 |
| | 48 | 60 | .80 | .82 | 1.32 | .33 | .95 | -.1 | .93 | -.2 | 1.06 | .31 | .24 | 64 | 64 |
| | 47 | 60 | .78 | .80 | 1.21 | .32 | .94 | -.2 | .92 | -.2 | 1.08 | .33 | .25 | 97 | 97 |
| | 47 | 60 | .78 | .80 | 1.21 | .32 | .99 | .0 | 1.02 | .1 | 1.01 | .25 | .25 | 113 | 113 |
| | 46 | 60 | .77 | .79 | 1.11 | .32 | .91 | -.5 | .86 | -.5 | 1.15 | .41 | .27 | 17 | 17 |
| | 46 | 60 | .77 | .78 | 1.11 | .32 | .88 | -.6 | .86 | -.6 | 1.18 | .43 | .26 | 65 | 65 |
| | 45 | 60 | .75 | .77 | 1.02 | .31 | 1.00 | .0 | 1.01 | .1 | .99 | .25 | .26 | 43 | 43 |
| | 43 | 60 | .72 | .73 | .83 | .30 | .96 | -.2 | .91 | -.4 | 1.11 | .34 | .27 | 44 | 44 |
| | 43 | 60 | .72 | .73 | .83 | .30 | .76 | -1.9 | .65 | -2.1 | 1.55 | .65 | .27 | 140 | 140 |
| | 43 | 60 | .72 | .73 | .83 | .30 | 1.00 | .0 | 1.01 | .1 | .99 | .26 | .27 | 143 | 143 |
| | 43 | 60 | .72 | .73 | .83 | .30 | .96 | -.3 | .87 | -.7 | 1.14 | .37 | .27 | 147 | 147 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 43 | 60 | .72 | .73 | .83 | .30 | .94 | -.3 | .94 | -.3 | 1.12 | .35 | .27 | 91 91 |
| | 43 | 60 | .72 | .73 | .82 | .30 | 1.01 | .1 | 1.03 | .2 | .96 | .24 | .27 | 75 75 |
| | 42 | 60 | .70 | .72 | .75 | .29 | .92 | -.5 | .84 | -.9 | 1.22 | .41 | .28 | 148 148 |
| | 42 | 60 | .70 | .72 | .74 | .29 | 1.04 | .3 | 1.18 | 1.1 | .84 | .18 | .28 | 199 199 |
| | 41 | 60 | .68 | .70 | .66 | .29 | .99 | .0 | 1.01 | .1 | 1.01 | .28 | .28 | 141 141 |
| | 40 | 60 | .67 | .68 | .57 | .29 | .85 | -1.3 | .80 | -1.4 | 1.45 | .52 | .29 | 16 16 |
| | 40 | 60 | .67 | .68 | .57 | .29 | .92 | -.6 | .93 | -.4 | 1.20 | .38 | .28 | 74 74 |
| | 39 | 60 | .65 | .66 | .50 | .28 | .78 | -2.2 | .72 | -2.2 | 1.74 | .62 | .29 | 139 139 |
| | 39 | 60 | .65 | .66 | .50 | .28 | .87 | -1.2 | .82 | -1.3 | 1.43 | .48 | .29 | 146 146 |
| | 39 | 60 | .65 | .66 | .49 | .28 | .81 | -2.0 | .75 | -2.0 | 1.66 | .59 | .28 | 116 116 |
| | 39 | 60 | .65 | .66 | .49 | .28 | 1.00 | .0 | .99 | .0 | 1.01 | .28 | .28 | 117 117 |
| | 39 | 60 | .65 | .66 | .49 | .28 | .79 | -2.1 | .74 | -2.1 | 1.69 | .61 | .29 | 1 1 |
| | 39 | 60 | .65 | .66 | .49 | .28 | .83 | -1.7 | .78 | -1.8 | 1.59 | .56 | .28 | 175 175 |
| | 38 | 60 | .63 | .65 | .42 | .28 | 1.03 | .3 | 1.05 | .4 | .88 | .24 | .29 | 142 142 |
| | 38 | 60 | .63 | .65 | .41 | .28 | .95 | -.4 | .96 | -.2 | 1.16 | .35 | .28 | 104 104 |
| | 38 | 60 | .63 | .65 | .41 | .28 | .98 | -.1 | 1.01 | .1 | 1.04 | .30 | .28 | 125 125 |
| | 38 | 60 | .63 | .64 | .41 | .28 | 1.09 | .9 | 1.16 | 1.2 | .63 | .14 | .30 | 194 194 |
| | 38 | 60 | .63 | .64 | .41 | .28 | .94 | -.6 | .94 | -.4 | 1.22 | .37 | .29 | 73 73 |
| | 37 | 60 | .62 | .63 | .34 | .28 | 1.05 | .6 | 1.06 | .5 | .79 | .21 | .30 | 27 27 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 37 | 60 | .62 | .63 | .34 | .28 | .90 | -1.1 | .88 | -1.0 | 1.42 | .45 | .30 | 137 137 |
| | 37 | 60 | .62 | .63 | .34 | .28 | 1.09 | .9 | 1.10 | .8 | .64 | .16 | .30 | 144 144 |
| | 37 | 60 | .62 | .63 | .34 | .28 | .92 | -.8 | .91 | -.8 | 1.33 | .41 | .29 | 105 105 |
| | 37 | 60 | .62 | .63 | .34 | .28 | .92 | -.8 | .88 | -1.0 | 1.36 | .42 | .29 | 107 107 |
| | 37 | 60 | .62 | .63 | .34 | .28 | 1.20 | 2.0 | 1.21 | 1.7 | .23 | .00 | .30 | 24 24 |
| | 36 | 60 | .60 | .61 | .27 | .28 | 1.01 | .1 | 1.02 | .2 | .95 | .28 | .30 | 129 129 |
| | 36 | 60 | .60 | .61 | .27 | .28 | .95 | -.4 | .90 | -.8 | 1.25 | .38 | .30 | 133 133 |
| | 36 | 60 | .60 | .61 | .26 | .28 | .99 | -.1 | .97 | -.2 | 1.08 | .32 | .30 | 2 2 |
| | 35 | 60 | .58 | .59 | .19 | .27 | 1.11 | 1.3 | 1.14 | 1.3 | .45 | .12 | .30 | 45 45 |
| | 35 | 60 | .58 | .59 | .19 | .27 | 1.03 | .3 | 1.03 | .3 | .88 | .26 | .30 | 126 126 |
| | 35 | 60 | .58 | .59 | .19 | .27 | 1.03 | .3 | 1.02 | .2 | .88 | .26 | .30 | 128 128 |
| | 35 | 60 | .58 | .59 | .18 | .27 | .98 | -.1 | .97 | -.2 | 1.09 | .32 | .29 | 63 63 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.17 | 1.9 | 1.24 | 2.3 | .08 | .03 | .30 | 135 135 |
| | 34 | 60 | .57 | .57 | .11 | .27 | .79 | -2.7 | .76 | -2.7 | 2.07 | .61 | .30 | 136 136 |
| | 34 | 60 | .57 | .57 | .11 | .27 | .82 | -2.4 | .80 | -2.3 | 1.96 | .57 | .29 | 78 78 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.10 | 1.2 | 1.12 | 1.2 | .45 | .13 | .29 | 111 111 |
| | 34 | 60 | .57 | .57 | .11 | .27 | .93 | -.9 | .91 | -.9 | 1.40 | .40 | .29 | 52 52 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.05 | .6 | 1.08 | .8 | .72 | .20 | .29 | 57 57 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.21 | 2.4 | 1.28 | 2.7 | -.17 | -.05 | .29 | 66 66 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.21 | 2.4 | 1.28 | 2.7 | -.17 | -.05 | .29 | 70 | 70 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.17 | 2.0 | 1.23 | 2.2 | .08 | .03 | .30 | 19 | 19 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.32 | 3.5 | 1.39 | 3.6 | -.65 | -.19 | .30 | 22 | 22 |
