VALIDATION OF AN ACADEMIC LISTENING TEST: EFFECTS OF "BREAKDOWN" TESTS AND TEST TAKERS' COGNITIVE AWARENESS OF LISTENING PROCESSES

BY

YOUNGSHIN CHI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Fred G. Davidson, Chair
Professor Carolyn Anderson
Assistant Professor Christopher Grindrod
Assistant Professor Peter Golato

Abstract

This study investigated the breakdown effect of a listening comprehension test, whether test takers are affected in comprehending lectures by impediments, and collected test takers' cognitive awareness on test tasks which contain listening breakdown factors how they perceived these impediments. In this context of the study, a "Breakdown" is a test task that contains impeding factors which could limit test takers' listening comprehension. Impeding factors used in this study are British accent, fast speech rate, and noise. The listening comprehension test contained four different types of lectures with two topics: dinosaurs and weather changes. In total, ninety-six test takers took the test and each group of twenty-four test takers were assigned to following test groups: the "regular", the "British", the "speech rate", and the "noise" groups. The name of each group represented the source of interference or impediment in the listening task. Along with test takers' performance on listening test, the design of the study was informed by test takers' responses to a questionnaire on their cognitive awareness of listening, metacognitive process, and test-taking strategy uses.

Both quantitative and qualitative analyses were used on test scores and responses on questionnaire. To estimate validity and reliability of the test, several quantitative analyses were conducted. Cronbach's alphas and confidence intervals of alphas were calculated for each test and the coefficient values were observed to vary among the tests. A Generalizability study and an analysis of variance (ANOVA) were conducted to measure the breakdown effect on the test. The result of generalizability study showed that variance components for interaction between person and items were larger than variance components for person, which indicated confounding effect. However, the analysis of variance showed that breakdown effect on test was not detected, but the effect of talk types was significantly different. Classical item analyses were also

conducted to observe differences on item functions among four tests. A few items were flagged as easy items, however difficulty and discrimination indices were varied among groups. To understand test takers' perspectives on second language listening, both quantitative and qualitative analyses were conducted. Test takers' responses on questionnaire showed test takers from different groups demonstrated similar metacognitive strategy uses, monitoring behaviors, and test-taking strategies.

Findings showed that there was no intentional breakdown effect on all groups. Low reliability coefficient and generalizability analysis showed that there was insignificant evidence to support breakdown effect of three impediments on test takers' listening proficiency. However, one of impediments, noise functioned as an impediment in the test. Some test takers from the noise group reported that their comprehension was impeded by noise. At the test level, talk type was identified as breakdown factor from both quantitative and qualitative analyses. At item level, regardless of item difficulty indices, lower item discrimination indices indicated that items functioned differently when an impediment was added in the test.

Listening passages functioned as a more influential factor in comprehension than the breakdown factors. Test takers who already had some background knowledge could perform better than those who do not have the knowledge. Furthermore, test takers with higher scores were able to use various types of metacognitive process than those with lower scores regardless of groups.

This study suggests that several considerations are needed to make authentic listening tasks when listening impediments are included in the second language test development such as degrees of impediments, threshold of the breakdown factor, and breakdown effects on both item

and test levels. Moreover, metacognitive process embedded listening items could contribute to understand second language learners' listening comprehension.

To my family

Acknowledgments

This dissertation was a long and challenging journey for me, and many people contributed to it. I owe my gratitude to all those people who have made this dissertation possible.

I would like to express my deepest gratitude to my advisor, Professor Fred Davidson. I have been fortunate to have an advisor who taught me how to explore on my own and gave me great advice whenever I needed. He showed me the world of language testing, and his patience and support helped me overcome many difficult situations and kept me focused on my study. I learned from him how to apply knowledge to practice, how to manage working relationships with colleagues, and how to think as a language tester. I am forever indebted to him for his continuous encouragement and guidance.

I wish to thank Professor Carolyn Anderson for inspiring and encouraging me to design the study. Without her comments and support it would have been very difficult to complete the dissertation. I sincerely thank to her for her support and encouragement.

My thanks go out to Professor Christopher Grindrod for insightful comments and constructive criticism at different stages of my research. His comments were thought-provoking and helped me focus on the idea of breakdown.

I am also grateful to Professor Peter Golato who helped me to think about the fundamental issues of second language listening from a psycholinguistic perspective. The question that he raised on my preliminary exam kept me on track to finish the study.

This dissertation would not have been possible unless Ann Spear supported me. She was my first writing teacher, and not only did she advise me on writing, but she also taught me to

think critically. I am thankful to her for encouraging the use of correct grammar and for commenting on revisions of this manuscript.

I am thankful to my advisors in Korea, Dr. Yongsoon Kang and Dr. Haemoon Lee for their support and care. I deeply appreciate their belief in me.

I would like to acknowledge Liam Moran and Kevin Kurasch who helped me to record the listening prompts for my test. Without their help, I could not have created the test.

Most importantly, none of this would have been possible without the support of my family: my parents, Nakyum, Kanghoon, and Tate. My parents and my sister Nakyum always prayed for me and showed endless support. Without their belief in me, I would never have been able to start or finish this long journey. I want to express gratitude and love to my husband, Kanghoon. He is my greatest ally. His support and care helped me overcome hindrances and stay focused on my study. I also want to send loves to my son, Tate. He craved time with me and disliked seeing me in front of the computer so much, but at the same time, he reminded me how much my dream and career are valuable to me. He was a great energizer and motivation for me to complete this dissertation.

Many thanks to the 96 volunteer test takers who participated in the study and shared their valuable opinions. I would like to express my great thanks to my awesome advocates: Angela Park, Sun Joo Chung, Michelle Moon, Jiyoung Kim, Jungsun Kim, Carsten Wilmes, Rick Partin, and members of the UIUC Foreign Language Assessment Group (FLAG) and Language Testing Research Group (LTRG). Your support and feedback were crucial to my study.

Finally, I appreciate the financial support from the ETS TOEFL Small Grants for Doctoral Research in Second Language Assessment that funded parts of the research discussed in this dissertation.

Table of Contents

CHAPTER 1: INTRODUCTION 1
Issues in Second Language Listening Tests 5
Definition of Breakdown 8
Rationale for the Study 8
Research Questions 9

CHAPTER 2: LITERATURE REVIEW 10
Chapter Overview 10
Views on Test Validation 10
Theoretical Background on Listening Comprehension 17
Issues in Developing Second Language Listening Tests 29

CHAPTER 3: METHODOLOGY 38
Chapter Overview 38
Design of the Study 38
Data Collection 39
Methodologies for Different Levels of Inference 46

CHAPTER 4: RESULTS 51
Chapter Overview 51
Preliminary Stage of Test Development 52
Findings for Different Levels of Inference 56

CHAPTER 5:DISCUSSION AND CONCLUSIONS 98
Chapter Overview 98
Findings and Discussions on Levels of Inferences 98
Limitations and Suggestions for Future Research 119

REFERENCES 123

Appendix A Test Specifications 130
Appendix B Needs Analysis 138
Appendix C Test Form 145
Appendix D Test Booklet 152
Appendix E Cognitive Questionnaire 157
Appendix F Reflection Questions 159
Appendix G Item Statistics 160

List of Tables

Table		Page
1	*Interpretive Argument for a Trait Interpretation (Kane, 2006)*	16
2	*Average Speech Rates for English (Tauroza and Allison, 1990)*	21
3	*Overview of the Research Activities*	39
4	*Participants' Length of Residency in US by Groups*	42
5	*Features of Listening Comprehension*	54
6	*One-Way ANOVA for Students' TOEFL Scores*	57
7	*Descriptive Statistics for Test Scores*	58
8	*Reliability Analyses for Four Tests*	59
9	*Standard Error of Measurements for All Tests*	60
10	*Frequency Table for Test Scores*	61
11	*Generalizability Study (p x i Design)*	62
12	*4 x 4 Latin Square Design*	64
13	*Result of 4 x 4 Latin Square Design*	65
14	*Tukey Comparison for Talk Type*	66
15	*Item Difficulty Analysis for Four Tests (p-value of Items)*	68
16	*Item Discrimination Analysis for Four Tests*	71
17	*Self-Assessment on Second Language Listening*	76
18	*Cognitive Awareness on Second Language Listening*	77
19	*Listening Process Approaches*	79
20	*Metacognitive Process in Second Language Listening: Planning*	81
21	*Metacognitive Process in Second Language Listening: Monitoring*	83

22 *Metacognitive Process in Second Language Listening: Strategy Uses* 86

23 *Metacognitive Process in Second Language Listening: Strategy Uses* 87

24 *Test-Taking Strategy Use in Second Language Listening* 89

25 *MIT Free Online Lecture 1* 138

26 *MIT Free Online Lecture 2* 138

27 *BBC Learning English Radio 1* 138

28 *BBC Learning English Radio 2* 139

29 *Princeton University Lecture* 139

30 *China Now Lecture Series: Can a Green Dragon Fly? China's Energy Challenges and Opportunities* 139

31 *Prions: A New Principle of Disease* 140

32 *Debate 1: The Seat* 140

33 *Debate 2: Fast Food* 140

34 *Discussion: Accident* 141

35 *News: Men More Attracted to Women in Red* 141

36 *Voanews: American History* 142

37 *News: Foreign Student Series-Getting a US Education from Home* 142

38 *Lecture: Dinosaurs* 143

39 *Lecture: Twister* 143

40 *Halloween Solar Storm Anniversary* 144

41 *Saturn's Cyclones* 144

List of Figures

Figure Page

1. Gender ratio of participants 40

2. Participants' academic status 41

3. Participants' years of residence in US 42

4. Participants' first language 43

5. Iterative process of test specifications 56

6. Frequency distribution of test scores 63

7. Item comparison of the regular group 72

8. Item comparison of the British group 72

9. Item comparison of the speech rate group 73

10. Item comparison of the noise group 73

Chapter 1

Introduction

Our listening comprehension supplies us a huge range of knowledge about our world, and our degree of skill in decoding aural information, especially spoken language, often determines our safety, our emotional and intellectual growth, and the success or failure of our social endeavors. In everyday life, however, we often encounter situations that cause our comprehension to break down because listening is a receptive skill that can be disrupted by competing incoming information. For instance, in an airplane, the pilot's announcement may be incomprehensible because the engine noise is too loud. Or the details of a speaker's message may not be clear to a listener unfamiliar with the topic. Why do people often encounter listening comprehension breakdown? Is it because of their listening proficiency, or is their listening proficiency impeded by some other listening factors? If this listening comprehension breakdown exists in the first language, how, and how often, does comprehension breakdown occur in the second language?

Listening comprehension breakdown is common enough in everyday life, but it can be critical in academic settings. At the university level, international students must listen and interact in English, their second language, and they often face listening challenges that greatly exceed those of their first language. When the listening environment includes new topics and unfamiliar formats of lectures as well as delivery in a second language, comprehension of the content of the message can be exponentially more difficult. For instance, learners' listening process may not fully activate if the topic of a lecture is new to them because they cannot use their prior knowledge to support their comprehension. Or they may need time to adjust to speakers' varied speaking styles, because even though a speaker produces simply structured

language, listeners might miss the information due to an unfamiliar accent or speech rate. For serious college students, this unexpected circumstance can increase anxiety and inactivate their listening process so they will miss crucial key words and concepts and misunderstand the information.

Although listening comprehension breakdown might be derived from the definition of listening comprehension, listening comprehension has not yet been defined among listening researchers (Wagner, 2002; Buck, 1999; Rubin, 1994). Exclusion or inclusion of "response" by a listener, listeners' memory and their comprehension strategies in a definition is still controversial because tangible evidence of these elements is difficult to capture. In addition, listeners' listening behaviors are unique, and this factor confounds our understanding of what listening comprehension really is. Since the relationship between these controversial elements and listening comprehension is not clearly explained, it is very difficult to define how second language listeners process the aural information they receive. Due to the unclear definition of listening, possible causes or reasons for listening comprehension breakdown have not yet been addressed in research.

Why are learners affected by these listening features when they attempt to comprehend messages delivered in their second language? Unlike processing first language texts, second language listening comprehension can be even more negatively influenced by impediments such as noise, accents, fillers, pauses, and fast speech rates. The biggest reason why second language learners are affected by listening impediments is that their second language listening competence is not advanced enough to overcome them, and this deficiency results in faulty perception and disrupted attention span.

Since second language listening comprehension can so easily be reduced by various listening features, listeners are required to learn strategies to overcome comprehension breakdown. When listeners listen to messages in their first language, they have an innate competency to comprehend; however, listeners must use a conscious process for second language listening, and so applying strategies to activate their schema should support and enhance their comprehension. Indeed, it has been shown that most listeners automatically rely on their prior knowledge when they process real time spoken information, but second language listeners need to actively invoke it (Flowerdew & Miller, 2005; Rubin, 1994). In addition, while L2 listeners often rely on bottom-up processing strategies, Vandergrift (2007) has reported that learners using top-down processing comprehend better than those using bottom-up processing. Therefore, a learner with an effective strategy would succeed in second language comprehension and would perform better than one who does not acquire an adequate strategy.

Though listening is known as an important language skill to communicate and understand language, it is neglected in second language curricula despite its role in learning and teaching. For second language teaching, research findings on reading comprehension have been applied in an attempt to enhance second language learners' listening comprehension. For example, because it is recommended that readers take notes about what they read, teachers emphasize and concentrate on note-taking exercises which use the information from note-taking for a follow-up activity (Flowerdew & Miller, 1997). The same type of activity is used for listening practice because learning how to get the gist of aural information and anticipate missing information could help students to understand the complexity of lectures. These learning tasks are mostly based on learning strategies for understanding academic lectures. However, they do not fully reflect the real academic situation because students must not only develop note-taking skills, but

also discussion, debate and presentation skills are needed in real academic settings. Furthermore, learners need the ability to process different sources of aural information, and these listening tasks fall short of providing the skills needed for the actual classroom environment and teaching materials.

This incongruity also appears in second language testing. Assessing second language listening has not been as extensively investigated compared to the domains of writing, speaking and reading because of the unique comprehension process involved. Most listening tests provide simple structured listening tasks and a restricted range of higher level listening items. Language testers have investigated the construct validity of their own tests, but they have not attended to the needs and opinions of stakeholders such as examinees, teachers, and parents, so unlike other language tests, listening tests are not fully linked to teaching and learning. Moreover, inferences drawn from listening test scores have not been effectively reflected in the curriculum or in improving students' listening capacity. Testing, learning and teaching are not fully connected because a realistic listening construct has not been adequately determined for either testing or teaching, and insufficient test inferences do not help teachers to provide appropriate and effective instruction. Language testers need a more detailed understanding of listening comprehension and listening traits in order to connect these three areas.

Findings from listening research are closely related to findings from the reading research because research on listening comprehension has depended to a great degree on findings borrowed from reading research. In particular, the general construct for listening comprehension is an adaptation of the principles of reading comprehension. However, although there is a relationship between reading and listening comprehension, listening demands distinct skills (Anderson, 1983). For example, listeners must process a speaker's information in real time

without the luxury of "looking back" or "re-listening" because unlike a reading text, a listening text exists in real time (Flowerdew, 1994). Listeners must also cope with phonological and lexico-grammatical challenges that are not a problem for readers. In a listening text, the speaker's unnecessary or redundant words, pauses and asides make it difficult for listeners to process the main idea, and understanding a segment of a listening text could require new information or material. Because of the real-time aspect, listeners cannot control the speed of the text, so they need strategies to process the meaning when they may only catch important words or retrieve the meaning. In sum, it is important to acquire effective listening strategies—different from reading strategies—to cope with the speed of text and possibly unfamiliar vocabulary and grammatical structures.

It is important for second language learners to acquire listening strategies that support and enhance their comprehension, and L2 teachers have emphasized the importance of listeners' own strategies and ways of comprehension. However, these aspects are rarely considered in test development. The reason that such strategies have not been considered is that testers aim to measure test takers' listening capacity while minimizing variance caused by impeding comprehension factors. Such factors are speakers' rate of speech, the listener's knowledge of the topic, and general knowledge of the world. To minimize the possible confounding factors, testers prefer to use multiple-choice items, monologue lectures, and structured conversations to measure second language learners' listening capacity.

Issues in Second Language Listening Tests

A valid language test would provide predictive information about how a learner will perform in a non-test setting. To examine the validity of the test, a validation process is required to collect evidence to support the inferences and decisions made on the basis of the test scores

(Cronbach, 1971; Messick, 1989). To provide valid and reliable inferences, the purpose of each test must be clearly defined. Validity and reliability are a great concern to language testers because they demonstrate the usefulness of the test. Validity has been identified as the most important quality of test use, and with it, meaningful inferences can be drawn from test scores (Bachman, 1990).

To test second language listening, it is important to reflect learners' cognitive operation and process of comprehension in the definition of a test construct. Developing a listening test construct is considered to be an important factor to ensure the validity of the test. Henning's (1992) study has suggested that there is no consensus on the best approach or technique to assess a listening construct. Buck (2001) described the practical constraints of construct definition, the foremost constraint being the purpose of the test. He also proposed the idea of the "default listening construct" that is useful and applicable to measure pragmatic and discourse knowledge. The disadvantage of Buck's construct is that it is not flexible enough to be tailored by test developers to fit individual testing contexts because he promoted the use of at least 100 listening items and longer listening tasks. He argued, however, that a longer test and many items can best measure the complete range of a test takers' listening proficiency.