| | 34 | 60 | .57 | .57 | .11 | .27 | 1.16 | 1.9 | 1.19 | 1.8 | .17 | .05 | .30 | 187 | 187 |
| | 33 | 60 | .55 | .56 | .04 | .27 | 1.02 | .2 | 1.04 | .4 | .89 | .27 | .30 | 33 | 33 |
| | 33 | 60 | .55 | .55 | .03 | .27 | .88 | -1.5 | .86 | -1.5 | 1.65 | .48 | .30 | 3 | 3 |
| | 32 | 60 | .53 | .54 | -.03 | .27 | .89 | -1.4 | .87 | -1.4 | 1.61 | .46 | .30 | 26 | 26 |
| | 32 | 60 | .53 | .54 | -.03 | .27 | 1.10 | 1.2 | 1.12 | 1.3 | .45 | .15 | .30 | 47 | 47 |
| | 32 | 60 | .53 | .54 | -.04 | .27 | .86 | -1.9 | .85 | -1.7 | 1.82 | .50 | .29 | 102 | 102 |
| | 32 | 60 | .53 | .54 | -.04 | .27 | .90 | -1.3 | .88 | -1.4 | 1.62 | .45 | .29 | 56 | 56 |
| | 32 | 60 | .53 | .54 | -.04 | .27 | 1.04 | .5 | 1.04 | .5 | .75 | .22 | .29 | 157 | 157 |
| | 32 | 60 | .53 | .54 | -.04 | .27 | 1.00 | .0 | 1.01 | .1 | .97 | .28 | .29 | 158 | 158 |
| | 32 | 60 | .53 | .54 | -.04 | .27 | 1.08 | 1.0 | 1.07 | .7 | .58 | .19 | .30 | 184 | 184 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | 1.08 | 1.0 | 1.10 | 1.1 | .52 | .17 | .31 | 48 | 48 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | .81 | -2.7 | .80 | -2.5 | 2.16 | .58 | .29 | 81 | 81 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | .88 | -1.6 | .86 | -1.7 | 1.76 | .48 | .29 | 54 | 54 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | .86 | -2.0 | .84 | -1.9 | 1.88 | .51 | .29 | 59 | 59 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | .96 | -.4 | .97 | -.3 | 1.22 | .34 | .29 | 159 | 159 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | 1.01 | .1 | .99 | .0 | .97 | .28 | .29 | 166 | 166 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 31 | 60 | .52 | .52 | -.11 | .27 | .99 | .0 | 1.00 | .0 | 1.04 | .31 | .30 | 178 | 178 |
| | 31 | 60 | .52 | .52 | -.11 | .27 | 1.09 | 1.1 | 1.16 | 1.7 | .39 | .14 | .30 | 196 | 196 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | 1.03 | .3 | 1.01 | .1 | .88 | .27 | .31 | 46 | 46 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | .85 | -1.9 | .84 | -1.9 | 1.84 | .52 | .31 | 49 | 49 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | 1.22 | 2.7 | 1.25 | 2.7 | -.29 | -.03 | .31 | 50 | 50 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | .88 | -1.7 | .87 | -1.5 | 1.75 | .48 | .30 | 101 | 101 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | .90 | -1.4 | .87 | -1.5 | 1.66 | .46 | .30 | 106 | 106 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | 1.09 | 1.2 | 1.10 | 1.2 | .44 | .15 | .30 | 119 | 119 |
| | 30 | 60 | .50 | .50 | -.18 | .27 | .94 | -.7 | .92 | -.9 | 1.38 | .38 | .29 | 167 | 167 |
| | 30 | 60 | .50 | .50 | -.19 | .27 | 1.16 | 2.1 | 1.21 | 2.2 | -.03 | .04 | .30 | 18 | 18 |
| | 30 | 60 | .50 | .50 | -.19 | .27 | 1.06 | .8 | 1.12 | 1.3 | .56 | .19 | .30 | 188 | 188 |
| | 30 | 60 | .50 | .50 | -.19 | .27 | 1.11 | 1.5 | 1.17 | 1.8 | .26 | .11 | .30 | 197 | 197 |
| | 29 | 60 | .48 | .48 | -.25 | .27 | 1.03 | .4 | 1.02 | .2 | .84 | .26 | .31 | 34 | 34 |
| | 29 | 60 | .48 | .48 | -.25 | .27 | .89 | -1.4 | .88 | -1.3 | 1.61 | .46 | .31 | 38 | 38 |
| | 29 | 60 | .48 | .48 | -.25 | .27 | 1.12 | 1.5 | 1.13 | 1.4 | .31 | .12 | .31 | 150 | 150 |
| | 29 | 60 | .48 | .48 | -.25 | .27 | .97 | -.3 | .98 | -.1 | 1.14 | .33 | .30 | 85 | 85 |
| | 29 | 60 | .48 | .48 | -.26 | .27 | .94 | -.7 | .96 | -.5 | 1.33 | .37 | .29 | 163 | 163 |
| | 29 | 60 | .48 | .48 | -.26 | .27 | .98 | -.2 | .97 | -.3 | 1.14 | .32 | .29 | 164 | 164 |
| | 29 | 60 | .48 | .48 | -.26 | .27 | 1.11 | 1.4 | 1.12 | 1.3 | .33 | .13 | .30 | 12 | 12 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 29 | 60 | .48 | .48 | -.26 | .27 | 1.03 | .4 | 1.03 | .4 | .81 | .25 | .30 | 180 | 180 |
| | 29 | 60 | .48 | .48 | -.26 | .27 | 1.04 | .6 | 1.04 | .5 | .74 | .23 | .30 | 181 | 181 |
| | 29 | 60 | .48 | .48 | -.26 | .27 | 1.15 | 1.9 | 1.20 | 2.1 | .04 | .06 | .30 | 185 | 185 |
| | 29 | 60 | .48 | .48 | -.26 | .27 | 1.07 | .9 | 1.13 | 1.4 | .49 | .17 | .30 | 186 | 186 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | 1.03 | .4 | 1.02 | .2 | .84 | .26 | .31 | 39 | 39 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | .88 | -1.5 | .86 | -1.6 | 1.71 | .48 | .29 | 109 | 109 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | .98 | -.2 | .96 | -.4 | 1.15 | .33 | .29 | 110 | 110 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | 1.05 | .6 | 1.06 | .7 | .71 | .21 | .29 | 114 | 114 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | .96 | -.5 | .94 | -.6 | 1.27 | .36 | .29 | 62 | 62 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | .83 | -2.4 | .80 | -2.3 | 2.04 | .56 | .29 | 152 | 152 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | .99 | .0 | .98 | -.2 | 1.07 | .31 | .29 | 161 | 161 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | 1.01 | .1 | 1.01 | .1 | .95 | .28 | .29 | 169 | 169 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | 1.09 | 1.1 | 1.12 | 1.3 | .44 | .14 | .29 | 171 | 171 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | .95 | -.6 | .97 | -.3 | 1.25 | .35 | .29 | 172 | 172 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | 1.01 | .0 | .99 | .0 | .99 | .29 | .30 | 5 | 5 |
| | 28 | 60 | .47 | .46 | -.33 | .27 | 1.04 | .5 | 1.07 | .7 | .73 | .23 | .30 | 7 | 7 |
| | 27 | 60 | .45 | .45 | -.40 | .27 | .93 | -.8 | .92 | -.9 | 1.36 | .40 | .31 | 35 | 35 |
| | 27 | 60 | .45 | .45 | -.40 | .27 | .98 | -.2 | .96 | -.4 | 1.13 | .34 | .31 | 41 | 41 |
| | 27 | 60 | .45 | .45 | -.40 | .27 | .95 | -.6 | .94 | -.6 | 1.30 | .37 | .29 | 100 | 100 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 27 | 60 | .45 | .45 | -.40 | .27 | .90 | -1.3 | .88 | -1.3 | 1.58 | .44 | .29 | 153 153 |
| | 27 | 60 | .45 | .45 | -.40 | .27 | .98 | -.2 | 1.01 | .1 | 1.06 | .30 | .29 | 156 156 |
| | 27 | 60 | .45 | .45 | -.40 | .27 | .94 | -.8 | .93 | -.8 | 1.37 | .39 | .29 | 168 168 |
| | 27 | 60 | .45 | .44 | -.41 | .27 | 1.01 | .1 | 1.02 | .1 | .92 | .27 | .30 | 25 25 |
| | 27 | 60 | .45 | .44 | -.41 | .27 | 1.19 | 2.3 | 1.25 | 2.5 | -.11 | -.01 | .30 | 192 192 |
| | 27 | 60 | .45 | .44 | -.41 | .27 | 1.07 | .9 | 1.12 | 1.2 | .56 | .18 | .30 | 193 193 |
| | 26 | 60 | .43 | .43 | -.47 | .27 | .93 | -.8 | .91 | -.8 | 1.36 | .41 | .30 | 36 36 |
| | 26 | 60 | .43 | .43 | -.47 | .27 | .95 | -.5 | .94 | -.5 | 1.25 | .38 | .30 | 40 40 |
| | 26 | 60 | .43 | .43 | -.47 | .27 | .94 | -.7 | .91 | -.9 | 1.34 | .39 | .29 | 103 103 |
| | 26 | 60 | .43 | .43 | -.47 | .27 | .79 | -2.8 | .76 | -2.7 | 2.11 | .61 | .29 | 108 108 |
| | 26 | 60 | .43 | .43 | -.47 | .27 | 1.14 | 1.7 | 1.18 | 1.8 | .22 | .06 | .29 | 118 118 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | .93 | -.9 | .92 | -.7 | 1.38 | .40 | .29 | 53 53 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | 1.03 | .4 | 1.03 | .3 | .83 | .24 | .29 | 60 60 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | .98 | -.2 | .97 | -.2 | 1.13 | .32 | .29 | 162 162 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | 1.06 | .7 | 1.06 | .6 | .71 | .20 | .29 | 173 173 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | 1.00 | .0 | 1.03 | .2 | .97 | .28 | .30 | 10 10 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | 1.22 | 2.6 | 1.33 | 3.0 | -.26 | -.07 | .30 | 179 179 |
| | 26 | 60 | .43 | .43 | -.48 | .27 | 1.03 | .3 | 1.02 | .2 | .86 | .25 | .30 | 191 191 |
| | 25 | 60 | .42 | .41 | -.55 | .27 | .89 | -1.3 | .87 | -1.3 | 1.56 | .46 | .29 | 72 72 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 25 | 60 | .42 | .41 | -.55 | .27 | .94 | -.7 | .94 | -.5 | 1.28 | .38 | .30 | 8 | 8 |
| | 25 | 60 | .42 | .41 | -.55 | .27 | .92 | -.9 | .91 | -.8 | 1.38 | .41 | .30 | 176 | 176 |
| | 25 | 60 | .42 | .41 | -.55 | .27 | 1.21 | 2.3 | 1.24 | 2.1 | -.02 | -.02 | .30 | 195 | 195 |
| | 24 | 60 | .40 | .39 | -.63 | .28 | 1.11 | 1.2 | 1.17 | 1.5 | .49 | .12 | .30 | 145 | 145 |
| | 24 | 60 | .40 | .39 | -.63 | .27 | .99 | .0 | 1.01 | .1 | 1.02 | .29 | .29 | 155 | 155 |
| | 24 | 60 | .40 | .39 | -.63 | .27 | 1.08 | .9 | 1.12 | 1.1 | .59 | .14 | .29 | 165 | 165 |
| | 24 | 60 | .40 | .39 | -.63 | .27 | 1.01 | .1 | 1.09 | .8 | .89 | .25 | .29 | 174 | 174 |
| | 24 | 60 | .40 | .39 | -.63 | .28 | .89 | -1.3 | .86 | -1.2 | 1.52 | .46 | .29 | 4 | 4 |
| | 24 | 60 | .40 | .39 | -.63 | .28 | .99 | .0 | 1.00 | .0 | 1.02 | .30 | .29 | 6 | 6 |
| | 23 | 60 | .38 | .37 | -.70 | .28 | 1.24 | 2.4 | 1.33 | 2.6 | -.04 | -.11 | .29 | 61 | 61 |
| | 23 | 60 | .38 | .37 | -.70 | .28 | .99 | .0 | .95 | -.4 | 1.08 | .33 | .30 | 28 | 28 |
| | 22 | 60 | .37 | .36 | -.78 | .28 | .89 | -1.1 | .85 | -1.2 | 1.42 | .46 | .29 | 76 | 76 |
| | 22 | 60 | .37 | .36 | -.78 | .28 | .96 | -.3 | 1.00 | .0 | 1.10 | .32 | .29 | 77 | 77 |
| | 22 | 60 | .37 | .36 | -.78 | .28 | .96 | -.3 | 1.02 | .1 | 1.08 | .32 | .29 | 82 | 82 |
| | 22 | 60 | .37 | .36 | -.78 | .28 | .92 | -.8 | .92 | -.6 | 1.28 | .40 | .29 | 84 | 84 |
| | 22 | 60 | .37 | .36 | -.78 | .28 | 1.02 | .2 | 1.06 | .5 | .91 | .24 | .29 | 88 | 88 |