Despite of the importance of the listening process, learners' cognitive operations are not considered as a factor in the test development. Learners' cognitive operations vary depending on the situation and types of aural information, and the test-taking environment especially changes the cognitive operation of listening. For example, Shang (2005) showed how test takers use metacognitive and cognitive strategies and how those strategies affect their performance. Teachers and researchers emphasize the importance of proper strategies in maximizing listening comprehension so why have strategy uses and cognitive processes been ignored in test

development? Testers and teachers provide listening tasks that reflect content features that learners should understand assuming that test scores can predict examinees' performance in non-testing situations or classroom settings. However, these tasks are often not authentic as real classroom lectures or conversations because testers have developed listening tasks with the intention of anticipating possible comprehension breakdown factors.

For a more valid language test, testers must decide on and control the degree of authenticity of audio tasks at the development stage. Authentic tasks replicate challenges faced in the real classrooms, and such tasks encourage the integration of teaching, learning and assessment because students learn in the process of performing them and are eventually able to apply target concepts to the real world. Nevertheless, highly authentic test tasks cause problems for test developers due to the construct-irrelevant variance they may introduce. Therefore, to avoid confounding variables, testers often use scripted or canned audio texts as listening input for comprehension questions.

Another problem for language testers is that unlike assessing reading, writing and speaking—which require that the learner only to read, write or speak—assessment of the listening domain requires that test takers respond with the other language skills because we cannot directly access the process of listening. Consequently, test developers prefer to use multiple choice questions to minimize the possible confounding effects from other language domains. One question for test developers is how authentic listening tasks can realistically be without intruding on a test's validity. Whether a particular listening task is authentic is not at issue. What is at issue is this question: what characteristics of the task cause comprehension to break down, and once such characteristics are articulated, what use can be made of that knowledge for test development? This is the driving force of the present proposed study.

Definition of Breakdown

"Breakdown," in the context of this study refers not to the test takers, but rather to the listening test. A "breakdown" is a test task that contains impeding factors that could limit the test takers' listening comprehension. Impeding factors used in the study are British accent, fast speech rate (reduced spectrum into 90%), and noise (15% of pink noise).

Rationale for the Study

This study attempts to understand test takers' second language listening proficiency by providing breakdown listening tests and a cognitive awareness questionnaire. The purpose of using breakdown tests is to get in a depth understanding of the effect of each breakdown listening factor and how test takers overcome intended listening impediments. Using impediments in listening tests is motivated by a central threat to language test validity, namely, the need to generalize to non-test settings. This research presumes that non-test listening settings are burdened with impediments.

This study also attempts to investigate test takers' cognitive awareness when they process aural information. Little research has been conducted on test takers' cognitive awareness in listening test situations, and few listening tests have implemented test takers' perspectives on listening or their strategy uses. This study taps their cognitive awareness of listening in attempt to suggest more appropriate listening prompts. Furthermore, the investigation of learners' cognitive strategy use and perspectives on listening could provide useful information to language testers, teachers and test takers. The inferences could be drawn from a comparison of test takers' performance and cognitive awareness between regular and breakdown test groups.

Research Questions

The following research questions for this study have been formulated to answer some of its concerns. These questions were developed according to the levels of inference from the argumentative approach to test validation. The questions for each level of inference are:

Research Question 1 [First level of inferences: Scoring]:

> To what extent does the newly developed listening comprehension test measure test takers' ability to understand academic lectures? How are the test takers' performances reflected in their test scores?

Research Question 2 [Second level of inferences: Generalization]:

> To what extent do test takers vs. test tasks contribute to the source of variance? How do the different types of tests (regular vs. breakdown) contribute to the source of variance?

Research Question 3 [Third level of inferences: Extrapolation]:

> To what extent do test takers themselves contribute to the source of variance? How do their thought processes differ between the regular group and the breakdown groups? Does their cognitive awareness impact on their test performance?

Research Question 4 [Fourth level of inferences: Implication]:

> To what extent are the implications associated with trait (second language listening) appropriate in this case? Does the evidence support the implications associated with the trait label, or should the trait (assessed by this test) be called something else?

Chapter 2

Literature Review

Chapter Overview

This chapter discusses three different topics: views on test validation, views on listening comprehension, and views on test development. The framework of my study is Kane's (2006) argument-based approach to validation. Along with this framework, a psycholinguistic view of the listening process is discussed as background for listening comprehension. Furthermore, factors affecting the listening process and listeners' use of strategies are discussed, as they may contribute to improvement of listening tasks and to a better understanding of test takers' process of listening. Finally, the chapter discusses critical issues of listening test development that make the design of a listening test challenging..

Views on Test Validation

For several decades, various models of validity have been introduced and adopted by researchers seeking better interpretations of test scores. Meanwhile, the concept of validity itself has expanded to collect evidence for better interpretation, decision, and meaningful consequences. Kane (2006) uses the term 'validation' instead of validity and highlights its uses in the field of measurement. First, validation involves the development of evidence to support the proposed interpretations and uses. Second, validation associates with an evaluation of the extent to which the proposed interpretations and uses are plausible and appropriate. Different views on test validation affect both the methodology of research and the interpretation of results.

Before Messick's validity, three concepts of validity were widely used: criterion validity, content validity and construct validity. The concept of validity has evolved from criterion validity to construct validity. The criterion model of validity was popular as the initial standard

of validity used between 1920 and 1950. Cureton (1951) defined the concept in terms of "the correlation between the actual test scores and the 'true' criterion score" (p.623), so that the test-criterion correlation corrected for unreliability in the criterion. The main focus of validation was on how well the actual task was performed, because they believed that a valid criterion provides an accurate estimate of the test. Two advantages were noticed using this model. First, criterion-related evidence is relevant to the plausibility of the proposed interpretations and uses. Second, the specified criterion helps to provide straightforward data analysis and results. However, conceptualization of a valid criterion was challenging for researchers. Ebel (1961) pointed out that the problem with the criterion model is in creating a validated criterion. Even though a second criterion could be identified as a basis for validating the initial criterion, the potential problem still remains: the criterion validation approach is tautological.

Another concept of validity was "content validity." This model interprets test scores based on a sample performance of an activity as an estimate of overall skill level in that activity. This model evolved from the criterion model in response to the question of validating one criterion by appealing to another. Two researchers proposed that a criterion be validated by establishing a rational link between the procedures used to generate criterion scores and the proposed interpretation or use of the scores (Cureton, 1951; Ebel, 1961). In practice, this model has been applied to some measures of academic achievement; however, it is limited in that it provides content-related validity evidence that is subjective and has a confirmatory bias.

Another difficulty was identified by Messick (1989) who argued that it is difficult for testers to draw interpretations of test scores because content validity evidence does not itself involve test scores. He also pointed out that the model has a limited role in test validation, mainly because it does not provide direct evidence for "inferences to be made from test scores"

(p.17). Kane (2006) also agreed that content-related evidence has a limited role despite its important role in validation. Moreover, content-related evidence has a positive role in terms of the representativeness of the test tasks and the generalizability of expected performances, but it needs other evidence to support the score interpretations that demand more than the basic interpretation.

The third model of validity is "construct" validity, a model still guide dominant for fields of measurement and testing. The concept of construct validity was first introduced by Cronbach and Meehl (1955). They offered the concept as an alternative to the criterion and content models, and suggested that it be used to measure some attributes or qualities which are not operationally defined. However, the authors did not provide a general organizing framework for validity in their paper.

The concept of construct validity has changed in the field of testing. In his seminal paper, Messick (1989) addressed dimensions of validity and how construct validity is a unitary concept in the evaluation of testing. According to Messick, validity has two dimensions: value of the assessment and function of the assessment. The value of the assessment is based on either evidence or consequence of the test. The function of the assessment provides information of interpretation or use of the assessment. Thus, construct validity provides an evidential basis for test interpretation and test use.

Caveats were raised, however about this framework of validity. Two problems were discussed concerning the evidential basis for test interpretation: construct under-representation and construct-irrelevant variance. Construct under-representation occurs when the construct does not properly measure and omits the nature of the construct. If we have a construct under-representation in the test, the test score cannot be interpreted as defined in the construct.

Construct-irrelevant variance is the inevitable test variance that we do not intend to measure. This variance can skew test scores making them irrelevant to the purpose of the test. Therefore, researchers have attempted to reduce the possible effects of these constraints on construct validity.

Fulcher (1999) emphasized that test items should be relevant to the domain and representative of the domain from Messick's (1989) view of construct validity. To ensure relevance and representativeness of construct, the process of domain specification should be pursued to provide evidence for construct validity. He also claimed that inaccuracy and inappropriate difficulty levels created systematic construct-irrelevant variance that could threat to test validity.

To further strengthen the validity of the test, Davidson and Lynch (2002) explained the construction of test specifications. Test specification could prevent the potential threats to validity. According to Davidson and Lynch (2002), a test specification is an iterative process rather than a fixed product. The specification reflects a diversity of belief forces such as culture, theory, bias, finances and other influences in a concrete and measurable manner. Based on feedback from all these different sources, the test specification can generate an operational test. However, because of its iterative characteristics, the test specification is able to continually update and improve that test.

This concept of validity is evidenced in Li's master thesis. Li (2006) explained an iterative test specification model using an audit trail. She claimed that the validity of a test will eventually be strengthened by ensuring the validity of test specifications. The validity of test specifications is made by feedback exchanges from language testing experts through versions of

test specifications. She attempted to prove that spec-driven language testing can be associated with validity.

Fulcher and Davidson (2007) described different approaches and definitions of validity and explained how validity theories have changed. They also introduced the concept of pragmatic validity that has no 'absolute' answer to the validity question. Validity arguments could have disagreements and/or other interpretations of the facts that challenge the arguments. They also suggested how to conduct a pragmatic validity investigation by deciding the appropriate explanation of the facts: *simplicity, coherence, testability,* and *comprehensiveness*. The argument should not speculate beyond the available evidence, but should allow making predictions about further interpretations based on the relationship between variables. Furthermore, the argument takes account of the available facts without minimizing unexplained phenomena.

The definition of validity framework has changed over time and many scholars have dedicated energy to the conversion. Cronbach's validity argument is to provide an overall evaluation of the intended interpretation and uses of test scores by generating a coherent analysis of all evidence. Messick's definition is similar to Cronbach's, but he claimed that validity requires adequacy and appropriateness of inference and action based on test scores (p.12). The definition of validity has now been modified again in the latest edition of the *Standards* (AREA et al. 1999) as "Validity logically begins with an explicit statement of the proposed interpretation of test scores along with a rationale for the relevance of the interpretation to the proposed use" (p.9). These definitions of validity have evolved to become the argument-based approach to validity. The argument-based approach to validity has come to reflect a general principle: by

providing both interpretive argument and validity argument, the interpretation would be more plausible and feasible.

In the argument-based validity approach, two kinds of arguments are needed in the validation process: an interpretive argument and validity argument. Kane (2006) differentiated these validity arguments:

> An *interpretive* argument specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances. The *validity* argument provides an evaluation of the interpretive argument (Cronbach, 1988). To claim that a proposed interpretation or use is valid is to claim that the interpretive argument is coherent, that its inferences are reasonable, and that its assumptions are plausible (Kane, 2006, pp. 23).

An interpretive argument gives a framework of validation by providing reasoning for the proposed interpretation and uses of test scores. If the interpretive argument includes statistical generalization, then the validity argument should evaluate the dependability of the generalization over occasions.

The current research questions are framed based on validity arguments especially for trait interpretations. For trait interpretation, an evaluation of the appropriateness of the target domain and an evaluation of the coherence and completeness of the interpretive argument are needed. The interpretive argument is outlined in the following table.

Table 1

Interpretive Argument for a Trait Interpretation (Kane, 2006)

Trait level	Description
Scoring	Scoring from observed performance to the observed score
	A1.1 The scoring rule is appropriate.
	A1.2 The scoring rule is applied as specified.
	A1.3 The scoring is free of bias.
	A1.4 The data fit any scaling model employed in scoring.
Generalization	Generalization from observed score to universe score
	A2.1 The sample of observation is representative of the universe of generalization.
	A2.2 The sample of observation is large enough to control random error.
Extrapolation	Extrapolation from universe score to target score
	A3.1 The universe score is related to the target score.
	A3.2 There are no systematic errors that are likely to undermine the extrapolation.
Implication	Implication from target score to verbal description
	A4.1 The implications associated with the trait are appropriate.
	A4.2 The properties of the observed scores support the implications associated with the trait label.

Four major inferences shown in this table are arguments for a trait interpretation. First, arguments for the scoring inference are for verifying that the scoring rule is appropriate from

observed performance to the observed score. Kane (2006) emphasized that the scoring inference relies on assumptions that the scoring criteria are appropriate and are applied as intended, that the process is free of bias, and that any statistical models (scaling, equating) employed in scoring are appropriate (p.34). Besides statistical modeling, congruence checking between test specifications, items and scoring rubrics will be helpful to collect inference.

Generalization inference can be extended whether the sample of observations is representative of the universe of generalization, or the sample of observations is large enough to control random error. Kane (2006) pointed out two effective ways to minimize large random error with a facet. First, a larger sample size could minimize random error and sampling variability. Second, the definition of "attribute" can be modified, and the effect of narrowing that definition is related to the test usefulness. In sum, standard errors of measurement from reliability studies will provide precision of estimates of the universe scores.

The assumption of extrapolation inference is that the universe score is related to the target score and relatively free of systematic and random error. Even though an actual score does not change, the interpretation of the score can be changed by reviewing empirical evidence of relationship between scores and target domain.

The last inference for a trait interpretation is the implication derived from target score to verbal description. The implication involves a detailed interpretation of suggestions or claims associated with the trait. By claiming an implication argument, factors affecting inconsistency of the trait could be revealed and the plausibility of the trait interpretation could be increased.

Theoretical Background on Listening Comprehension

Process of listening comprehension from a psycholinguistic perspective. In this section, the psycholinguistic view of the listening process is discussed because this perspective

crucially informs the collection of interpretive evidence in the test validation process. The psycholinguistic approach addresses research questions about how learners process meaning and what types of processes they prefer to use. In response to such questions, three processing approaches have been identified and developed for listening comprehension: top-down, bottom-up, and parallel. Listening comprehension was first believed to be a one-way, bottom-up process. However, a recent theory suggests that listening comprehension requires a more complex combination of top-down and bottom-up processing.

Top-down processing refers to a theory that learners derive meaning from and interpret a message using schemata. In other words, learners use compensatory strategies when they process language input. According to the top-down approach, learners/listeners are required to predict the meaning, use contextual clues and combine them with background knowledge. Listeners prefer to use the top-down process when they build a conceptual framework of comprehension using context and prior knowledge (Vandergrift, 2007). According to the top-down approach, listeners' successful speech perception depends on their active reconstruction of acoustic input. Most researchers argue that learners need to use the top-down process in order to enhance their comprehension ability. However, if learners rely too heavily on the top-down process, they might not catch specific vocabulary or meaning.

Bottom-up processes develop when listeners derive the meaning of a text based on its individual building blocks, sounds, words, and grammars. Listeners apply the bottom-up processing when they use linguistic knowledge to understand the meaning of a text. In the field of listening comprehension research, it is known that listeners with limited L2 competence heavily rely on bottom-up processing skills. If listeners depend on the bottom-up process, they are likely to encounter perceptual problems when they hear hesitations or pauses. Several

researchers have investigated the level of dependence on bottom-up processes among L2 listeners. For example, Conrad (1983) showed that non-native speakers were more likely to pay attention to syntactic information than native speakers. However, the findings of Conrad's study should be interpreted with caution because he/she did not consider L2 subjects' level of vocabulary knowledge. Since there is some evidence that supports a positive relationship between learners' listening comprehension ability and their vocabulary knowledge, the results of Conrad's study are limited. In contrast to the findings of Conrad's study, Vandergrift (2007) has suggested that less skilled listeners rely more heavily upon top-down processes in order to compensate for their limited ability in perception. Listeners' reliance on the top-down process reflects their lack of confidence in their ability to process the sounds of second language accurately. In other words, listeners prefer to match unfamiliar words with known words using top-down processing.

Parallel processing suggests that both top-down and bottom-up processes interact in a parallel form when listeners attempt to comprehend the meaning of a text. This processing depends on listeners' level of language capacity because they need to know how to use these processes or how to exchange one process with another. According to Rost (2002), understanding spoken language is an inferential process, and linguistic and world knowledge are interacting in parallel as listeners create a mental representation of what they hear. His view of listening comprehension is close to the parallel approach because he argues that learners use different types of processing simultaneously. O'Malley *et al.* (1989) found that effective listeners listen to larger chunks and are able to shift their attention to individual words when there is a breakdown in comprehension. They claim that this interaction is particularly evident when communication breaks down. Their study is significant in that it initiated the notion that

using several processing skills depends on the learners' motivation to comprehend second language listening.

Factors influencing listening comprehension. As researchers have gained a better understanding of the unique characteristics of listening skills, they have begun to pay more attention to the factors that affect listening comprehension. Several characteristics have been identified that affect or improve a learner's listening comprehension: 1) text characteristics, 2) task characteristics and 3) listener characteristics (Rubin, 1994). It is important to understand how these factors impact listening comprehension because they are closely related to the processing of aural input. Furthermore, they have been important in the field of language testing and second language research because they can either aid or inhibit listening comprehension.

Text characteristics. Text characteristics include such variables as acoustic (temporal variables), stress and rhythm patterns, redundancy (repetition), structural complexity, and text type. Many studies have reported the effect of these variables on listening comprehension. First, prior studies have examined the relationship among speech rate, pause and hesitation phenomena, and listening comprehension. Buck (2001) reported that the average speech rate of a native English speaker is 165 to 180 words per minute (w.p.m.). The threshold for comprehension loss is set between 250-275 w.p.m., beyond which comprehension decreases in accordance with a function of mental aptitude and difficulty level. However, Tauroza and Allison's study (1990) showed that the normal speed of British English speakers varied depending on the types of speech. Table 2 shows that the speed of typical interactive speech, conversations and interviews, is relatively faster than that of linear speech such as monologue and lecture. From the table, it is also noted that the speed of lectures to non-native speakers is much slower, which suggests that lectures are presented more slowly to help listeners'

comprehension. While Tauroza and Allison showed different speed rates of various types of speech, there is limited evidence on how speech rates affect listeners' comprehension. Several studies have suggested that different rates of text may lead to effective comprehension, but it appears that speech rates alone cannot explain the level of comprehension (Tauroza & Allison, 1990, Buck, 2001). Additional variables such as the types of texts used and the amount of knowledge required should be considered when an investigation of L2 listening comprehension is conducted.

Table 2

Average Speech Rates for English (Tauroza and Allison, 1990)

Text Type	Words/minute	Syllables/minute	Syllables/second	Syllables/word
Radio/ Monologue	160	250	4.17	1.6
Conversations	210	260	4.33	1.3
Interviews	190	250	4.17	1.3
Lectures to NNS	140	190	3.17	1.4

Second, it has been shown that syntactic, morphological modification affects learners' comprehension. Previous research reports that morphological and syntactic modifications in the input could ease comprehension. Rubin (1994) reviewed representative studies on the effects of morphological modification. According to him, one study showed that when a text contained a repeated noun, it helped learners' recognition and recall. He also reported the results of another study in which high-intermediate learners benefited from speech modifications while low-intermediate learners did not. According to Rubin, irrelevant redundancies in the text could give learners more cognitive load and input to process, especially for those with low capacity.

In addition to speech rates and linguistic modification, structural complexity and spoken text types have been identified as factors affecting listening comprehension. Unlike written texts, spoken texts tend to be less syntactically complex and more redundant. They also include more hesitations and pauses than written texts, which provide listeners more processing time and make them easier to understand. According to Brown (1985), expository texts are more difficult than narrative texts due to their complex structure. He also argues that narrative texts in which events are described in a disrupted order are more difficult to understand than narrative texts that present events in a chronological order. However, there is limited evidence to support his argument that the types of texts affect learners' comprehension ability. More investigation is needed to identify what types of texts could better assist learners' comprehension and how the length of texts affects their understanding.

Task characteristics. Task characteristics are less directly related to listening comprehension than the other two characteristics. However, few studies have been conducted on the effects of different task types on listening. Shohamy and Inbar (1991) investigated how question types affected L2 listening tasks and found that participants performed better on questions that provided local cues in the text than on those presenting global cues. Participants who were able to answer global questions correctly answered local questions. However, test takers who answered local questions were not always able to correctly respond to global questions. The results of the study demonstrated that learners' performance was affected by their level of language proficiency. It appeared that the effects of task types on performance were greater for learners with a low proficiency level due to their lower language ability.

In his study, Robinson (2001) distinguished the complexity of cognitively defined tasks, learners' perceptions of the difficulty of the given tasks, and the interactive conditions under

which tasks are performed. Even though his study was not directly targeted for listening comprehension, it is significant in that he examined listeners' perspectives on the relationships between speakers and listeners. His study suggested that cognitive complexity of tasks has a significant influence on learners' outcome. Therefore, it is important to determine the effects of complexity on listeners' comprehension.

Listener characteristics. Listener characteristics can impact listening comprehension in positive and negative ways. Researchers include language proficiency level, memory, background knowledge, and aptitude in this category (Rubin, 1994). In the field of language learning and teaching, these considerations are very important when we consider improving learners' listening comprehension. If learners have strong working memory and in-depth background knowledge, they probably comprehend listening texts more easily than those without strong working memory or background knowledge. Hence, some teachers emphasize listener characteristics when they teach listening. Nonetheless, when creating listening tests, language testers try not to provide items which require strong working memory because those items would leave doubt about whether they are measuring listeners' listening comprehension or their working memory span.

Even if we consider some characteristics as confounding variables, listener characteristics have a very strong impact on listening comprehension. First, language proficiency is a major variable because people cognitively process based on what they have learned and perceived. Researchers suggest that cognitive processing varies depending on learners' knowledge of the language. But it is not clear what role linguistic competence, sociolinguistic competence and cultural background knowledge play at different proficiency levels (Rubin, 1994). Rubin

highlighted this as a very intriguing issue for language researchers and testers because several variables inter-correlate with one another.

The second learner variable is memory and its complex relationship with listening comprehension. Studies on memory are ongoing, and we still lack tangible evidence that shows how memory affects listening comprehension. Especially in second language acquisition, we do not clearly understand how language is processed during listening. In order to acquire evidence, many researchers have used an introspective data collection method (Vandergrift, 2007; Goh, 2000). Call (1985) found that short-term memory was important in listening comprehension, and different types of input were correlated with listening comprehension as well. Even though Call highlighted the strong relationship between short-term memory and listening comprehension, she did not eliminate the possible influence of variables such as language proficiency level and input selection. Rubin (1994) argued that studies related to short-term memory should be reconceptualized from a new perspective. Short-term memory tends to be measured using recall protocol or recognition, but studies should be more focused on levels of activation.

The third learner variable, background knowledge is also a crucial factor in listening comprehension, as it is especially closely related to cognitive processing. According to process theory (Flowerdew & Miller, 2005), when learners listen to messages with unfamiliar noises or words, they first try to find an overall schema. Regardless of proficiency levels, listeners always try to associate background knowledge with the input. Brown and Yule (1983) describe this prior knowledge as schemata, and it functions by leading listeners to expect or predict elements of the discourse. They discuss two principles to relate the new information to the previous knowledge: the principle of analogy and the principle of minimal change. The principle of analogy means that listeners expect things to be the same as they were. The principle of minimal

change means listeners expect things to be as similar as possible to how they were. Listeners form inferences based on these principles and use them to interpret spoken language.

Many studies have been conducted on background knowledge and listening comprehension by using a recall protocol. Background knowledge has been operationalized in various ways: cultural knowledge, religious knowledge, technical knowledge, topic familiarity, and contextual visuals. Schmidt-Rinehart (1994) conducted a study whose main purpose was to investigate the effects of topic familiarity on second language listening comprehension. Findings support the hypothesis that background knowledge contributes significantly to students' comprehension. However, this study couldn't explain whether proficient listeners use schemata-based processing to the same degree as less-proficient listeners. Chang and Read (2006) investigated the effects of four types of listening support by providing a preview of the test questions, repetition of the input, background knowledge about the topic, and vocabulary instruction. This study showed that the most effective type of support overall was to provide information about the topic by repeating the input. Most learners also gave positive feedback on providing topical knowledge, but some learners had a hard time concentrating on the test input due to previously provided topic knowledge input. Overall the study showed that learners' proficiency levels interact significantly with the types of listening support provided.

Several studies show evidence that certain characteristics provide useful resources for comprehension. Awareness of the strength of those characteristics would be very helpful for teachers, but the question still remains for language testers whether we should factor these variables into test development or not.

Learners' listening strategies. Attention to listening strategies has risen as researchers have begun paying attention to learning strategies and teaching methods that promote strategies.

Several researchers use O'Mally & Chamot's (1990) classification system: metacognitive strategies (strategies concerned with planning, regulating, and managing learning), cognitive strategies (strategies that make use of prior knowledge to facilitate comprehension), and social and affective strategies (strategies such as questioning, positive self-talk, and anxiety management). Underlying this research on listening strategies is the belief that strategy instruction can improve learners' listening comprehension. As a result, strategy instruction has become the core of many listening instructional programs (Mendelsohn, 1998).

Metacognitive strategies involve planning, monitoring, and evaluating comprehension to help learners to view the big picture process. Comprehension monitoring (Vandergrift, 1996) is a superordinate framework that directs other metacognitive strategies such as selective attention, inferring, and elaboration. First, O'Mally, Chamot, and Kupper (1989) explained that monitoring is a key process that distinguishes good learners from poor learners. Monitoring consists of maintaining awareness of the task demands and information content. Attention has two different types: selective attention (focusing on specific information), and direct attention (focusing more generally on the task demands). Teachers often emphasize the use of selective attention to help learners succeed in second language listening comprehension. Last, elaboration relates the new information to old information that is stored in memory or interconnecting portions of the new text. Students use elaboration by evoking prior knowledge and forming inferences.

The common approach to studying strategy use is to determine what strategies learners use, how frequently they use those strategies, and what purpose is served by the strategies. Researchers have investigated the relationship between frequency of strategy use and proficiency levels of learners to determine the effectiveness of their listening comprehension. O'Malley *et*

al. (1989) conducted a study to investigate the relationship between both effective and ineffective listeners and their strategy use. This study was mainly focused on the mental processes and strategies used by learners who were designated as effective or ineffective listeners by their teachers. Since this categorization was made by the participants' teachers, the process was subjective and could be a limitation of the study. Results showed that effective listeners use more self-monitoring, elaboration, and inferencing while they are listening. The authors claim that analyses of strategic processing offer three important research conclusions. First, the frequency and type of strategy use determine whether learners are effective or ineffective. Second, they believe that strategic modes of processing could be taught by teachers. And third, the use of strategic processing can enhance learning. Demand has risen to define the terms "effective listener" and "proficient listener." Failure to provide such definitions is a potential weakness of most studies on listening comprehension.

Vandergrift (1996) also investigated the relationship between learners' proficiency level and use of metacognitive strategies. He found that higher proficiency learners more frequently use metacognitive strategies. In sum, this evidence strengthens the importance of strategy use and strategy-based instruction in listening comprehension.

Identifying different strategy patterns at different proficiency levels is very useful. Young (1997) investigated the possible existence of a sequence of use of listening comprehension strategies by ESL learners. Her results were similar to those of other researchers. First, listeners who used more strategies frequently applied inferring or elaboration. When their background knowledge was activated, they also used summarization to reinforce their interpretation of the text. In addition, effective listeners used metacognitive strategies such as self-monitoring or self-evaluation to control their comprehension and to evaluate their strategy

use. Second, listeners seemed to use the same sequence of strategy choices. Listeners who used elaboration and inferring to activate their schemata also used summarization to consolidate their understanding. These results reveal that listeners use cognitive strategies to help activate their background knowledge, and they use metacognitive strategies for directing and monitoring their comprehension.

Unlike metacognitive strategies, cognitive strategies, involve solving learning problems by considering how to store and retrieve information. This involves active manipulation of the learning task as well as rehearsal, organization, and elaboration. Learners use rehearsal strategy by repeating the names of objects or items that have been heard, or practicing a longer language sequence. Organization helps learners to group information that leads to enhanced comprehension and retention. O'Malley *et al*. claimed that elaboration is a particularly significant strategy because of the demonstrated benefits to comprehension. Unlike Vandergrift's idea of the superordinate category, O'Malley *et al*. suggested that elaborative strategies are considered as a superordinate category for other strategies such as inferring, transfer, deduction, imagery and summarization.

A third strategy category, social and affective strategies, has scarcely been studied by second language researchers. However O'Malley *et al*. claimed that this strategy category plays an important role in second language instructional systems designed to entail cooperative learning, questioning for clarification, and affective control over the learning experience. Young showed evidence that social strategies appear to be optional in the processing of information, and listeners use them in interactions with other speakers.

Most researchers have investigated the effectiveness and usefulness of metacognitive and cognitive strategies. They have agreed that elaboration and inferring techniques are important

strategies for successful comprehension. Most advanced learners are aware how to use metacognitive and cognitive strategies effectively. Still, affective and social strategies need more attention from researchers and teachers.

Issues in Developing Second Language Listening Tests

Construction of listening comprehension tests. In language testing, test developers design various tasks to engage certain aspects of skill and knowledge, but they have no detailed guidelines to show what skills, abilities and knowledge are reflected in the listening tasks they create. Given the complexity of listening processes, language test developers struggle to figure out how well test-takers engage with their tasks and how easily test-takers can process listening. It is difficult to define the operational construct of listening and to assess higher level of listening skills due to practicality and authenticity of the test.

One challenge in second language listening assessment is to define a measurable and acceptable construct for the test. Defining a construct is the fundamental and is the initial step in test development. According to Buck (2001), there are two ways of defining a construct: a) define a construct at the theoretical or conceptual level, and b) define a construct by operationalizing it as test tasks. Theoretical knowledge of the target language and knowledge of the target-language situation are necessary to define the test construct. Buck (2001) also pointed out that operationalization of the construct could be inadequate in some situations because specific content may not be appropriate for all testing contexts. In listening test development, the construct needs to be defined both theoretically and operationally so that the test measures examinees' implicit listening ability. Implicit definitions are needed to reflect the intended definition of the trait because other traits could be measured unintentionally. Even though the

construct definition does not appear in the actual test items, it should underlie and inform each one.

Test developers also need to consider practical constraints that have a large impact on construct definition like budget and administrative logistics (Buck, 2001). First, they must decide whether they will use collaborative or non-collaborative listening tasks. Because of practical constraints like saving money and providing input in a convenient way, language test developers generally prefer to use non-collaborative tasks in their tests. This decision is directly related to a trade-off between authenticity and practicality. Non-collaborative listening items have listeners decode auditory meaning in a fixed manner and they lack authenticity (since listeners always listen as a third person, they get no opportunity to negotiate meaning), but they are much easier and less expensive to administer. Brindley (1998) argued against this trend and said that test developers should employ collaborative listening tests due to the importance of interactive assessment.

It is difficult to control a listening task so that it leads to one possible interpretation in the test. In actual communication, when a speaker conveys meaning to a listener, the meaning is likely to be interpreted in several ways. Buck (2001) explained that effective daily communication does not usually need precise understanding because listeners often manage well with a rough approximation of speakers' intended meaning. However this situation makes it difficult for test developers to create spoken texts, because normal variation in interpretation could cause a problem in test design. To minimize this possible problem, explicit guidelines on spoken texts should be written in the test specification.

In the test specification, not only do we need to explicitly indicate a construct definition, but also we need to create rational guidelines for task response. Often language testers choose to

use multiple-choice items to minimize the impact of other ability use, but Brindley (1998) claimed that the validity of a listening test is threatened by tasks and texts that require the use of language skills other than listening. Test-takers may have to read written instructions or questions (stimuli) and provide oral or written responses. Brindley also emphasized that if test developers have not carefully intended to design a listening test, the test would end up with assessing other skills. Doing so may trigger "construct-irrelevant variance" which was initially proposed by Messick (1989) (see Brindley, 1989; Buck, 2001). Test developers should be aware of construct-irrelevant variance and be careful not to discriminate against learners' proficiency levels when we adopt possibly confounding formats that engage other language abilities.

Some researchers have stated that it is impossible to construct a pure listening test that is uncontaminated by other skills (Buck, 2001). Call (1985) discussed in her paper the confounding effects of individual factors including memory and topic familiarity. It is evident that orally presented items could be affected by motivation, attention and memory.

To minimize the effect of memory and maximize the measurement of listening capacity, test developers have introduced open-ended questions in the listening test. Open-ended questions require test-takers to construct a response with less reading or less memorization, but it requires writing ability. Buck (2001) suggested that the use of the first language to construct a response may be useful to measure L2 listening comprehension because it would eliminate the L2 writing effect. However, it is not feasible to use first language for responses in a heterogeneous group of different first languages.

Listeners use both higher level and lower levels of listening processes when they use interactive, parallel processing. It is difficult to distinguish different levels of processing or to attribute test responses to only one skill (Brindley, 1989). Researchers agree on the notion of

indistinguishable differences between inferring and language processing. They have emphasized that test developers need to consider this variable as an inference for item construction. Test items should be designed to constrain the range of possible responses because text interpretation is subjective to individual variation.