| | 22 | 60 | .37 | .36 | -.78 | .28 | .83 | -1.8 | .78 | -1.9 | 1.65 | .55 | .28 | 58 | 58 |
| | 22 | 60 | .37 | .35 | -.78 | .28 | .96 | -.4 | .96 | -.2 | 1.14 | .34 | .29 | 14 | 14 |
| | 21 | 60 | .35 | .34 | -.86 | .28 | .94 | -.5 | .90 | -.7 | 1.21 | .38 | .28 | 79 | 79 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 21 | 60 | .35 | .34 | -.86 | .28 | .99 | .0 | .96 | -.2 | 1.06 | .30 | .28 | 55 55 |
| | 21 | 60 | .35 | .34 | -.86 | .28 | 1.14 | 1.3 | 1.21 | 1.5 | .52 | .07 | .30 | 30 30 |
| | 21 | 60 | .35 | .34 | -.86 | .28 | 1.09 | .8 | 1.24 | 1.7 | .61 | .10 | .28 | 9 9 |
| | 20 | 60 | .33 | .32 | -.94 | .29 | 1.00 | .0 | .97 | -.1 | 1.02 | .29 | .28 | 86 86 |
| | 20 | 60 | .33 | .32 | -.94 | .29 | 1.06 | .5 | 1.04 | .3 | .85 | .21 | .29 | 31 31 |
| | 20 | 60 | .33 | .32 | -.94 | .29 | 1.18 | 1.5 | 1.26 | 1.7 | .46 | .01 | .29 | 37 37 |
| | 19 | 60 | .32 | .30 | -1.02 | .29 | .93 | -.5 | .92 | -.4 | 1.17 | .38 | .29 | 29 29 |
| | 19 | 60 | .32 | .30 | -1.02 | .29 | .94 | -.4 | .97 | -.1 | 1.13 | .36 | .29 | 138 138 |
| | 18 | 60 | .30 | .28 | -1.10 | .29 | 1.04 | .3 | 1.04 | .2 | .91 | .20 | .27 | 67 67 |
| | 18 | 60 | .30 | .28 | -1.11 | .29 | .95 | -.3 | .97 | -.1 | 1.11 | .35 | .29 | 149 149 |
| | 18 | 60 | .30 | .28 | -1.11 | .29 | 1.18 | 1.3 | 1.23 | 1.3 | .56 | .00 | .27 | 15 15 |
| | 17 | 60 | .28 | .27 | -1.19 | .30 | 1.03 | .2 | 1.10 | .6 | .90 | .19 | .27 | 80 80 |
| | 17 | 60 | .28 | .27 | -1.19 | .30 | .91 | -.6 | .92 | -.4 | 1.19 | .40 | .27 | 112 112 |
| | 17 | 60 | .28 | .27 | -1.19 | .30 | 1.17 | 1.2 | 1.27 | 1.5 | .60 | -.02 | .27 | 123 123 |
| | 17 | 60 | .28 | .27 | -1.19 | .30 | .89 | -.8 | .86 | -.7 | 1.24 | .43 | .26 | 51 51 |
| | 17 | 60 | .28 | .27 | -1.19 | .30 | 1.13 | .9 | 1.25 | 1.4 | .68 | .03 | .26 | 69 69 |
| | 17 | 60 | .28 | .27 | -1.19 | .30 | .91 | -.6 | .96 | -.2 | 1.17 | .38 | .26 | 71 71 |
| | 16 | 60 | .27 | .25 | -1.28 | .30 | .93 | -.4 | .97 | -.1 | 1.10 | .34 | .26 | 89 89 |
| | 16 | 60 | .27 | .25 | -1.29 | .30 | .98 | -.1 | .92 | -.3 | 1.07 | .33 | .28 | 32 32 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 60 | .27 | .25 | -1.29 | .30 | .95 | -.2 | .92 | -.3 | 1.10 | .33 | .26 | 11 11 |
| 16 | 60 | .27 | .25 | -1.29 | .30 | .93 | -.4 | .92 | -.3 | 1.13 | .36 | .26 | 13 13 |
| 15 | 60 | .25 | .23 | -1.37 | .31 | .94 | -.3 | .85 | -.7 | 1.14 | .38 | .26 | 83 83 |
| 15 | 60 | .25 | .23 | -1.38 | .31 | .98 | .0 | .99 | .0 | 1.03 | .29 | .27 | 131 131 |
| 14 | 60 | .23 | .21 | -1.48 | .32 | .92 | -.4 | .82 | -.8 | 1.16 | .41 | .27 | 130 130 |
| 13 | 60 | .22 | .20 | -1.57 | .32 | .99 | .0 | .91 | -.3 | 1.03 | .28 | .25 | 90 90 |
| 11 | 60 | .18 | .17 | -1.80 | .34 | 1.16 | .7 | 1.58 | 1.9 | .73 | -.13 | .23 | 115 115 |
| 11 | 60 | .18 | .17 | -1.80 | .34 | 1.20 | .9 | 1.55 | 1.8 | .70 | -.17 | .23 | 122 122 |
| 10 | 60 | .17 | .15 | -1.92 | .36 | 1.10 | .5 | 1.22 | .8 | .87 | .03 | .22 | 68 68 |
| 10 | 60 | .17 | .15 | -1.92 | .36 | .97 | .0 | .92 | -.1 | 1.04 | .27 | .22 | 121 121 |
| 10 | 60 | .17 | .15 | -1.92 | .36 | 1.05 | .3 | 1.01 | .1 | .95 | .16 | .22 | 21 21 |
| 7 | 60 | .12 | .10 | -2.36 | .41 | 1.07 | .3 | 1.32 | .8 | .91 | .02 | .19 | 20 20 |
| 6 | 60 | .10 | .09 | -2.53 | .44 | .90 | -.2 | .61 | -.9 | 1.11 | .40 | .18 | 154 154 |
| 6 | 60 | .10 | .09 | -2.54 | .44 | 1.06 | .2 | 1.41 | .9 | .92 | .02 | .18 | 190 190 |
| 5 | 60 | .08 | .07 | -2.74 | .47 | 1.06 | .2 | 1.10 | .3 | .95 | .05 | .17 | 87 87 |
| 5 | 60 | .08 | .07 | -2.75 | .47 | 1.10 | .3 | 1.62 | 1.2 | .88 | -.11 | .17 | 182 182 |
| 4 | 60 | .07 | .06 | -2.99 | .52 | 1.04 | .2 | 1.05 | .2 | .97 | .08 | .15 | 177 177 |
| 4 | 60 | .07 | .06 | -2.99 | .52 | 1.01 | .1 | .85 | .0 | 1.00 | .16 | .15 | 189 189 |
| 30.5 | 60.0 | .51 | .51 | -.11 | .31 | 1.00 | .0 | 1.00 | .0 | | .27 | | Mean (Count: 200) |

```
|    11.5      .0     .19    .20 |   1.09   .12 |  .10  1.1   .18  1.2 |      |   .17       | S.D.
(Population)   |
|    11.6      .0     .19    .20 |   1.09   .12 |  .10  1.1   .18  1.2 |      |   .17       | S.D.
(Sample)       |
+-------------------------------------------------------------------------------------------------
Model, Populn: RMSE .33  Adj (True) S.D. 1.04  Separation 3.12  Strata 4.50  Reliability .91
Model, Sample: RMSE .33  Adj (True) S.D. 1.04  Separation 3.13  Strata 4.51  Reliability .91
Model, Fixed (all same) chi-square:  1514.6  d.f.: 199  significance (probability): .00
Model,  Random (normal) chi-square:  161.1  d.f.: 198  significance (probability): .97
-------------------------------------------------------------------------------------------------
```