For instance, Henning (1991) detected three problems of the TOEFL listening comprehension component. First, the TOEFL listening comprehension test relies more on short-term memory load than on comprehension. Second, the use of a reading response format invalidates the test as a measure of listening comprehension. Finally, too many items require the recall of minute details rather than requiring higher-level processing strategies. Based on these problems, he compared TOEFL listening comprehension item quality under a variety of conditions of stimulus repetition, response-option reading length, and cognitive processing hierarchy. No evidence was found that memory was associated with either stimulus passage length or nonrepetition of stimulus passage, which will negatively affect item quality or task validity. Even though passage length was confounded with the number of items per passage, the test with two-sentence passages tended to be more reliable than one-sentence passages. In addition, the longest passages tended to be perceived as more difficult than the shorter ones. He found that item response choice of length was significantly related to item difficulty. Furthermore, the test contained more lower-level processing items than higher-level processing items. The Rasch model fit showed lower-order items (involving understanding of utterances at the literal level) has greater response validity than higher-order items involving inference and critical evaluation. In sum, findings showed that reduced length of reading response choices could make better item quality and comprehension hierarchy. Above all, Henning's study shows that it is difficult to assess higher-level listening skills via a multiple-choice format because it

provides fixed response choices to interpretive inference questions. A challenge for test developers is to find a tangible relationship between test-takers' inferences and test developers' intentions of higher-level listening items.

Test usefulness. Many researchers and test developers have emphasized the importance of reflecting authenticity in test tasks and items in whatever language domain is being tested. However, Brindley (1989) pointed out that the use of authentic samples of naturally occurring speech could be problematic because of poor sound quality, lack of contextualization or heavy processing load.

Bachman (1991) categorized two types of authenticity used in language testing: situational authenticity and interactional authenticity. Situational authentic tasks share characteristics of real-world target-language use tasks. Bachman defined situational authenticity as the perceived relevance of the test method to the features of a specific target language use. If the target test-takers are from the field of business, for example, we can include technical terms or business topics to increase situational authenticity. Unlike real-life approaches, situational authenticity has distinctive features that characterize the target language use tasks. The other type of authenticity that Bachman proposed was interactional authenticity that resides in the interaction between the test-takers and test task. It requires considering both test tasks' characteristics and test-takers' language ability. Interactional authentic tasks engage test-takers' metacognitive, cognitive strategies, topical knowledge and affective schemata, the same abilities as the target-language use tasks. One way of assessing interactional authenticity is to observe test-takers and to request self-report on the strategies they used for completing the test tasks.

In addition to issues of effectiveness and adequateness of authenticity, the difficulty level of task is another issue related to authenticity. Difficulty levels of tasks are often related to

authenticity because authentic materials can be difficult and may discourage low-proficiency learners' motivation to solve the tasks. Ur (1984) stated that certain difficulties may surface to cause frustration and demoralization when teachers use non-scripted, authentic language tapes because it is particularly difficult for beginner level students to identify different voices and to cope with frequent overlaps in segments of authentic language. To minimize possible trouble and to maximize authenticity, test developers use semi-scripted texts as suitable listening passages (Buck, 2001). Speakers could add more hesitations to make it authentic and keep the natural speed to make every listening passage equivalent.

To increase authenticity, test developers could also control the input by designing items that are passage-dependent. In a listening comprehension test, successful completion of tasks depends on comprehension of audio and text input. Buck (2001) addressed two reasons why test tasks may lack passage dependency. First, test tasks themselves often provide information that could be a clue to the content of the passage. Second, test-takers might be able to use their background knowledge or common sense to respond to the questions. This can happen easily in a multiple-choice format, and test-takers' ability is overestimated because they find ways to compensate for a lack of comprehension. Test developers often face this issue as a challenge because they have to create test tasks that are not passage dependent.

Test developers always carefully think about the topic, types of text, types of presentations, and numbers of speakers in the spoken texts. Additional dilemmas are how to construct realistic and authentic spoken texts. In an authentic situation, speakers make frequent pauses, hesitations, and listeners also hear background noises. Furthermore, speakers may change their speech rate depending on situations or emotional changes. Speakers recording voice mail messages tend to speak faster than usual because they are speaking on the phone by

themselves. In a phone conversation, the speech rate could vary depending on the situation, so providing a consistent speech rate in test input may not fully reflect the real phone conversation. In this case, although the test developer tries to design the input to be authentic, test operationalization could be problematic.

In real life, listeners also largely depend on visual information such as facial expression, posture and movement. Messages result from these and other non-verbal or visual cues. When a listener engages in listening, the vocal message goes through the short-term memory first and the listener focuses on the auditory and/or visual stimulus and decodes the meaning. To approximate the natural process, some test developers have used context or content visuals as listening comprehension stimuli. However, this issue is still controversial in several ways: whether those content visuals actually help listeners' listening comprehension, and whether these are authentic enough. Ginther (2002) investigated the relative effect on comprehension of two types of visuals for the mini-talk in the TOEFL. Content visuals such as pictures related to actual content slightly enhanced listeners' comprehension, but context visuals that contained scenes for the upcoming verbal exchanges were not significantly helpful for learners. Buck (2001) discussed the positive and negative side of using visual along with audio input. He emphasized that the content of a video/visual aid should be the same as the audio input, because otherwise test-takers will be so confused by the mismatch as to spend more time processing meaning. Furthermore, if we include video/visual input, test-takers have to watch/study it, listen to the audio script and read questions and responses simultaneously. Therefore we might end up with adding one more input that is not very helpful for test-takers' comprehension.

Due to the uncertain effectiveness of content visuals, many test developers still use linear listening tasks and recorded spoken text. However several studies have been introduced in the

field of communication that reveals the effectiveness of nonlinear listening. Mesbah (2006) examined the effect on cognitive processes of the mode of listening to radio news. This study revealed several challenges and considerations of using nonlinear listening for testing. The stimulus was a real newscast that was manipulated into four versions: traditional radio news cast, online newscast played with one click, linear interactive netcast with a click for news item, and a support activity condition in which additional links for details were added to each link. As technology has evolved, people's method of getting news has changed from traditional print and broadcast sources to reading and listening to online news. This study tried to examine the effectiveness of online interactivity by using an authentic way of listening to the netcast.

A major quality that differentiates new media from traditional media is its interactivity, and this is what test developers aim to reflect in their test items. Mesbah looked at two factors, the potential cognitive efficacy of nonlinear and interactive audio presentations. The results showed that listening to online news improves both recall and comprehension. Participants were able to retrieve more details and to understand development of actions more easily when they had more control over the pace of information. They also showed less engagement with traditional radio listening than with Internet listening. Nonlinear listening to news provides easy elaboration of news content. However, adding multiple links and providing background knowledge did not result in a better memory performance. Some participants expressed frustration with the various links and information because of their heavy cognitive load. The simultaneous multiplicity of sources creates a more complicated cognitive situation in which lots of variables interact to facilitate or impede the processing of information.

Test developers face the need to make authentic tasks to measure examinees' listening capacity. They have to consider not only the construct of the test, but also the authenticity and

interactiveness of language use when they provide audio input. However, it was difficult to find a study that looks at the effect of impediments and authenticity of test items using various listening features. This study will attempt to meet these challenges.

Chapter 3

Methodology

Chapter Overview

The current study employed three major research phrases: test development, administration, and materials development. This research was planned in 2009, and the test development and data analyses were conducted in 2010. Detailed descriptions of materials that we used for the current study are explained. In addition, arguments for research questions and levels of inferences are addressed in this chapter.

Design of the Study

This study was designed based on a pragmatic perspective of mixed method design. The purpose of this study was to create an authentic listening comprehension test for ESL learners and to see the complex listening processes at play the testing situation. First, a web search was conducted to find different types of listening tests and tasks. A needs analysis was then conducted to analyze listening tasks on the web and to collect validity evidence for the listening test. These tasks were analyzed to determine the factors of authenticity that could impede listeners' comprehension. Based on the findings of this analysis, test specification, tasks, and a cognitive awareness questionnaire were developed.

A pilot study was conducted to verify degrees of impediments, item quality, and specificity of questionnaire items. Comments from potential test takers, an audio recording engineer and language testers were used to refine the elements of questionnaire and sample listening tasks. In addition, test takers' point of view contributed to strengthen validity evidence of the test. Based on input from test takers, the test specifications were revised periodically. When the test specifications were ready to operationalize, the main study was conducted to

assess learners' second language listening proficiency and to complete the collection of validity evidence. After the data collection, several data analyses were conducted to answer different levels of validity questions.

Table 3

Overview of the Research Activities

Time (year, month)	Activity	Description
2009.4	Needs analysis	Analyzed possible breakdown listening on the web
2009.11	Test development	Chose appropriate texts, edited for listening prompts, and created listening comprehension items
2009.12	Listening prompts recordings	Recorded listening prompts, edited, manipulated, and included impediments
2009.12	Pilot study	Selected appropriate impediments, revised listening comprehension items and cognitive questionnaire
2010.1-2010.2	Data collection	Administered four listening tests to 96 test takers
2010.3	Data analysis	Analyzed data at test level, item level, and cognitive awareness

Data Collection

Participants. Sixty-eight international graduate students and twenty-eight international undergraduate students enrolled at the University of Illinois at Urbana-Champaign participated in this study. The participants were recruited via email and advertisement. Volunteers individually visited the site and took the test. Each test was administered by a following order: regular, British accent, speech rate, and noise tests. Among these participants, fifty-five were males and forty-one were females (see Figure 1). As Figure 2 shows, 71% of participants were graduate students

and 29% were undergraduate students. Each group had participants who lived in the United States for less than 1 year, between 1 and 5 years, and more than 5 years. Twenty-nine participants have lived in United States for less than 1 year, 29 participants for between 1 and 5 years, and 38 participants for more than 5 years (see Figure 3). A majority of participants (n = 90) were from Asia (Korea, China, Taiwan, Malaysia, India, Japan, and Uzbekistan) and others (n=6) were from North America and Europe (see Figure 4).

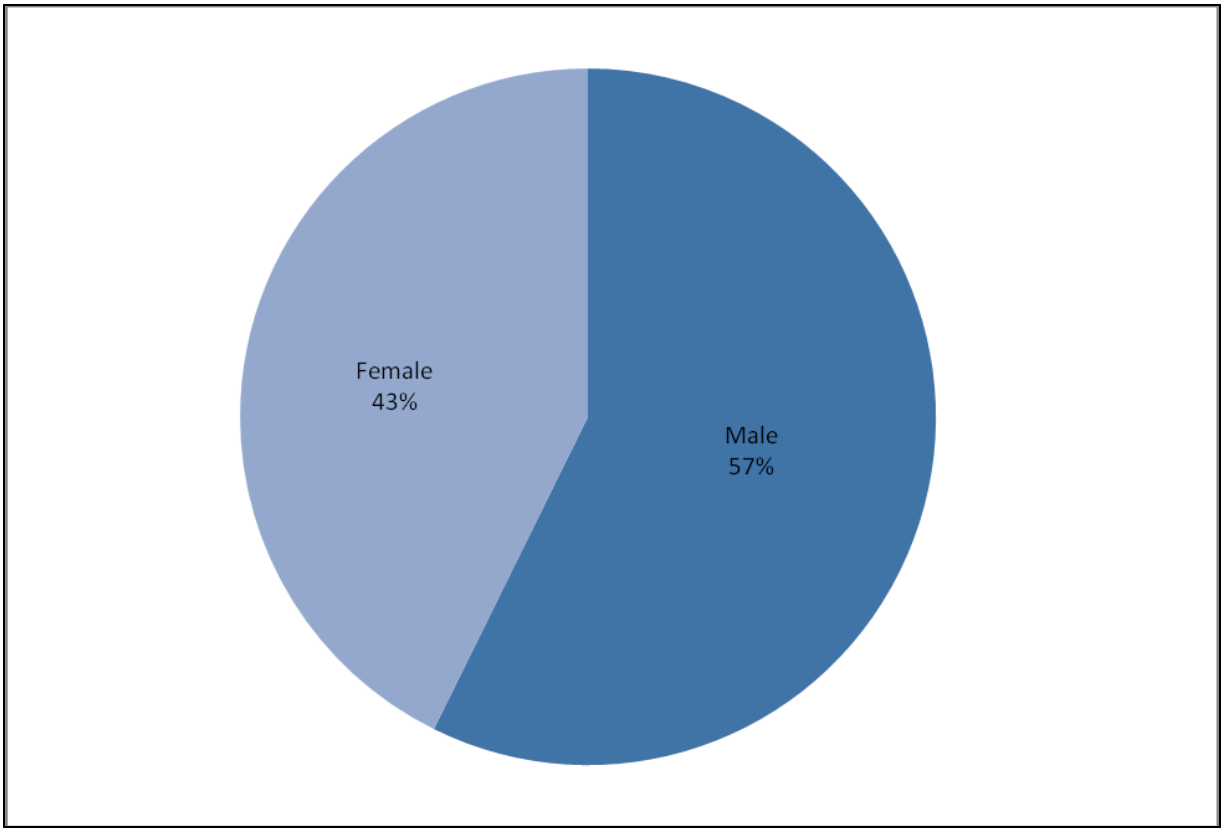

Figure 1. Gender ratio of participants

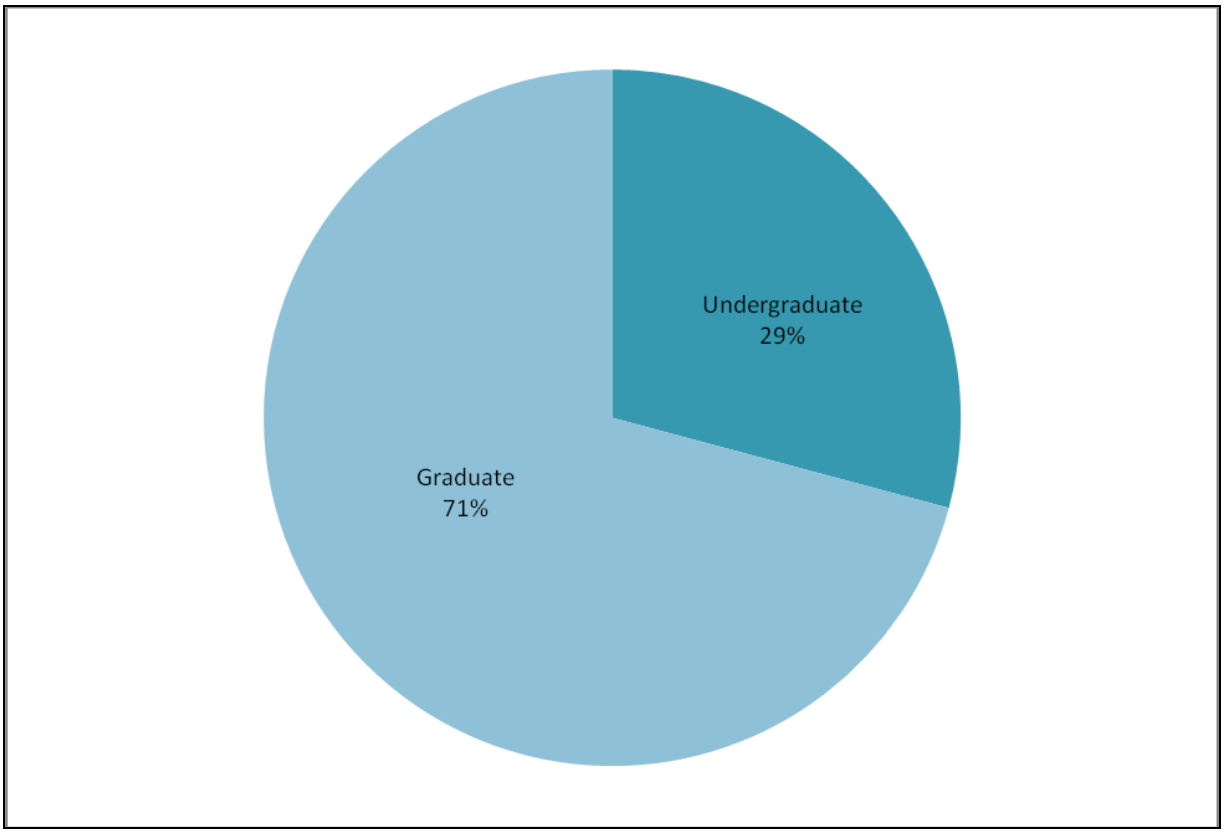

Figure 2. Participants' academic status

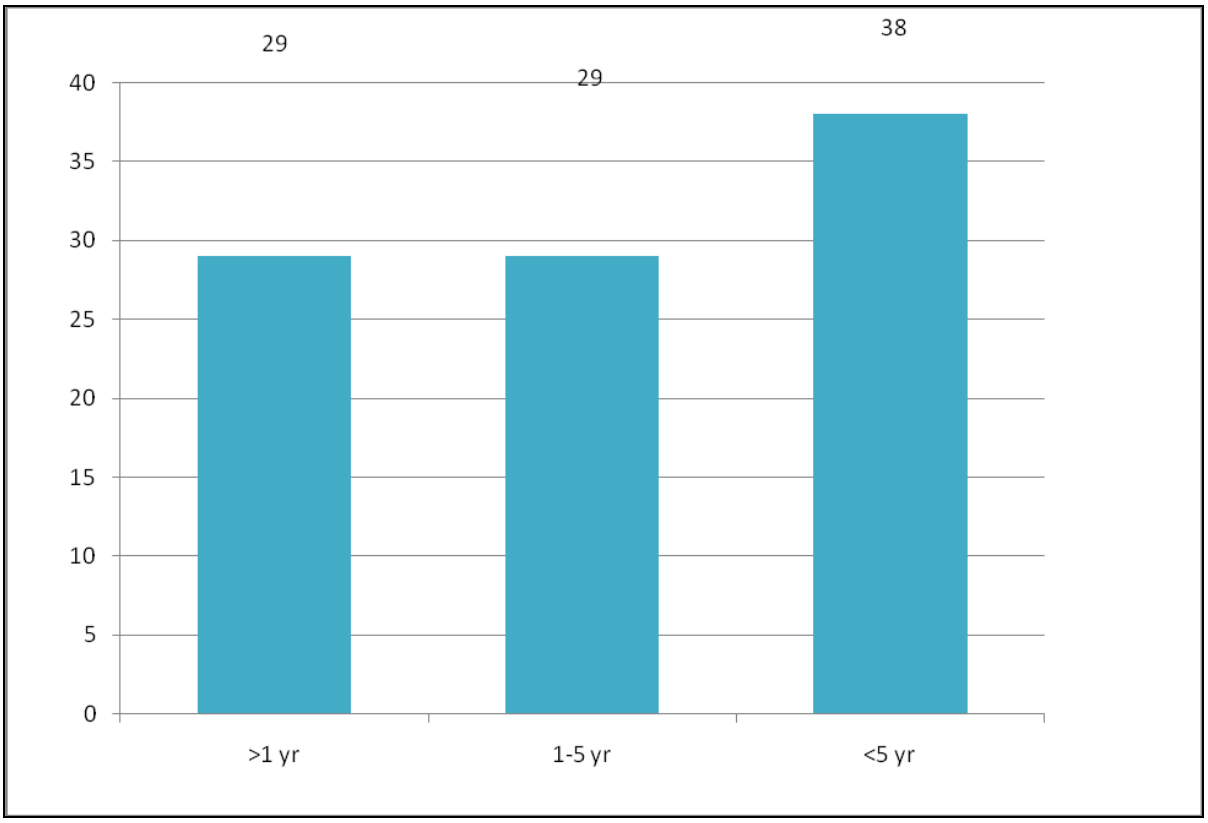

Figure 3. Participants' years of residence in US

Table 4

Participants' Length of Residency in US by Groups

Group	>1 year	1-5 years	<5 years
Regular	34% (10)	17% (5)	24% (9)
British	28% (8)	17% (5)	29% (11)
Speech Rate	14% (4)	34% (10)	26% (10)
Noise	24% (7)	31% (9)	21% (8)
Total	100% (29)	100% (29)	100% (38)

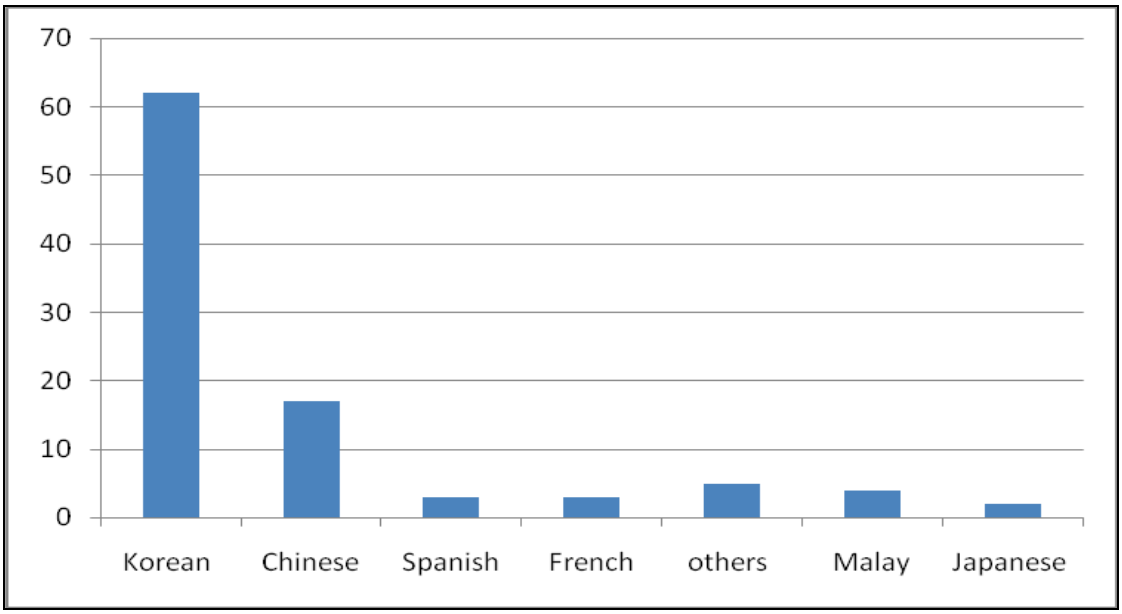

Figure 4. Participants' first language

The listening test. One listening comprehension test was developed, but a total of four test forms was administered to participants: three test forms with different impediments (British accent, speech rate, and background noise) and one test form with no impediment.

The purpose of including impediments was to investigate how listeners' strategies differ under different authentic conditions when answering comprehension questions. To investigate listening breakdown, three impediments were chosen to be included in the listening prompts. First, British accent was chosen. Although British English is another variation that is related to American English, most international students in this study have trained and studied American English might have difficulty understanding a British accent. Second, a fast speech rate was used as an impediment. When a speaker speaks slowly, a listener has a relatively long time to process the information. On the other hand, if a speaker speaks faster, the listener will have less time to comprehend the spoken information. For this reason, I selected a fast speech rate as an impediment in the study. Initially, the actor tried to speak fast when he read the scripts, but it was very difficult to maintain a certain speech rate. Therefore, to realize a faster yet consistent

speech rate, I manipulated the rate using an audio editing program called Audacity. The last impediment included in the study is background noise. Brown, white, and pink noises were tested to verify their feasibility. Both brown and white noises were inappropriate to use for the prompts, because the static noise of brown and white noise was too strong so it was difficult to comprehend the speaker's voice. These samples were not good examples for authentic listening tasks, so I did not select brown or white noise for this study. Therefore, 15 percent of pink noise was added to the recorded prompts.

Beside the purpose of measuring effects of breakdown, this test also attempted to measure test takers' cognitive awareness of listening and their strategic listening skills for comprehending explicit and implicit spoken information. To measure these psycholinguistic and linguistic features, two topics were selected: dinosaurs and climate change. These topics could be unfamiliar to listeners, however the comprehension questions were designed without bias and considered the possibility of prior knowledge use. Two types of talks were used: presentation and lecture. Each topic was presented in presentation and lecture version. A total of four talks were provided, and five comprehension questions were provided under each talk. Each comprehension question appeared in a multiple-choice format. Though listening prompts were manipulated with impediments, the order of tasks, topics, and listening comprehension items were counter-balanced.

Recording. A trained actor and an audio engineer participated in recording the listening samples for the test. The native male actor was an undergraduate student majoring in theater, and he had acting and voice over experience. He also had experience in voice recording for language test development and received coaching from the audio engineer and me. The audio engineer was a PhD student majoring in computational linguistics and worked as a video/audio

engineer at the UIUC ATLAS (Applied Technologies for Learning in the Arts and Sciences).

The actor recorded the lectures in a sound-attenuating booth using a unidirectional condenser microphone mounted on a boom and a digital recorder. After recording at a sample rate of 44.1 kHz (16 bit), MP3 files of the recordings were created. The actor practiced versions of the listening prompts prior to recordings, and he focused on any words or sentences produced with errors or mispronunciation. After the recordings were made, the audio engineer edited them and changed the speech rates. The researcher added the noise to the recordings. The audio engineer and I reviewed the recordings several times to verify the clarity and levels of impediments.

Impediments. As briefly mentioned before, three impediments were included in the test. To minimize confounding effect of speakers, one speaker recorded all listening prompts. He recorded both British and American accents**.** On the American accent listening prompt, the researcher manipulated the speech rate and noise.

Speech rates were tested in several versions by reducing the spectrum of recording. For example, 80% and 75% reduced speech samples were not appropriate for operational prompts because the speaker's voice sounded artificial and the content of the speech was not comprehensible. Thus, a 90% reduced speech sample was selected for the operational listening prompt.

For noise speech sample, three types of noise could be added to the recording: white, brown, and pink noises. Three types of noise were implemented at the testing stage, and the pink was selected as an operational noise because this noise was caused relatively less fatigue than other two noises. The noise speech sample was tested by adding 5%, 15%, and 30% of pink noise in the sample. Pink noise with 5% and 30% turned out as bad speech samples because one

was not distinctive and the other was too noisy to comprehend the content of the speech. Hence, the speech sample with 15% of pink noise was selected.

Cognitive awareness questionnaire. A questionnaire was designed to collect test takers' (1) background information, (2) self-reported TOEFL scores, (3) metacognitive and cognitive awareness, (4) listening strategy uses, and (5) test-taking strategy uses. A total of nineteen questions were provided to understand test takers' cognitive awareness and test-taking strategy. The response format was a scale of 1 to 3 (disagree, agree, and I don't remember). Before the test administration, participants were informed that the questionnaire would be provided at the end of the test. The estimated time for completing the questionnaire was about 15 minutes.

Frequency distributions were calculated based on the responses from test groups. To examine the association between test groups and test takers' responses, a Chi-Square test for independence was conducted using test takers' responses (agree/disagree only) from four test groups (the regular, British accent, speech rate, and noise groups).

Reflection session. After each participant filled out the questionnaire, the researcher asked a few questions to participants who agreed to participate reflection session. The questions that were asked to participants were mainly about the test format, task difficulty, and test taking strategy use. In this reflection session, the researcher tried to get more information on their thought processes and their experience of listening tests. The questions were asked in a casual manner, and this session took 5 to 15 minutes per test-taker.

Methodologies for Different Levels of Inference

The procedures of the study were conducted based on Kane's (2006) argument approach for trait interpretations. Each research question reflected one step of Kane's inferences: scoring, generalization, extrapolation, and implication.

On the first level of inference: Scoring. For *scoring* inference, the research question was designed to verify whether listening tests were consistent and how accurately those tests measured test takers' listening proficiency. Both qualitative and quantitative methods were used to answer the first research question as stated below:

Research question 1

> To what extent does the newly developed listening comprehension test measure test takers' ability to understand academic lectures? How are the test takers' performances reflected in their test scores?

Assumption

> The iterative process of test specifications ensures quality of test items and the consistency of the test. However, breakdown factors may cause an issue with the consistency of the test scores.

Versions of test specifications were developed to ensure the details of the test and the quality of breakdown factors. As a qualitative approach, feedback from colleagues with language testing expertise who reviewed the first version of the test specification was reflected in the second version of test specifications. Elements of breakdown factors were revised based on the colleagues and the audio engineer in the next draft of the test specifications. Based on the final version of the test specifications, test items were revised and operationalized.

The data from the study was examined to check the consistency and accuracy of the test. Two sets of analyses were conducted. First, descriptive statistics and standard error of measurement were used to describe performance of the test takers and to interpret their test scores. Second, Cronbach's alpha and confidence interval of alpha were calculated to check the reliability of the test.

On the second level of inference: Generalization. The second research question reflects the next level of inference, generalization, to see if the samples of observations are representative

of the universe of possible observations. The following question and assumption were generated for collecting evidence of generalizability, which would provide a warrant for generalization to the universe score.

Research Question 2

To what extent do test takers vs. test types contribute to the source of variance? How do those different types of tests (regular vs. breakdowns) contribute to the source of variance?

Assumption

The sample observations are representative of the universe of observations, and the tests do provide a reliable estimate of listening ability. At the test level, the sample observation from different groups can be representative of the universe of observation regardless of breakdown factors.

At this level, quantitative analyses were used to collect evidence of the generalization inference. At the test level, a Latin square design was used to measure the performance difference between the four tests. Generalizability theory was implemented to check sources of variance by using GENOVA to estimate the relative contribution of variation between tests to variation in test scores. At the item level, classical item analysis was conducted to investigate how each item functions differently among the four versions of the test.

On the third level of inference: Extrapolation. For this level, quantitative analysis was used to provide evidence of validity. The research question of the inference was to determine the correlation between test takers' cognitive awareness and their test performance.

Research Question 3

To what extent do test takers contribute to the source of variance? How do their thought processes differ between the regular group and the breakdown groups? Does their cognitive awareness impact on their test performance?

Assumption

> Test takers who succeed in performing regular or breakdown listening tasks will have less difficulty understanding academic lectures in real situations. The processes involved in the test tasks could be the same as the processes of real academic listening.

The cognitive awareness questionnaire was provided to collect participants' reports of their various strategies on the different listening tests (regular and breakdown). Test takers answered the questionnaire immediately after completing the test to preserve their short-term memory and feeling about the test.

The fourth level of inference: Implication. For this level of inference, the main goal was to compare individual test performance on different tests. The trait in this study was second language listening itself, and the construct of the test was to measure second language listening comprehension. Traditional measurement of second language listening has relied on other traits or a broad sampling of observation to measure listening comprehension.

Research Question 4:

> To what extent are the implications associated with trait (second language listening) appropriate in this case? Does the evidence support the implications associated with the trait label?

Assumption

> This second language listening test is designed to reflect listening in terms of content, task types, procedures, context, and scoring. Samples of test takers show similar performance on different listening tasks, unless the test measures an irrelevant variance of the trait.

Both quantitative and qualitative methods were employed for this level of inference. To check irrelevance of trait, findings of quantitative analyses were used to address the research question. As a qualitative method, an interview from the reflection session was conducted with

test takers. The purpose of these interviews was to collect test takers' opinions on the test in general, its difficulty and other issues of the test.

Chapter 4

Results

Chapter Overview

In this chapter, the three stages of test validation are explained. At the preliminary stage of test development, findings from the needs analysis and the pilot study are reported to show how the features of the test changed during the test development stage. At this stage, the test format, listening construct and breakdown factors were examined and finalized. For instance, findings were reflected in the test format particularly in a change of numbers of speakers, topics of passages and the types of breakdown factors. After the test administration, data were analyzed based on the levels of inferences. First, the scoring inference shows the reliability and scoring patterns between the four groups (regular, British, speech rate, and noise). The reliability and scoring patterns were varied among these groups due to a possible breakdown effect. Second, the generalization inference states the interpretation from the observed score to a claim about the expected listening performance over a larger universe of observations in the testing procedure. The effects of the breakdown in the test were expected to surface in the observed score, and this claim could support the effect in the universe observation because the error variance was larger than the person variance component in the generalizability analysis, possibly indicating confounding variables. Third, the extrapolation inference demonstrates evidence that the test takers' cognitive awareness helped to process second language listening regardless of impediments in the test. The findings from the questionnaire showed that most metacognitive strategy uses were independent to the type of test group. Test takers' metacognitive strategy or listening process approaches were not changeable but were more as individualized. Finally, the trait implication inference shows that some aspects of an under-represented trait were measured,

so some findings were not sufficient to draw a conclusion about the effect of breakdown in second language listening comprehension. However, implications associated with the trait were supported by test takers' reflections on the test.

Preliminary Stage of Test Development

Before test development, a needs analysis and pilot study were conducted to verify applicable breakdown factors in an operationalized test. For the needs analysis, listening features and impeding factors were examined. Twenty free online lectures were reviewed to investigate these features. After the needs analysis was completed, the pilot study was conducted. The purpose of the pilot study was to select suitable breakdown listening samples for the operationalized listening test. The researcher and an audio engineer had several discussions and chose appropriate samples among recorded listening samples. With these samples, an initial test form was developed. The initial test form was administered to a non-native and a native speaker of English to verify the feasibility of the test. Their feedback was reflected on the final form of the test.

Findings of the needs analysis. The purpose of the needs analysis was to analyze possible impeding listening factors that might impede test takers' comprehension. In this analysis, the nature and variety of the tasks were considered when reviewing the free online lectures. A global approach was used to examine each one, and twenty were chosen from sources such as university online courses, second language learning websites, and the British Broadcasting Corporation (BBC).

Table 5 shows features of second language listening comprehension from the literature. The lectures were reviewed with reference to these features. The level of difficulty of each lecture was determined from a linguistic perspective to decide whether the content and sentence

formation were too difficult for the second language learners. The researcher also looked into the sound quality of the recording, speakers' accents, speech rate, noise, and authenticity of the lecture. These features were carefully reviewed to decide an appropriate degree of breakdown. Lastly, the length of each lecture was reviewed to determine an appropriate length for future lectures for the study.

The needs analysis showed that listening comprehension of these lectures can be adversely affected by a lengthy audio text, background noises, fast speech rates, multiple speakers, or time. Most lectures were recorded as delivered in the classroom, so the length of the lectures ranged from 50 minutes to two hours. Without visuals or texts, understanding the details of a lengthy audio text could be challenging for second language learners. Since these lectures were recorded in the classroom, random noises such as recording noises (static noise), chatting, paper shuffling and microphone noise, were detected. Certain kinds of words were not clear due to these noises and speakers' speech style. Speakers' speech rate was a distinctive feature of the lectures. Their speech rates were varied, and most speakers spoke a bit faster with a lot of pauses and fillers. These features made the lectures very authentic.