# Appendix F
# FACETS Item Measurement Report

Items Measurement Report (arranged by mN).

| Total Score | Total Count | Obsvd Average | Fair(M) Average | Model Measure | Model S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim. Discrm | Correlation PtMea | Correlation PtExp | Nu | Items |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 46 | 200 | .23 | .19 | 1.51 | .19 | 1.16 | 1.5 | 1.25 | 1.3 | .78 | .26 | .40 | 58 | 58 |
| 50 | 200 | .25 | .21 | 1.38 | .18 | 1.12 | 1.2 | 1.10 | .6 | .84 | .31 | .40 | 43 | 43 |
| 61 | 200 | .31 | .27 | 1.03 | .17 | 1.03 | .3 | 1.11 | .8 | .92 | .36 | .40 | 20 | 20 |
| 63 | 200 | .31 | .28 | .97 | .17 | 1.01 | .1 | 1.13 | 1.0 | .95 | .38 | .40 | 19 | 19 |
| 66 | 200 | .33 | .30 | .90 | .17 | 1.25 | 3.0 | 1.46 | 3.3 | .44 | .19 | .41 | 40 | 40 |
| 69 | 200 | .34 | .31 | .82 | .16 | 1.06 | .8 | 1.02 | .1 | .89 | .37 | .41 | 46 | 46 |
| 73 | 200 | .37 | .34 | .71 | .16 | 1.08 | 1.1 | 1.06 | .6 | .83 | .35 | .41 | 37 | 37 |
| 75 | 200 | .38 | .35 | .66 | .16 | 1.09 | 1.4 | 1.15 | 1.3 | .75 | .33 | .41 | 56 | 56 |
| 76 | 200 | .38 | .36 | .63 | .16 | 1.02 | .3 | 1.03 | .2 | .94 | .38 | .40 | 6 | 6 |
| 80 | 200 | .40 | .38 | .52 | .16 | .96 | -.6 | 1.04 | .4 | 1.09 | .43 | .40 | 5 | 5 |
| 80 | 200 | .40 | .38 | .52 | .16 | .98 | -.4 | .95 | -.5 | 1.08 | .42 | .40 | 18 | 18 |
| 81 | 200 | .41 | .39 | .51 | .16 | 1.00 | .0 | .98 | -.2 | 1.01 | .41 | .41 | 59 | 59 |
| 81 | 200 | .41 | .39 | .50 | .16 | .94 | -1.0 | .89 | -1.1 | 1.20 | .46 | .40 | 8 | 8 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 83 | 200 | .41 | .40 | .45 | .16 | 1.02 | .4 | 1.01 | .1 | .93 | .38 | .40 | 14 | 14 |
| | 84 | 200 | .42 | .41 | .43 | .16 | .90 | -1.7 | .90 | -1.1 | 1.30 | .48 | .40 | 27 | 27 |
| | 86 | 200 | .43 | .42 | .38 | .16 | .97 | -.5 | .95 | -.5 | 1.10 | .42 | .40 | 29 | 29 |
| | 87 | 200 | .44 | .42 | .35 | .16 | .97 | -.5 | 1.00 | .0 | 1.07 | .42 | .40 | 2 | 2 |
| | 87 | 200 | .44 | .42 | .35 | .16 | .96 | -.7 | 1.00 | .0 | 1.11 | .43 | .40 | 10 | 10 |
| | 88 | 200 | .44 | .43 | .33 | .16 | .97 | -.5 | 1.00 | .0 | 1.08 | .43 | .41 | 57 | 57 |
| | 88 | 200 | .44 | .43 | .33 | .16 | .94 | -1.1 | .90 | -1.1 | 1.22 | .45 | .40 | 17 | 17 |
| | 89 | 200 | .44 | .43 | .31 | .16 | 1.00 | .0 | 1.09 | 1.0 | .96 | .40 | .41 | 53 | 53 |
| | 91 | 200 | .46 | .45 | .26 | .16 | .95 | -.8 | .92 | -.8 | 1.17 | .45 | .41 | 38 | 38 |
| | 91 | 200 | .46 | .45 | .26 | .16 | 1.03 | .6 | 1.12 | 1.2 | .82 | .36 | .41 | 55 | 55 |
| | 92 | 200 | .46 | .45 | .24 | .16 | 1.08 | 1.5 | 1.19 | 1.9 | .68 | .33 | .41 | 52 | 52 |
| | 94 | 200 | .47 | .46 | .19 | .16 | .97 | -.5 | .93 | -.8 | 1.12 | .44 | .41 | 48 | 48 |
| | 94 | 200 | .47 | .47 | .18 | .16 | .98 | -.4 | .94 | -.7 | 1.10 | .42 | .40 | 15 | 15 |
| | 94 | 200 | .47 | .47 | .18 | .16 | .98 | -.3 | 1.04 | .5 | 1.03 | .40 | .40 | 24 | 24 |
| | 94 | 200 | .47 | .47 | .18 | .16 | 1.08 | 1.4 | 1.13 | 1.4 | .70 | .32 | .40 | 26 | 26 |
| | 97 | 200 | .49 | .48 | .11 | .16 | .92 | -1.5 | .88 | -1.3 | 1.30 | .47 | .41 | 49 | 49 |
| | 97 | 200 | .49 | .48 | .11 | .16 | 1.00 | .0 | .99 | .0 | 1.00 | .40 | .41 | 60 | 60 |
| | 98 | 200 | .49 | .49 | .09 | .15 | .95 | -1.0 | .97 | -.3 | 1.19 | .44 | .40 | 22 | 22 |
| | 99 | 200 | .50 | .49 | .07 | .16 | .88 | -2.3 | .84 | -1.8 | 1.43 | .50 | .40 | 42 | 42 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 102 | 200 | .51 | .51 | -.01 | .16 | .88 | -2.4 | .82 | -2.1 | 1.46 | .51 | .40 | 35 35 |
| | 104 | 200 | .52 | .52 | -.06 | .15 | .97 | -.6 | .90 | -1.1 | 1.17 | .43 | .39 | 12 12 |
| | 105 | 200 | .52 | .53 | -.08 | .16 | .99 | -.1 | .99 | -.1 | 1.02 | .40 | .40 | 32 32 |
| | 105 | 200 | .52 | .53 | -.08 | .15 | 1.06 | 1.1 | 1.07 | .8 | .77 | .34 | .39 | 28 28 |
| | 107 | 200 | .54 | .54 | -.13 | .16 | .91 | -1.8 | .90 | -1.1 | 1.31 | .47 | .40 | 54 54 |
| | 111 | 200 | .56 | .57 | -.22 | .16 | .97 | -.6 | .94 | -.6 | 1.13 | .42 | .39 | 25 25 |
| | 112 | 200 | .56 | .57 | -.25 | .16 | 1.04 | .7 | 1.03 | .3 | .88 | .36 | .39 | 30 30 |
| | 113 | 200 | .56 | .58 | -.27 | .16 | 1.01 | .2 | 1.06 | .6 | .93 | .38 | .40 | 31 31 |
| | 117 | 200 | .58 | .60 | -.37 | .16 | .94 | -1.0 | .92 | -.8 | 1.20 | .43 | .38 | 13 13 |
| | 118 | 200 | .59 | .61 | -.39 | .16 | .98 | -.3 | .91 | -.8 | 1.10 | .41 | .38 | 7 7 |
| | 118 | 200 | .59 | .61 | -.39 | .16 | 1.19 | 3.3 | 1.28 | 2.6 | .31 | .20 | .38 | 11 11 |
| | 123 | 200 | .62 | .64 | -.52 | .16 | 1.01 | .2 | 1.04 | .3 | .95 | .37 | .39 | 47 47 |
| | 124 | 200 | .62 | .64 | -.54 | .16 | 1.02 | .3 | 1.03 | .2 | .94 | .36 | .38 | 3 3 |
| | 124 | 200 | .62 | .64 | -.54 | .16 | 1.00 | .0 | 1.01 | .0 | 1.01 | .38 | .38 | 16 16 |
| | 125 | 200 | .63 | .65 | -.57 | .16 | 1.06 | .9 | 1.14 | 1.3 | .80 | .33 | .39 | 34 34 |
| | 126 | 200 | .63 | .65 | -.59 | .16 | .98 | -.3 | .93 | -.6 | 1.08 | .40 | .38 | 23 23 |
| | 127 | 200 | .63 | .66 | -.62 | .16 | 1.05 | .8 | 1.07 | .6 | .85 | .33 | .37 | 9 9 |
| | 128 | 200 | .64 | .67 | -.65 | .16 | .89 | -1.8 | .81 | -1.7 | 1.33 | .48 | .38 | 45 45 |
| | 130 | 200 | .65 | .68 | -.70 | .16 | .95 | -.8 | .93 | -.5 | 1.13 | .42 | .38 | 51 51 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 132 | 200 | .66 | .69 | -.75 | .16 | 1.05 | .7 | 1.02 | .2 | .89 | .33 | .37 | 21 21 |
| 133 | 200 | .67 | .69 | -.78 | .16 | 1.00 | .0 | .95 | -.3 | 1.03 | .39 | .38 | 50 50 |
| 135 | 200 | .68 | .70 | -.83 | .16 | .99 | -.1 | .94 | -.4 | 1.04 | .38 | .37 | 1 1 |
| 135 | 200 | .68 | .71 | -.83 | .16 | 1.08 | 1.1 | 1.13 | 1.0 | .82 | .31 | .38 | 36 36 |
| 137 | 200 | .69 | .72 | -.89 | .17 | .93 | -.9 | .84 | -1.3 | 1.18 | .44 | .37 | 44 44 |
| 138 | 200 | .69 | .72 | -.91 | .17 | .90 | -1.4 | .79 | -1.6 | 1.25 | .47 | .37 | 41 41 |
| 140 | 200 | .70 | .73 | -.97 | .17 | .97 | -.3 | .97 | -.1 | 1.05 | .39 | .37 | 39 39 |
| 144 | 200 | .72 | .76 | -1.08 | .17 | .88 | -1.4 | .80 | -1.4 | 1.23 | .47 | .36 | 33 33 |
| 155 | 200 | .77 | .81 | -1.42 | .18 | 1.04 | .4 | 1.10 | .6 | .92 | .28 | .34 | 4 4 |
| 101.7 | 200.0 | .51 | .51 | .00 | .16 | 1.00 | -.1 | 1.00 | .0 | | .39 | | Mean (Count: 60) |
| 24.6 | .0 | .12 | .14 | .62 | .01 | .07 | 1.1 | .12 | 1.1 | | .06 | | S.D. (Population) |
| 24.8 | .0 | .12 | .15 | .63 | .01 | .07 | 1.2 | .12 | 1.1 | | .07 | | S.D. (Sample) |