One study has supported the notion that the number of speakers and their gender could impede listening comprehension (Lumley & O' Sullivan, 2005). The number of speakers and number of interactions between these speakers complicated the cognitive load. Not only the number of speakers, gender could be a factor because listeners often report that certain voices are preferable for comprehension, and same gender speakers could be confused. This analysis showed that using same gender speakers could challenge test takers' comprehension if there are no extra materials or visual aids to distinguish each speaker.

Some of the sample lectures provided comprehension questions that were simple and structured and intended to help listeners get the main idea easily. However, these questions were not appropriate for judging second language learners' comprehension levels because they were simple and only pointed to the main idea. Since the lectures were authentic, designing comprehension questions that reflect the test construct could be challenging.

Table 5

Features of Listening Comprehension

Feature	Description
Linguistic Difficulty	Whether length of sentences, words, or verbs make it difficult to decode the meaning
Accent	Whether a speaker's accent makes it difficult to understand the context
Topic	Whether topic of audio texts makes it difficult to understand the context
Speakers	How many speakers are talking in the text? Is the number of speakers affecting comprehension?
Speech Rate	Is the speech rate too fast or slow to understand the context?
Authenticity	Is the text authentic in a real academic setting?
Noise/Audio Quality	Is the audio quality good enough to understand the context?
Types of tasks	Whether the task is presentation, lecture, or discussion
Context	What kinds of topic are used? (i.e., Practical, social, professional and abstract topics, particular interests, or special fields of competence)
Time	Is time length appropriate to understand the context of spoken text?

Pilot study. Prior to the test administration, the pilot study was conducted to enhance the validity of the test specifications and to select appropriate listening samples for the operationalized test. As Li (2006) suggested in her thesis, four versions of test specifications were created to enhance the validity of the test. The first version (ver. 0.2) represented the

outlines of the test and features that the researcher wanted to include in the test. The second version was refined with special focus on the format of breakdown and the regular test. In the test specifications of this version, not only item types were considered but also the measurement design. The third version was shaped in terms of breakdown types and talk types of the test, and the final version was the operationalized stage of test development. Based on test items from the third version of test specifications, the test forms were provided to three students. Students' comments were reflected in the final version of the test specifications. In this version, the final format of the test and information of operationalized breakdown factors of the test were included.

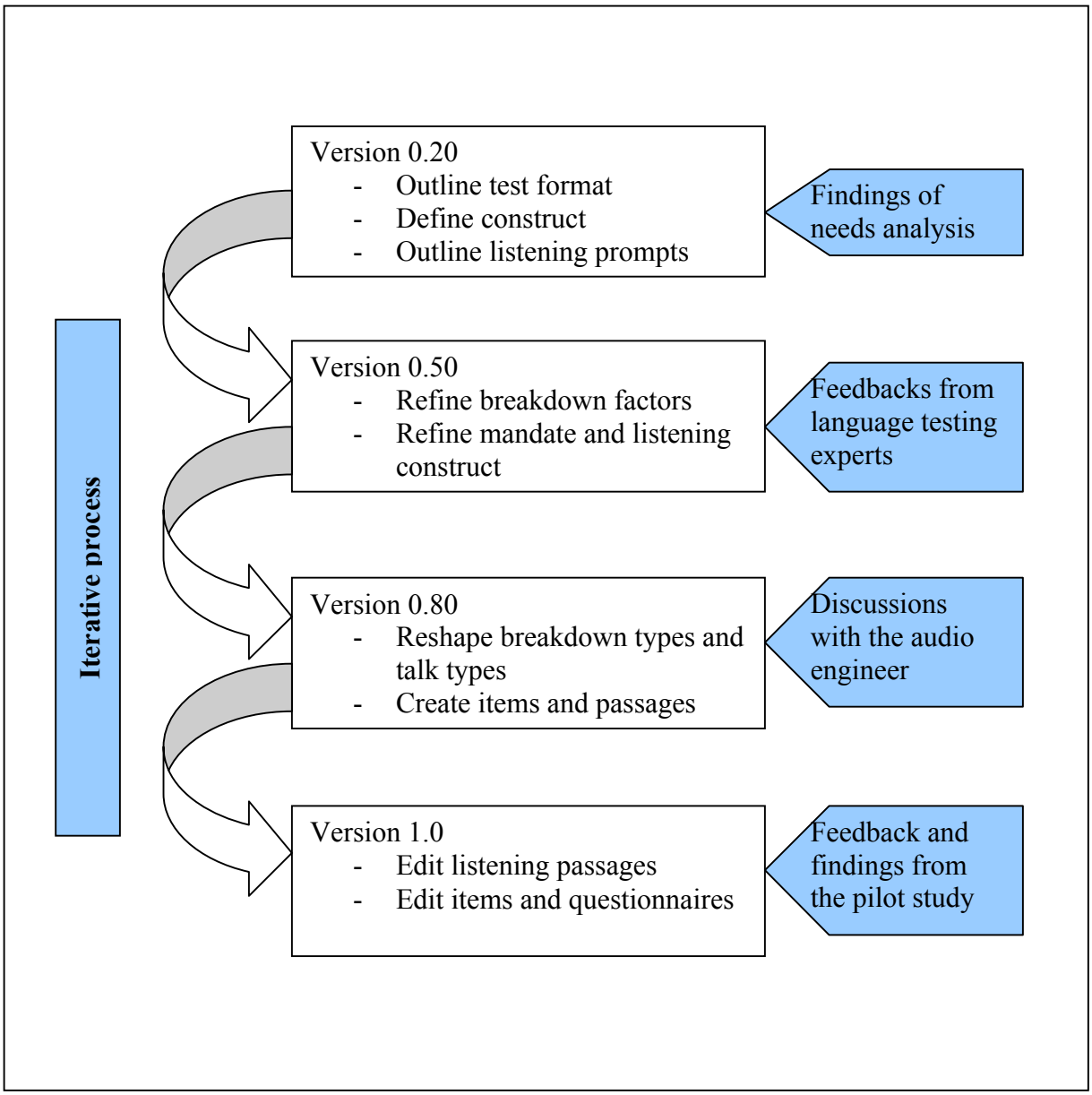

Figure 5. Iterative process of test specifications

Findings for Different Levels of Inference

After the pilot study was completed, an operational test was administered to 96 international students. The following research questions were addressed based on participants' test performance, responses on the cognitive questionnaire, and their reflections on the test. These four research questions were formulated based on Kane's validation framework.

On the first level of inference: Scoring. The first research question was formulated to investigate the validity of the test. To support the claim, test specifications and scoring of the test were investigated.

Participants were recruited individually so it was important to compare proficiency level of each group. Self-reported TOEFL score was used to compare participants' English proficiency level. A one-way ANOVA test was used to test for students' English proficiency among four groups. The null hypothesis was that test takers' TOEFL scores in each group are not different across the four groups. Test takers' TOEFL score did not differ significantly across the groups, F (3, 80) = 0.36, p =.78. Hence, we can assume that each group was formed with students who have fairly similar English proficiency level.

Table 6

One-Way ANOVA for Students' TOEFL Scores

	Sum of Squares	*df*	Mean Square	F	Sig.
Model	154.33	3	51.45	0.36	0.78
Error	11376.55	80	142.21		
Total	11530.89	83			

Descriptive statistics on test-takers' performance, the reliability of each test, and the standard error of measurement were used to answer the first research question. Table 7 shows that minimum and maximum scores among these four tests were not significantly different. The lowest mean (M=14.04) was from the regular group and the highest mean (M=15.13) from the British group. The standard deviation (SD = 2.42) was same for the British and the speech rate groups, but the regular group showed a larger standard deviation (SD =2.68) than other groups. The noise group showed a smallest standard deviation (SD=2.02) of all the groups. Though

there was no breakdown factor in the regular group, this group had a lower mean score and a larger standard deviation compared to the other groups.

Table 7

Descriptive Statistics for Test Scores

Test	N	Min	Max	Mean	SD
Overall	96	8.00	20.00	14.67	2.40
Regular	24	8.00	19.00	14.04	2.68
British	24	10.00	20.00	15.13	2.42
Speech Rate	24	9.00	19.00	15.08	2.42
Noise	24	10.00	18.00	14.46	2.02

Table 8 shows Cronbach's alphas for the four tests. Cronbach's alpha coefficient for the regular, British, and Speech rate versions was around 0.50, which is considered as poor internal consistency of items. Moreover, extremely low values on the noise test indicated that items were measured inconsistently, or possibly that other dimensions were detected besides second language listening. From this analysis, one of the breakdown factors, noise, exhibited a substantially distinct measurement result in comparison to the other three tests.

95% confidence intervals on the reliability value for each test showed that the width from the lower bound and upper bound of three tests (Regular, British, and Speech Rate groups) was similar, however, that of noise test was notably larger.

Table 8

Reliability Analyses for Four Tests

Test	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items	95% Confidence Interval	
				Lower Bound	Upper Bound
Regular	0.48	0.56	19	0.18	0.78
British	0.54	0.56	19	0.27	0.81
Speech Rate	0.49	0.56	19	0.19	0.79
Noise	0.08	0.10	20	-0.45	0.61

The standard error of measurement (SE*m*) was also calculated to estimate errors and to interpret each individual's test score (see Table 9). Using reliability coefficients from table 8 and standard deviation from table 7, the standard error of measurement was calculated for each test. The values of the standard error of measurement showed that the noise and regular tests had a larger SE*m* than the other two tests. Usually, the higher a test's reliability coefficient, the smaller the test's standard error of measurement, though the reliability coefficient of the regular test was higher, the value of the standard error of measurement was similar with the noise test.

Table 9

Standard Error of Measurements for All Tests

Test Type	Standard Error of Measurement
Regular	1.93
British	1.64
Speech rate	1.73
Noise	1.94

Summary of research question 1. From a standpoint of test consistency, the newly developed listening tests did not successfully measure second language listeners' listening ability. Mean performance of test takers from the regular, British accent, and speech rate groups was similar than performance of test takers from the noise group. The test items for each group were identical but the noise group's performance displayed extreme inconsistency as evidenced by its reliability analysis. Perhaps the noise condition produced measurement noise, as well.

On the second level of inference: Generalization. The following research question was formulated to generalize the listening process from the observed performance. Generalizability theory and GENOVA was the basis for estimating the source of variance of test types and test takers.

The generalizability study for the listening test had a crossed *p x i* design (person by item). Twenty items were administered to 24 individuals, and both items and persons were considered as random samples from the universe of items and the population of persons. A total of four analyses were conducted to estimate sources of variance for persons, test items, and interaction

between person and items. The estimated variance components from this study reflected the magnitude of error in generalizing from a person's score on a single item to the universe score.

Table 10

Frequency Table for Test Scores

Total score	Regular	British	Speech rate	Noise
8	1	0	0	0
9	0	0	1	0
10	0	1	0	1
11	5	3	2	2
12	1	0	1	0
13	3	0	1	3
14	2	4	2	6
15	4	3	3	5
16	3	5	9	4
17	4	7	2	1
18	0	0	2	2
19	1	0	1	0
20	0	1	0	0

Table 11 shows the results of the generalizability analysis of the person-by item-design. First, for the variance component for persons, the estimated universe-score variance accounted for less than 5% of the total variance across the four tests. The results show that the tests did not accurately measure individuals' listening proficiency. This low variance component may have

resulted from scores being tightly clustered around the grand mean. This clustering of scores is reflected in the lower estimated variance component for universe scores. Second, the variance components for items across the four tests were less than 20% of the total variance. The variance components for items were larger than those for persons but still relatively small. Table 10 and Figure 6 show the frequency of the test scores. It is clear that the scores are tightly clustered around the scores of 14 and 15, the mean of the tests. These results support the low variance component revealed by the generalizability analysis.

Lastly, the largest components were the residual interactions between items and persons. According to Shavelson and Webb (1991), large residuals can be interpreted in three ways: a) large interaction between items and persons, b) sources of error variability in the measurement that the one-facet, the *p x i* measurement could not capture, and c) both interaction and error variability of the measurement. In this study, breakdown factors might be captured in the variation of the *p x i* measurement, so the variance component was large. However, this does not confirm the breakdown effects because the residual for the regular test (control group) was even larger than that of the other tests.

Table 11

Generalizability Study (p × i Design)

Variance Component	Regular	British	Speech Rate	Noise
Person	3.96%	4.72%	3.86%	0.35%
Item	11.18%	15.36%	14.88%	16.97%
Person × Item	84.86%	79.92%	81.26%	82.68%

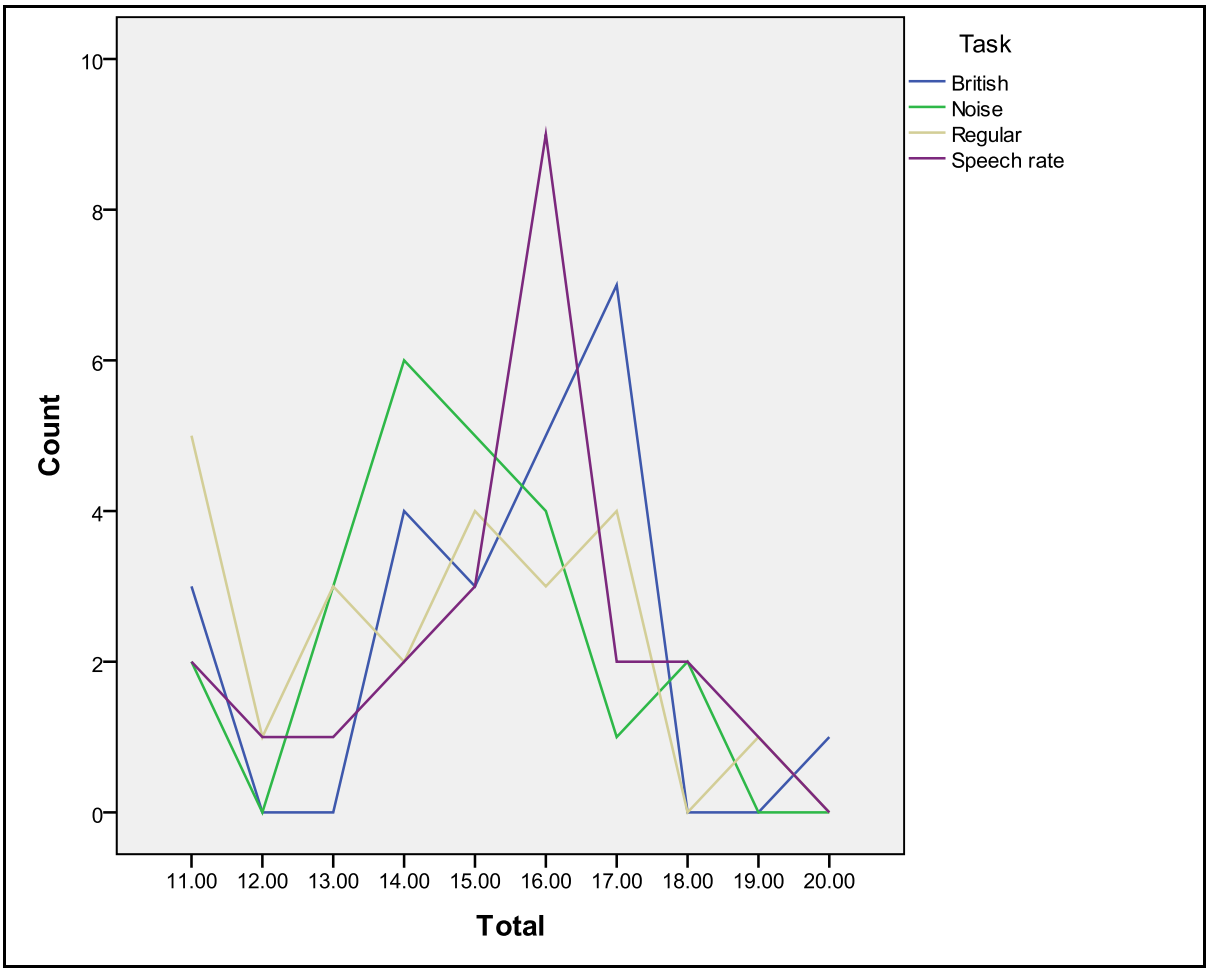

Figure 6. Frequency distribution of test scores

After the generalizability analysis, the experimental design was used to estimate the difference between the test types and talk types. Table 12 shows a 4-by-4 Latin Square design in which each cell has six different test takers' scores on different tests and prompts. The four tests were no impediment, accent, speech rate and noise impediments, and each test had four types of talks: presentation1, lecture 1, presentation 2, and lecture 2. The test group and talk types are fixed effects, and the score is considered as a random effect.

Table 12

4 × 4 Latin Square Design

Type	Regular	British	Speech Rate	Noise
Presentation1	A	B	C	D
Lecture 1	B	C	D	A
Presentation 2	C	D	A	B
Lecture 2	D	A	B	C

The null hypotheses of this analysis are: a) there is no significant difference in test takers' performance (scores) between the test types, and b) there is no significant difference in test takers' performance between prompt types (talk). Table 13 shows a significant difference in the model. In particular, there is a significant difference in talk types. In other words, the p-value for talk types rejected the null hypothesis, indicating a performance difference between the talk types. However, the p-value for the test types supported the hypothesis. This indicates that there is no performance difference between the test types.

Table 14 displays the differences in talk types. The Tukey comparison shows that the mean for presentation 2, *global cooling* and lecture 1, *dinosaur* are significantly higher than the means of lecture 3, *size of dinosaur* and presentation 4, *climate change*, and that there is no significant difference between lecture 3 and presentation 4.

Table 13

Result of 4 × 4 Latin Square Design

Source	*df*	*MS*	*F*	*Pr>F*
Model	6	339.50	6.78	0.01
Error	9	50.08		
Total	15			
Test	3	92.08	1.84	0.21
Talk	3	586.92	11.72	0.00

Table 14

Tukey Comparison for Talk Type

Talk comparison	Difference between means	Simultaneous 95% Confidence Limits		
Presentation 2- Lecture1	12.00	-3.62	27.62	
Presentation 2- Presentation 4	14.00	-1.62	29.62	
Presentation 2- Lecture 3	29.50	13.88	45.12	*
Lecture 1- Presentation 2	-12.00	-27.62	3.62	
Lecture 1- Presentation 4	2.00	-13.62	17.62	
Lecture 1- Lecture 3	17.50	1.88	33.12	*
Presentation 4- Presentation 2	-14.00	-29.62	1.62	
Presentation 4- Lecture 1	-2.00	-17.62	13.62	
Presentation 4- Lecture 3	15.50	-0.12	31.12	
Lecture 3- Presentation 2	-29.50	-45.12	-13.88	*
Lecture 3- Lecture 1	-17.50	-33.12	-1.88	*
Lecture 3- Presentation 4	-15.50	-31.12	0.12	

Note: a comparison significant at the 0.05 level is indicated by*.

I investigated if there was a breakdown effect among the four tests using the generalizability analysis and the experimental design. Results of these analyses do not explicitly show the effect of the breakdown factors, so I looked at the item level of the test to determine whether there was one item that was influenced by the breakdown factors. Two analyses were conducted: item difficulty and item discrimination analyses. The purpose of the classical item analysis is to diagnose whether items function appropriately. Since items were the same among

the four tests, an underlying assumption was that the items would function similarly if there was no breakdown effect. The item difficulty index is a measure of the proportion of examinees who answered the item correctly. It can be range between 0 and 1 and higher p-value indicates an easier item. The item discrimination index is a measure of how well an item is able to distinguish between examinees with higher scores and lower scores, because it is a correlation index between the item score and the total score. The range of the discrimination index is -1.0 to 1.0; however any item discrimination index below 0 suggests a problem.

Table 15 shows that a few items were too easy for the test takers. The rule of thumb (Bachman, 2004) for identifying good items is to use those whose difficulty falls between a p-value of 0.25 and 0.75, or 0.20 and 0.80. For instance, items 1, 2, 7, 9, 10, and 19 are marked as easy items. If the difficulty value of these items is too high, in terms of item quality, they would not be good items to test second language listening appropriately.

The item difficulty indices functioned similarly across four tests. The items were marked as easy items (item number 1, 2, 7, 9, 10, and 19) functioned similarly among the four tests. Most items that fell in a range of p-values between 0.25 and 0.75 functioned similarly among the four tests even though the p-values are different, however item number 11 and 14 show lower p-value on three breakdown tests.

Table 15

Item Difficulty Analysis for Four Tests (p-value of Items)

Item	Regular	British	Speech rate	Noise
1	0.88	1.00	0.96	0.96
2	0.83	0.79	0.83	0.88
3	0.63	0.96	0.88	0.96
4	0.42	0.63	0.38	0.50
5	0.79	0.63	0.63	0.58
6	0.67	0.67	0.71	0.75
7	1.00	0.96	1.00	0.96
8	0.75	0.75	0.83	0.71
9	0.96	0.92	0.96	0.96
10	0.83	0.92	0.96	0.83
11	0.42	0.25	0.33	0.21
12	0.54	0.42	0.58	0.67
13	0.63	0.67	0.75	0.75
14	0.63	0.83	0.71	0.42
15	0.54	0.83	0.67	0.50
16	0.67	0.79	0.83	0.79
17	0.79	0.71	0.79	0.79
18	0.58	0.75	0.67	0.63
19	0.96	0.92	0.92	0.88
20	0.50	0.63	0.58	0.58

Item difficulty analysis shows that items functioned similarly among the four tests, the regular and the three breakdown tests. An item discrimination analysis was conducted to determine how well each item discriminated between persons who scored high on the test as a whole and persons who scored low on the test as whole. A common rule of thumb is to include items that have discrimination indices equal to or greater than 0.30 (Bachman, 2004). In the regular test, the following items have strong discrimination indices: items 2, 8, 13, 14, 16, 17, 18 and 20. In the British test, items 2, 4, 7, 8, 9, 15, 17 and 18 have good discrimination indices. In the Speech rate test, most items had relatively large discrimination indices except items 5, 9, 10, 12, 13, 15, 16, 17 and 19. For the noise test, items 1, 3, 6, 14, 15, and 17 have good discrimination indices. Though the items were the same, each test's item discrimination indices were different, but certain items were marked as items with lower discrimination indices across five tests (item 5, 10, 12 and 20).

Item difficulty and discrimination analyses flagged items 10 and 20 as not performing well. These items were too easy, and they did not have discrimination power between individuals with higher scores and individuals with lower scores. I also found that the difficulty of items in the four tests showed similar p-values even though there were impediments in the test. However, different item discrimination indices across four tests indicate that some individuals with lower scores were hindered by the impediments in the test.

Though I provided the same items to the four groups, item statistics of the twenty items were different which lead to the possibility of a breakdown effect at the item level. Hence, item comparison of the item difficulty and item discrimination of each group was plotted to find out different patterns of each group. Figures 7 to 10 show different patterns of item comparison for each group.

The item comparison patterns changed over breakdown groups. A couple of items with discrimination indices ≈ 0 were found in the regular and noise groups (see Figure 7 and 10). Some difficult items were such that almost everyone got them wrong, and some items were so easy that almost every test taker got them right. Furthermore, negative discrimination indices were found in the British and speech rate groups (see Figure 8 and 9), regardless of whether the item focused on main idea or on details. These breakdown effects were detected in the item discrimination indices rather than in the item difficulty indices. Though test takers are higher proficiency students, perhaps they could get an easy item wrong if they were bothered by these listening impediments.

Table 16

Item Discrimination Analysis for Four Tests

Item/Group	Regular	British	Speech rate	Noise
1	.170	0.00	.451	.387
2	.452	.446	.322	.157
3	.189	-.161	.465	.268
4	.214	.350	.301	.023
5	-.193	.013	.151	-.131
6	.059	.060	.455	.296
7	.216	.345	.482	-.288
8	.422	.677	.000	.055
9	.211	.258	.267	-.182
10	.178	.133	-.110	.178
11	.102	-.058	.296	-.183
12	-.247	.075	-.070	-.095
13	.301	.172	.029	-.039
14	.472	.217	.334	.262
15	-.046	.314	-.249	.441
16	.387	-.080	.171	-.082
17	.326	.358	.013	.330
18	.422	.307	.402	-.223
19	-.220	.195	-.247	-.181
20	.452	-.056	.322	-.336

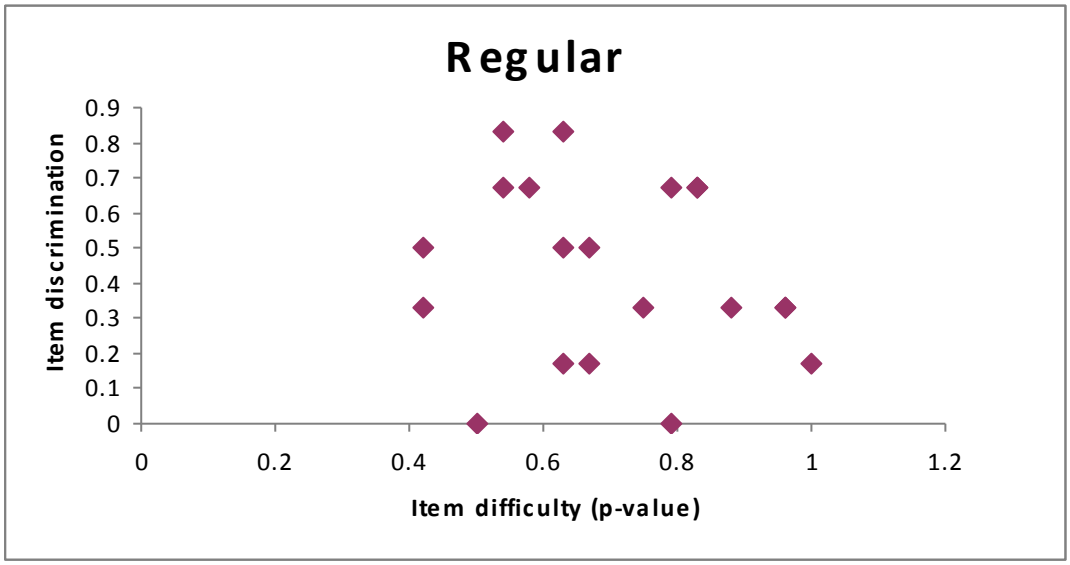

Figure 7. Item comparison of the regular group

Figure 8. Item comparison of the British group

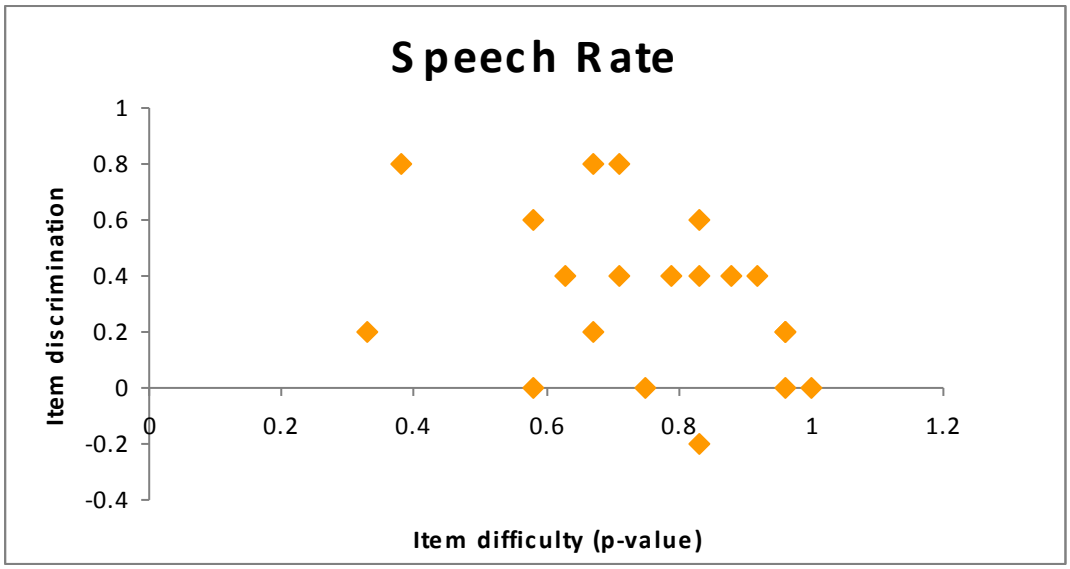

Figure 9. Item comparison of the speech rate group

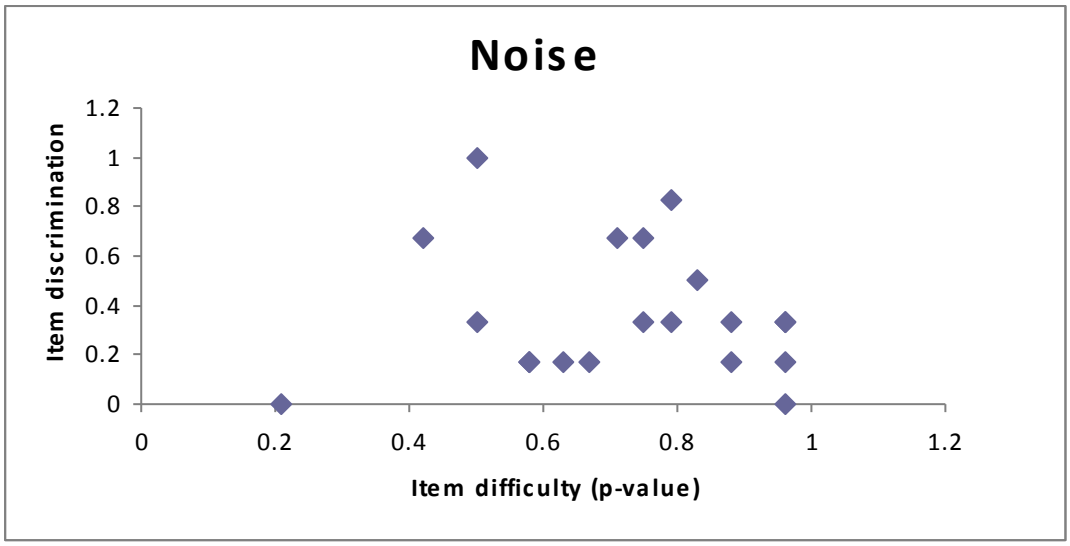

Figure 10. Item comparison of the noise group

Summary of research question 2. The assumption was that test takers' listening proficiency test scores should not be affected under different listening condition. Their performance on the test could be generalized as their usual listening proficiency in academic setting. However, the findings of analyses showed that the test measured test takers' listening proficiency. First, Generalizability theory showed that the error variance was larger than systematic variance in all cases. This finding indicates that the tests were not precise at

measuring the intended construct (listening). Also, findings of the ANOVA analysis showed that there was no intended breakdown effect on test level such as British accent, speech rate, or noise. Though there was no group difference in test type, the group difference was found in listening passages (talk type). These lack of results on the ANOVA could be due to the large error variance, previously noted.

Second, although a breakdown effect was not detected at the test level, there was some evidence of it at the item level. Items of each test functioned unexpectedly, given that the items were parallel across the test versions. In particular, some easy items had low item discrimination values beyond what would normally be expected at that item difficulty range. The British and Speech rate tests had items with negative item discrimination value that shows possible misunderstanding of the question.

Whether it was due to the intentional breakdown or not, the large error variance in the tests indicated that the test with regular and breakdown factors was not appropriate as a sample to generalize universal scores. The tests may have been measuring a second language listening construct, but it is clear that the tests were also producing unsystematic variance, and possibly also measurement of unknown trait (although that cannot be precisely determined based on the instrumentation of this study). Item types and listening passages were revealed as possible causes of group difference, to the extent that such differences are noteworthy given the unsystematic variance.

On the third level of inference: Extrapolation. Previous research questions focused on the development of the test and its validation. The following research questions were raised to understand test takers' cognitive awareness of their performance. To answer this question, a cognitive questionnaire was analyzed to investigate the third level of inference.

Research Question 3

To what extent do test takers contribute to the source of variance? How do their thought processes differ between test takers who listen to regular tests and breakdown tests? Does their cognitive awareness impact on their test performance?

Assumption

Test takers who succeed in performing breakdown listening tasks will have less difficulty understanding academic lectures in real situations. The processes involved in the test tasks could be the same as the processes of real academic listening.

The cognitive questionnaire was composed of four parts: self-awareness of listening proficiency, listening process, metacognitive strategy use, and listening test-taking strategies. Responses to the questionnaire were reported by test group and by the entire group. To find the association between test groups and test takers' responses on questionnaire, a Chi-square test for independence was conducted.

First, cognitive questions that focused on awareness of second language listening were asked (see Table 17). Most test takers answered that listening comprehension was challenging for them. In particular, a majority of participants from the regular and the noise groups reported that they were not confident listening in English. The participants from the British group reported that they were confident in second language listening. For the speech rate group, over 50% of participants reported that second language listening was challenging for them.

Based on the responses on the question, a Chi-square test for independence was conducted. The null hypothesis for this test was that there is no association between test groups and responses (yes/no) on awareness of second language listening. The Chi-square value was 17.15 with $df = 3$. The significance level was less than 0.05 so there was a significant difference between two variables. Therefore, responses from different test groups were associated with the responses on the self-awareness question.

Table 17

Self-Assessment on Second Language Listening

Question	Disagree	Agree	Don't remember
Listening comprehension in English is a challenge for me			
Overall	39.6% (38)	59.4% (57)	1.0% (1)
Regular	16.7% (4)	83.3% (20)	0% (0)
British	70.8% (17)	29.2% (7)	0% (0)
Speech Rate	45.8% (11)	54.2% (13)	0% (0)
Noise	25.0% (6)	70.8% (17)	4.2% (1)

Table 18 shows the participants' responses on questions about their awareness of second language listening. Unlike their self-assessment of second language listening, a majority of participants reported that they did not feel nervous when they listened to the test. Based on their response, it seems that the breakdown factors did not make them nervous. This indicates that the testing environment was not stressful, and the breakdown factors were not significantly disturbing to test-takers' comprehension of the lectures.