Model, Populn: RMSE .16 Adj (True) S.D. .60 Separation 3.76 Strata 5.35 Reliability .93
Model, Sample: RMSE .16 Adj (True) S.D. .61 Separation 3.79 Strata 5.39 Reliability .94
Model, Fixed (all same) chi-square: 830.1 d.f.: 59 significance (probability): .00
Model, Random (normal) chi-square: 55.1 d.f.: 58 significance (probability): .58

# Appendix G
# FACETS Condition Measurement Reports

Input Measurement Report (arranged by mN).

| Total Score | Total Count | Obsvd Average | Fair(M) Average | Model Measure | Model S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim. Discrm | Correlation PtMea | Correlation PtExp | N Input |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2909 | 6000 | .48 | .48 A | .00 | .03 | .99 | -1.3 | 1.00 | .1 | 1.03 | .46 | .45 | 1 Audio |
| 3193 | 6000 | .53 | .54 | -.24 | .03 | 1.01 | .9 | 1.01 | .3 | .97 | .44 | .45 | 2 Video |
| 3051.0 (Count: 2) | 6000.0 | .51 | .51 | -.12 | .03 | 1.00 | -.2 | 1.00 | .2 | | .45 | | Mean |
| 142.0 (Population) | .0 | .02 | .03 | .12 | .00 | .01 | 1.1 | .00 | .1 | | .01 | | S.D. |
| 200.8 (Sample) | .0 | .03 | .04 | .17 | .00 | .02 | 1.6 | .00 | .1 | | .01 | | S.D. |

Model, Populn: RMSE .03 Adj (True) S.D. .12 Separation 3.99 Strata 5.65 Reliability .94
Model, Sample: RMSE .03 Adj (True) S.D. .17 Separation 5.73 Strata 7.97 Reliability .97
Model, Fixed (all same) chi-square: 33.8 d.f.: 1 significance (probability): .00

Notetaking Measurement Report (arranged by mN).

| Total Score | Total Count | Obsvd Average | Fair(M) Average | Model Measure | Model S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim. Discrm | Correlation PtMea | Correlation PtExp | N Notetaking |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2973 | 6000 | .50 | .49 A | .00 | .03 | 1.02 | 1.8 | 1.04 | 2.0 | .93 | .43 | .45 | 1 Hand |
| 3129 | 6000 | .52 | .53 | -.13 | .03 | .97 | -2.2 | .97 | -1.6 | 1.07 | .47 | .45 | 2 typed |
| 3051.0 (Count: 2) | 6000.0 | .51 | .51 | -.06 | .03 | 1.00 | -.2 | 1.00 | .2 | | .45 | | Mean |
| 78.0 (Population) | .0 | .01 | .02 | .06 | .00 | .02 | 2.1 | .04 | 1.8 | | .02 | | S.D. |
| 110.3 (Sample) | .0 | .02 | .02 | .09 | .00 | .03 | 2.9 | .05 | 2.6 | | .03 | | S.D. |

Model, Populn: RMSE .03 Adj (True) S.D. .06 Separation 1.96 Strata 2.94 Reliability .79
Model, Sample: RMSE .03 Adj (True) S.D. .09 Separation 2.94 Strata 4.25 Reliability .90
Model, Fixed (all same) chi-square: 9.6 d.f.: 1 significance (probability): .00

# Appendix H
# Item Logit and Standard Error Values by Condition

| | Audio Input | | Video Input | | Handwritten Notes | | Typed Notes | |
|---|---|---|---|---|---|---|---|---|
| **Item** | **Logit** | ***SE*** | **Logit** | ***SE*** | **Logit** | ***SE*** | **Logit** | ***SE*** |
| **A1** | –0.36 | 0.47 | +0.17 | 0.46 | –0.34 | 0.46 | +0.19 | 0.47 |
| **A2** | –0.48 | 0.47 | 0.00 | 0.43 | +0.62 | 0.43 | +1.10 | 0.47 |
| **A3** | –0.26 | 0.46 | –0.07 | 0.44 | –0.42 | 0.44 | –0.24 | 0.46 |
| **A4** | +0.50 | 0.58 | –2.02 | 0.55 | +0.49 | 0.59 | –2.03 | 0.54 |
| **A5** | –0.73 | 0.46 | +0.96 | 0.46 | –0.82 | 0.45 | +0.88 | 0.47 |
| **A6** | –0.46 | 0.46 | +0.66 | 0.45 | –0.43 | 0.45 | +0.68 | 0.47 |
| **A7** | –0.07 | 0.45 | +0.69 | 0.44 | –0.58 | 0.43 | +0.18 | 0.46 |
| **A8** | –0.80 | 0.46 | +0.97 | 0.46 | –1.11 | 0.45 | +0.66 | 0.46 |
| **A9** | +0.81 | 0.45 | –0.93 | 0.48 | +1.39 | 0.47 | –0.35 | 0.46 |
| **A10** | +0.74 | 0.45 | +0.66 | 0.45 | –0.43 | 0.45 | –0.51 | 0.45 |
| **A11** | –0.27 | 0.45 | +0.89 | 0.45 | –0.77 | 0.43 | +0.39 | 0.46 |
| **A12** | +0.37 | 0.44 | –0.19 | 0.43 | +0.24 | 0.43 | –0.32 | 0.45 |
| **A13** | –0.30 | 0.44 | –0.56 | 0.45 | +0.45 | 0.44 | +0.19 | 0.45 |
| **A14** | –0.58 | 0.47 | +0.07 | 0.44 | +0.25 | 0.43 | +0.90 | 0.47 |
| **A15** | +0.38 | 0.44 | –1.92 | 0.48 | +0.34 | 0.43 | –1.96 | 0.50 |
| **A16** | +0.05 | 0.45 | –0.56 | 0.45 | +0.07 | 0.44 | –0.54 | 0.46 |
| **A17** | +0.73 | 0.45 | –0.43 | 0.44 | +0.35 | 0.43 | –0.81 | 0.46 |
| **A18** | +0.07 | 0.46 | –0.22 | 0.44 | +0.54 | 0.44 | +0.25 | 0.46 |
| **A19** | +0.39 | 0.52 | +0.56 | 0.45 | +0.67 | 0.49 | +0.83 | 0.48 |
| **A20** | –0.78 | 0.49 | +1.09 | 0.50 | –0.97 | 0.49 | +0.90 | 0.49 |
| **A21** | +0.79 | 0.46 | –0.69 | 0.47 | +0.78 | 0.46 | –0.70 | 0.47 |
| **A22** | –0.26 | 0.45 | +0.76 | 0.44 | –0.22 | 0.43 | +0.81 | 0.46 |
| **A23** | –0.26 | 0.45 | +0.35 | 0.45 | –0.32 | 0.44 | +0.29 | 0.46 |
| **A24** | –0.34 | 0.44 | –0.50 | 0.44 | –0.12 | 0.43 | –0.29 | 0.46 |
| **A25** | –0.58 | 0.45 | –0.28 | 0.43 | –0.41 | 0.43 | –0.11 | 0.45 |
| **A26** | +0.62 | 0.45 | –0.30 | 0.43 | +0.62 | 0.43 | –0.30 | 0.45 |
| **A27** | +0.20 | 0.45 | +0.36 | 0.31 | –0.03 | 0.44 | +0.22 | 0.45 |
| **A28** | +0.50 | 0.45 | –0.59 | 0.44 | +0.06 | 0.43 | –1.03 | 0.46 |
| **A29** | +0.41 | 0.45 | +0.07 | 0.44 | +0.35 | 0.44 | +0.01 | 0.45 |
| **A30** | +0.87 | 0.45 | –0.27 | 0.44 | +0.80 | 0.43 | –0.34 | 0.45 |
| **B1** | –0.15 | 0.43 | +0.11 | 0.45 | –0.06 | 0.45 | +0.21 | 0.44 |
| **B2** | +0.49 | 0.45 | –0.11 | 0.46 | +1.18 | 0.47 | +0.58 | 0.44 |
| **B3** | +0.55 | 0.47 | –0.12 | 0.50 | +0.34 | 0.48 | –0.32 | 0.50 |
| **B4** | +0.02 | 0.44 | –1.02 | 0.48 | +0.87 | 0.48 | –0.17 | 0.44 |
| **B5** | +0.88 | 0.45 | –0.59 | 0.45 | +1.29 | 0.47 | –0.18 | 0.43 |
| **B6** | +0.62 | 0.46 | +0.11 | 0.48 | +0.29 | 0.46 | –0.22 | 0.48 |
| **B7** | –0.52 | 0.47 | +0.04 | 0.45 | –0.20 | 0.47 | +0.36 | 0.45 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **B8** | +0.47 | 0.45 | –0.02 | 0.44 | +0.23 | 0.46 | –0.27 | 0.43 |
| **B9** | –0.89 | 0.48 | –0.23 | 0.48 | –0.69 | 0.50 | –0.04 | 0.46 |
| **B10** | +0.03 | 0.47 | –0.09 | 0.47 | –0.32 | 0.49 | –0.45 | 0.45 |
| **B11** | –0.20 | 0.45 | +0.25 | 0.50 | 0.00 | 0.47 | +0.45 | 0.48 |
| **B12** | +0.15 | 0.44 | +0.45 | 0.45 | –0.29 | 0.45 | +0.01 | 0.43 |
| **B13** | –0.42 | 0.50 | +0.57 | 0.53 | –1.18 | 0.54 | –0.20 | 0.49 |
| **B14** | +0.31 | 0.45 | 0.00 | 0.49 | +0.20 | 0.47 | –0.11 | 0.48 |
| **B15** | +0.60 | 0.44 | +0.70 | 0.49 | +0.25 | 0.45 | +0.35 | 0.49 |
| **B16** | –0.30 | 0.47 | +1.18 | 0.48 | –0.96 | 0.50 | +0.52 | 0.45 |
| **B17** | +0.12 | 0.44 | –0.22 | 0.47 | +0.48 | 0.46 | +0.14 | 0.45 |
| **B18** | +0.17 | 0.44 | +0.56 | 0.45 | –0.20 | 0.46 | +0.19 | 0.43 |
| **B19** | –0.03 | 0.44 | –0.14 | 0.44 | +0.02 | 0.45 | –0.09 | 0.43 |
| **B20** | –0.30 | 0.45 | –0.89 | 0.47 | +0.13 | 0.48 | –0.45 | 0.44 |
| **B21** | +0.60 | 0.45 | –0.46 | 0.48 | +0.72 | 0.47 | –0.34 | 0.46 |
| **B22** | –0.31 | 0.44 | +0.26 | 0.44 | –0.29 | 0.45 | +0.28 | 0.43 |
| **B23** | –0.52 | 0.44 | +0.02 | 0.45 | –0.80 | 0.45 | –0.27 | 0.43 |
| **B24** | –0.16 | 0.44 | –0.20 | 0.45 | +0.65 | 0.46 | +0.29 | 0.43 |
| **B25** | –0.61 | 0.45 | +0.66 | 0.45 | –0.61 | 0.46 | +0.66 | 0.44 |
| **B26** | –0.93 | 0.46 | +0.06 | 0.45 | –0.73 | 0.46 | +0.26 | 0.46 |
| **B27** | –0.61 | 0.44 | +0.01 | 0.45 | –0.60 | 0.45 | +0.02 | 0.43 |
| **B28** | –0.42 | 0.51 | +0.05 | 0.54 | –0.90 | 0.53 | –0.43 | 0.51 |
| **B29** | –0.63 | 0.47 | +0.48 | 0.45 | –0.20 | 0.46 | +0.90 | 0.45 |
| **B30** | +0.42 | 0.44 | –0.21 | 0.44 | –0.19 | 0.45 | –0.82 | 0.43 |

# Appendix I
# Inductively Developed Thematic Categories

| Opinion Code | Thematic Category | Key Terms, Phrases, or Ideas | Sample Response |
|---|---|---|---|
| Q1. Which of lecture styles did you prefer, the audio or the audio with video? Why? | | | |
| **Video** | --- | Video, audio with video, audio with picture | --- |
| | *Aided in comprehension* | Understand, unfamiliar, key words | I prefer audio with video because it helps me understand more being about to see pictures and diagrams. |
| | *Provided enhanced focus* | Focus, concentrate, pay attention | I prefer the audio with picture. Because first of all the picture make me pay attention. Second, it is easier for me to understand the lectures with picture and information showing. |
| | *Created greater authenticity* | Classroom, imagine, realistic, felt real | I prefer the video because I can imagine myself better to be in a classroom. |
| | *Easier to get back on track when lost* | Get lost, keep up, catch up, (key words AND catch up), miss [something] | Audio with video. It gives more material to catch up if I miss anything. |
| **Audio** | --- | Only audio, audio-only, audio lecture | --- |

| | | | |
|---|---|---|---|
| | *Easier to focus* | Attention, focus, concentrate, pay attention | I prefer lecture because I can be more focused. Somehow, the slides with only letters are distracting. If the slides were shoing only pictures, it would be helpful. |
| | *Difficult due to note-taking type* | Any reason with reference to note-taking condition | Audio because I can take note using my pen. |
| **Both** | --- | Both, Either, No preference | Both, because I'm an auditory learner. It's enough so long as the information is given clearly and I can take notes on it. |

Q2. Did you find it difficult to take notes while the video was playing? If so, why? Please explain the nature of any difficulties you had.

| | | | |
|---|---|---|---|
| **Video did not distract** | --- | No, it helped, was helpful | --- |
| | *Visual aids were helpful* | Visuals, PowerPoint, ppt, aids, slides, gestures, information, take notes | No, it helped; it provided me the high-level points to summarize. |
| | *Made lectures authentic* | Felt real, natural, in school | It is a common way in all class in school, so it was natural |
| **Video Distracted** | --- | Yes, It was ok, but…, Difficult | --- |
| | *Focused more on visuals than listening* | Ppt, distracted, reading, copying notes | Yes…because I am trying to copy the words from ppt, and forgot what teacher said… |

| | | | |
|---|---|---|---|
| | *Unable to divide attention* | Notes with movie/video, cannot focus, look more at… | It difficult to take notes when video was playing. I cannot focus on what her talking about. |
| | *Video lectures were faster* | Video faster, speaking fast, audio slower | Difficult. Speaking more fast then listening with no picture. |
| | *Context videos distracting* | Texts on wall, distract | Some texts on the wall, "do not drink…" that distract me from the lecture make video difficult to concentrate |

Q3. Did the presence of video make it easier to remember lecture information, more difficult to remember lecture information, or have no effect on your memory?

| | | | |
|---|---|---|---|
| **Aided Recall** | --- | Yes, Helped, Remember, Recall easier | --- |
| | *Slide images activated image memory* | Image memory, slides, pictures, helped answer, visual facts, easier to remember | The pictures did help. I could remember the pictures more and this helped me anser the questions. |
| | *Allowed visualization of lecture during recall* | Imagine, visualize, rebuild lecture | Yes, I could imagine easier. |
| **Did not aid recall** | --- | No video, not really, hard, difficult, only remember ppt, easier with no video | --- |

| | | | |
|---|---|---|---|
| | *Lack of concentration hindered later recall* | Concentrate, information, focus, make connections, lose focus | I can only remember what shows on ppt. only a little form what teacher said. And it was hard to make connection to the sentence form ppt. |
| | *Video helped concentration but not recall* | Remember, later, not cannot during question | First I concentrated, so I can remember about that, but I cannot final questions. |
| | Listening material was too difficult | Difficult information, terminology, cannot understand | Not really. Because I cannot understand the terminology so I cannot keep tracking in memorization. |
| **Not sure** | --- | Maybe | --- |
| | *Only some pictures were helpful* | Some pictures, helpful, words, remember, not easy | Maybe, I am not sure. I think some pictures helpful, but words weren't easy to remember. |

Q4. What did you find yourself paying the most attention to in the video lecture?

| | | | |
|---|---|---|---|
| **N/A** | | | |
| | *PowerPoint slides* | Slides, ppt, PowerPoint, screen, pictures, text | Mostly the screen because there are key points on the screen, so it is helpful to understand what they talking about the time. |

| | | | |
|---|---|---|---|
| | *Instructor's gestures and speech* | Hands, face, mouth, accent (i.e., emphasis), intonation, pitch | The content, I liked the way the lecturer explained beyond the slides of the presentation. Also she uses her hands for emphasizing, it was good but sometimes distracting. The slides were thorough. |
| | *Subject matter of the lecture* | Content, Topic, Comparison of specific topics, interest, | I was paying the most attention towards the lecture about alien life because it was more interesting compared to others. |
| | *Listening for key words, NOT visuals* | Key words, questions, words | I focused on the questions in the lectures, such as HOW, WHY, WHEN |
| | *Did not look at video* | Not look, avoid looking | I avoided looking at it. |

Q5. Which method of note-taking do you prefer to use while listening to a lecture or while listening on a test, handwriting or typing? Why?

| | | | |
|---|---|---|---|
| **Handwriting** | --- | Handwriting, Prefer Handwriting, Like to write down | --- |
| | *Speed was increased* | Faster, easier, used to handwriting, slow to type | Yes, I prefer handwriting because it is easier than typing because I'm too slow to type English. |
| | *Aided in memory* | Easier to Remember, Recall, Memorize faster | I like to write down the things the teacher says that are not written on the powerpoints. I feel I then will remember the topics easier. |

| | | | |
|---|---|---|---|
| | *Provides better platform for taking notes* | Organize, Draw Arrows, Use L1, Write in different directions, Draw circles | By handwriting. I like to use Korean and English and draw some picture in my paper. |
| | *Greater Comfort* | More comfortable, comfort | I feel more comfortable by handwriting. It is easier to control than typing, but typing is faster than handwriting. |
| | *Noise of typing* | Noise | I am not good at typing and if everyone types, there might be more noise. Noise distraction from others may be another factor. |
| | *Classroom policies and Personal Practice* | Instructor preference, [expressing a manner of inconvenience] | By handwriting, because I don't like to bring my heavy computer to class. |
| **Typing** | --- | Typing, Typing is easier, Prefer typing | --- |
| | *Greater speed and facility* | Faster, Easier to organize, adding information, easier to focus | Typing. It makes it easier for me to focus on the lecture or screen. |

| | | | |
|---|---|---|---|
| | *Easier to read later* | Read clearer, bad handwriting, neater notes | I prefer typing them because I can read it clearer due to my bad hand writing and I don't have to worry about losing it. |
| | *Greater comfort* | Same as handwriting | Typing. I feel more comfort using a computer. |
| **Both** | --- | --- | --- |
| | *Dependent upon Context* | Depends on class | It depends on the class. If the class is in small class, I would hand write, however if the class is in lecture style, I would prefer to take notes by typing…… |