The Chi-Square test also confirmed the finding of the frequency table. Chi-Square value was 1.08 with $df = 3$ and p-value was 0.78. The significance level was more than 0.05 so the null hypothesis was retained. Therefore, the test groups and test takers' responses on this question were independent. Regardless of test type, test takers did not feel nervous during the test.

Table 18

Cognitive Awareness on Second Language Listening

Question	Disagree	Agree	Don't remember
I don't feel nervous when I listen to the listening passage			
Overall	36.5% (35)	60.4% (58)	3.1% (3)
Regular	37.5% (9)	62.5% (15)	0% (0)
British	33.3% (8)	62.5% (15)	4.2% (1)
Speech Rate	29.2% (7)	62.5% (15)	8.3% (2)
Noise	45.8% (11)	54.2% (13)	0% (0)

The purpose of part two of the questionnaire was to measure different approaches to the listening process and to compare listening processes among the test groups (see Table 19). First, the listening process approach was investigated to determine whether test takers used a top-down or a bottom-up approach to understand the listening passages. One question asked whether they translated word by word while listening. Most respondents disagreed that they translated words into their first language while listening in English. The second question asked whether the test takers focused on each sentence or on the overall context of the passage. The findings from both questions showed that test takers did not use the bottom-up approach.

In Table 19, the first question was not appropriate for a Chi-Square Test because the 20% of expected values were less than 5. Thus, Chi-Square test was conducted for the second question. The null hypothesis of this test was that there is no association between test groups and listening process approaches. The Chi-Square value was 1.61 with $df = 3$ and p-value was 0.66 that exceeded the significance level of 0.05. Since the p-value was more than the significance

level, the null hypothesis was retained. Therefore, there is no relationship between test groups and listening process approach.

In other words, they were not paying attention to each word or sentence while they were listening to the prompts. The possible reason for using the top-down approach was that the length of the passages was long, so it couldn't have been easy for them to use the bottom-up approach. However, the bottom-up approach was required in order get detailed information in the test. These findings raise a question about how the test takers succeeded on detail questions if they only used the top-down approach.

Test-takers' metacognitive process of listening was also investigated in the questionnaire. Several questions were provided to verify whether they used metacognitive knowledge while taking the test. Researchers have claimed that the learners who have metacognitive knowledge have better understanding when they listen to a new concept or knowledge in a second language (Goh, 2000).

Table 19

Listening Process Approaches

Question	Disagree	Agree	Don't remember
I translate word by word as I listen			
Overall	88.5% (85)	7.3% (7)	3.1% (3)
Regular	83.3% (20)	4.2% (1)	8.3% (2)
British	95.8% (23)	4.2% (1)	0% (0)
Speech Rate	87.5% (21)	8.3% (2)	4.2% (1)
Noise	87.5% (21)	12.5% (3)	0% (0)
When I listen to the passage, I tend to focus on understanding the meaning of each sentence rather than the overall meaning of the text			
Overall	71.9% (69)	24.0% (23)	3.1% (3)
Regular	70.8% (17)	29.2% (7)	0% (0)
British	79.2% (19)	16.7% (4)	4.2% (1)
Speech rate	75.0% (18)	20.8% (5)	4.2% (1)
Noise	62.5% (15)	29.2% (7)	4.2% (1)

Three questions were provided to understand test takers' planning and monitoring processes when they received the listening prompts. Table 20 shows that a majority of respondents planned how they were going to listen before they started the test. Almost 90% of the test takers previewed the comprehension questions before they listened to the passage because the questions helped them to plan their listening.

To test the association between test groups and metacognitive process, only first question was applicable for the Chi-Square test and this was not used for other two questions due to insufficient value of some cells. The Chi-Square test showed that there is no association between test groups and metacognitive process especially planning ($X^2 = 0.38$, $df = 3$, $p = 0.94$). Regardless of test type, the test takers used similar planning in order to trigger their metacognition.

Table 20

Metacognitive Process in Second Language Listening: Planning

Question	Disagree	Agree	Don't remember
Before I start to listen, I have a plan in my mind for how I am going to listen			
Overall	32.3% (31)	59.4% (47)	8.3% (8)
Regular	33.3% (8)	54.2% (13)	12.5% (3)
British	29.2% (7)	66.7% (16)	4.2% (1)
Speech Rate	33.3% (8)	54.2% (13)	12.5% (3)
Noise	33.3% (8)	62.5% (15)	4.2% (1)
I read comprehension question first before I listen to the passage			
Overall	10.4% (10)	86.5% (83)	1.0% (1)
Regular	20.8% (5)	79.2% (19)	0% (0)
British	4.2% (1)	91.7% (22)	0% (0)
Speech rate	8.3% (2)	87.5% (21)	4.2% (1)
Noise	8.3% (2)	87.5% (21)	0% (0)
Before I listen to the passage, I try to predict the content of the passage by reviewing comprehension			
Overall	17.7% (17)	79.2% (76)	2.1% (2)
Regular	16.7% (4)	83.3% (20)	0% (0)
British	12.5% (3)	83.3% (20)	4.2% (1)
Speech rate	20.8% (5)	75.0% (18)	4.2% (1)
Noise	20.8% (5)	75.0% (18)	0% (0)

As Table 21 shows, another two questions were asked to determine whether test takers monitored their listening/test-taking processes while they were taking the test. Their responses indicate that most were able to adjust their interpretation when they perceived an error. Though most of the test takers were able to adjust their interpretations, 50% of the test takers from the regular group were monitoring themselves constantly, while most test takers in other three groups did not. This also indicates that these test takers did not fully comprehend the whole lectures.

A Chi-Square test was conducted to examine the relationship between the test groups and metacognitive process particularly monitoring ($X^2 = 4.45$, $df = 3$, $p = 0.22$). The p-value of second question in Table 21 was more than the significance level (0.05). Therefore, the null hypothesis was retained. It suggests that the test groups and monitoring process are not related.

Table 21

Metacognitive Process in Second Language Listening: Monitoring

Question	Disagree	Agree	Don't remember
As I listen, I quickly adjust my interpretation if I realize that it is not correct			
Overall	16.7% (16)	60.4% (58)	22.9% (22)
Regular	25.0% (6)	58.3% (14)	16.7% (4)
British	16.7% (4)	66.7% (16)	16.7% (4)
Speech Rate	8.3% (2)	62.5% (15)	29.2% (7)
Noise	16.7% (4)	54.2% (13)	29.2% (7)
As I listen, I periodically ask myself if I understand everything that I heard			
Overall	56.3% (54)	36.5% (35)	7.3% (7)
Regular	45.8% (11)	50.0% (12)	4.2% (1)
British	54.2% (13)	45.8% (11)	0% (0)
Speech rate	58.3% (14)	29.2% (7)	12.5% (3)
Noise	66.7% (16)	20.8% (5)	12.5% (3)

The third part of the questionnaire investigated test takers' metacognitive strategy use. Table 22 shows the four questions provided to verify their strategy use. The first three questions regarded test takers' strategy use when they did not understand a certain part of the passage. The last question was provided to investigate their different strategy uses on the short and long passages.

The first question concerned the use of a guessing strategy, and almost 90% of test takers from all groups reported that they used guessing strategy when they did not understand the passage. Their responses confirmed that guessing is a common strategy for language learners to fill the comprehension gap when they perceived comprehension obstacles.

The second question regarded listening passively. More test takers from the regular and the noise groups responded that they listened passively when they could not understand the passage. Over 50% of the British test takers also responded that they used passive listening when they did not understand the passage. The speech rate group showed insignificant difference in responding to this question. The Chi-Square value was 3.20 with $df = 3$ and the p-value (0.36) exceeded the significance level (0.05). Therefore, the null hypothesis was retained. A Chi-Square test for independence also showed that there is an association between test groups and passive listening. As was the case with the guessing strategy, the passive listening strategy was also not relevant to the test types.

In response to the third question, all of the test takers reported that they did not give up listening to the lectures even though they perceived comprehension obstacles. A majority of test takers responded that they tried to comprehend the meaning of the passage even though they could not understand it. A Chi-Square test showed that there was no association between the test

groups and listening strategy ($X^2 = 2.28$, $df = 3$, $p = 0.52$). Regardless of test types, the test takers completed the test even though they encountered listening barriers.

To the question asking about different strategy use depending on the length of passages only the British group agreed that they used different strategies when listening to short and long passages. On the other hand, most of the speech rate group reported that the length of the passage did not cause them to change their strategy. However, a Chi-Square test showed that responses from different group were not related to the response on this question. The Chi-Square value was 6.13 and the p-value (0.11) exceeded the significance level of 0.05 so the two variables were not independent of each other.

Table 22

Metacognitive Process in Second Language Listening: Strategy Uses

Question	Disagree	Agree	Don't remember
When I listen to the listening passage, if I don't understand something, I guess what the word or phrase might mean based on the context			
Overall	5.2% (5)	91.7% (88)	3.1% (3)
Regular	8.3% (2)	91.7% (22)	0% (0)
British	0% (0)	95.8% (23)	4.2% (1)
Speech Rate	0% (0)	91.7% (22)	8.3% (2)
Noise	12.5% (3)	87.5% (21)	0% (0)
When I listen to the passage, if I don't understand something, I find myself thinking about the segment and passively listening			
Overall	28.1% (27)	59.4% (57)	12.5% (12)
Regular	25.0% (6)	70.8% (17)	4.2% (1)
British	25.0% (6)	58.3% (14)	16.7% (4)
Speech rate	41.7% (10)	45.8% (11)	12.5% (3)
Noise	20.8% (5)	62.5% (15)	16.7% (4)

Table 23

Metacognitive Process in Second Language Listening: Strategy Uses

Question	Disagree	Agree	Don't remember
When I listen to the passage, if I don't understand something, I just give up trying to comprehend the passage			
Overall	75.0% (72)	21.9% (21)	3.1% (3)
Regular	70.8% (17)	29.2% (7)	0% (0)
British	75.0% (18)	20.8% (5)	4.2% (1)
Speech rate	87.5% (21)	12.5% (3)	0% (0)
Noise	66.7% (16)	25.0% (6)	8.3% (2)
I use different listening strategies when I listen to long or short talks			
Overall	38.5% (37)	43.8% (42)	15.6% (15)
Regular	33.3% (8)	45.8% (11)	20.8% (5)
British	20.8% (5)	58.3% (14)	16.7% (4)
Speech rate	58.3% (14)	33.3% (8)	8.3% (2)
Noise	41.7% (10)	37.5% (9)	16.7% (4)

The last part of the questionnaire asked about test takers' test-taking strategy. When I administered the test to the test takers, I attempted to create a comfortable test-taking environment to observe test takers' various strategies. The test takers were allowed to take notes, preview comprehension questions, and/or use their own test taking strategy. The purpose was to scrutinize as many strategy uses from them as possible and to verify the most effective strategy in listening assessment. As Table 24 shows, the regular, the speech rate, and the noise groups did not use note-taking strategy; however some test takers from the British group did use the note taking strategy. To confirm the relationship between the test groups and test taking strategy, a Chi-Square test was conducted. The Chi-Square test also showed that different group can have different response on this question. The Chi-Square value was 5.43 with $df = 3$ and the p-value (0.14) exceeded the significance level of 0.05 so the two variables were independent of each other.

As the literature has supported (Schmidt–Rinehart, 1994), most test takers from all groups agreed that background knowledge on topics was very helpful to understand each passage. The Chi-Square test also showed that test takers' responses on background knowledge could be changed regardless on what test groups they were in. The null hypothesis of this test was that there is no association between the test group and test taking strategy. The value of Chi-Square was 2.90 with $df = 3$, and the p-value (0.41) exceeded the significance level (0.05). Thus, the null hypothesis was retrained.

At the test development stage, I considered this listening element during passage selection and incorporated it into the test specifications. Most technical terms were explained in the passage, and the initial assumption was that if a test taker has good listening proficiency, background knowledge might not play a crucial role in the assessment. Still, respondents

reported that background knowledge played a big role in second language listening assessment because they could process the knowledge faster regardless of proficiency level.

Table 24

Test-Taking Strategy Use in Second Language Listening

Question	Disagree	Agree	Don't remember
Note-taking is very helpful when I listen to the passage			
Overall	57.3% (55)	32.3% (31)	8.3% (8)
Regular	54.2% (13)	37.5% (9)	8.3% (2)
British	41.7% (10)	41.7% (10)	12.5% (3)
Speech Rate	79.2% (19)	16.7% (4)	4.2% (1)
Noise	54.2% (13)	33.3% (8)	8.3% (2)
Background knowledge of the topics was helpful to answer the comprehension questions			
Overall	24.0% (23)	71.9% (69)	2.1% (2)
Regular	16.7% (4)	79.2% (19)	4.2% (1)
British	20.8% (5)	70.8% (17)	4.2% (1)
Speech rate	37.5% (9)	62.5% (15)	0% (0)
Noise	20.8% (5)	75.0% (18)	0% (0)

Summary of research question 3. Findings of the cognitive questionnaire showed that test takers' thought process between the regular group and breakdown groups was not significantly different. Test takers' awareness of second language listening proficiency was varied among groups. In particular, the regular and the noise groups reported that listening in English was challenging. In terms of metacognitive strategy uses and test taking strategy, there

was no significant association between test groups and test takers' responses. Hence, there was no evidence that test takers' cognitive process or strategy uses was hindered by the impediments.

The fourth Level of Inference: Implication. Three research questions were formulated to investigate validity and reliability of the test, test takers' cognitive awareness, and their strategy uses. The last research question was formulated to identify implications of the second language listening test. Findings from this question complete a picture of test validation and suggest implications for second language listening.

Research Question 4:

> To what extent are the implications associated with trait (second language listening) appropriate in this case? Does the evidence support the implications associated with the trait label?

Assumption

> Second language listening test is designed to reflect listening in terms of content, test types, procedures, context, and scoring. Samples of test takers show similar performance on different listening tasks unless the test measures irrelevant variance of the trait.

At this level, three aspects of test validation are discussed: test level, item level, and cognitive awareness. Based on the findings of the previous research questions, the test-takers' interviews are added to understand the phenomenon and to strengthen validity claims.

Test level. The experimental design was used to measure the effects of the tests; breakdown tests and the regular test. The previous finding of the experimental design showed no significant effect between test types and breakdown factors. To confirm this finding, I interviewed test takers after they finished the test. One of the reflection questions asked for an overall impression of the test and what they perceived as the distinctive features of the test. From this question, I also expected to hear test takers' views on the effect of the breakdown

factors. Test taker A, who took the regular test and received score 16 (out of 20), pointed out that the test format is very similar to the TOEFL test.

> Test taker A: It is very similar to the TOEFL. The topics were not easy for me, but the test was familiar to me so I was comfortable taking the test.

Test taker B, who took the British test and received score 15 (out of 20), agreed that the test format was similar to the TOEFL, and he also pointed out that the British accent was easy to understand. Hence, the sound was not disturbing when he comprehended the passages.

> Test taker B: It is like the TOEFL. Items were familiar because I took the TOEFL a few times. The items and the style of the passages are like the TOEFL. My test, the speaker was British. His speech did not bother me at all. I think he speaks slowly so it was not so difficult to understand him, but the first topic, dinosaur, was difficult for me because I don't know much about dinosaur.

Test taker C, who took the speech rate test and received score 14 (out of 20), said the following about the overall impression of the test.

> Test taker C: I don't know much about dinosaur and weather. The topics were so hard so I don't think I did well on the test. The speaker in the test spoke a bit fast but I was not bothered by the speed of his speech. I missed several words in the passages but I understood overall meaning of the passages.

Test taker C mentioned that the topics of the passages were not familiar, so the impression of the test was not positive. Due to the topic difficulty, the speech rate did not function as a breakdown factor for this test taker. From this statement, we could assume that topic familiarity could be more influential than the breakdown factors for some test takers.

Test taker D, who got score 19, also confirmed the important relationship between topic familiarity and listening comprehension. She mentioned that the test was not difficult because she was familiar with the topics and the content of the passages was predictable.

> Test taker D: The test was good. I knew about both topics so it was easier for me to answer the questions. Questions were also predictable. I did not really notice the speed of the passage.

Unlike other test takers who reported that they were not bothered by the breakdown factors, test takers who took the noise test reported that they were bothered by the noise in the passages. And the noise factor affected their overall impression of the test.

> Test taker E: It was really hard. Static noise was really disturbing. I missed a lot of the lectures. I tried to concentrate first and second lectures but I got tired at the end. I don't think this is a good listening test. And I don't think I did well on the test.

> Test Taker F: It was challenging. Noise was really disturbing. But somehow I managed it and I was getting used to it. From the second lecture, the noise made me concentrate more on the lectures. Still I think I did not do well on the test. I don't understand why you include noise in it.

According to test taker E and F, neither test taker E (score 10) nor F (score 17) could comprehend the lectures as much as they normally could due to the noise. Test taker E's TOEFL iBT score was 110, and test taker F's TOEFL iBT score was 120. However, test taker F found a way to manage the noise and understand the content of lectures. Based on both TOEFL and the test score, test taker F's listening comprehension ability was better than that of test taker E. The better performance within the same condition may have resulted from test taker F's listening strategy to manage the background noise or she was a better listener.

The interviews showed that among the breakdown factors only noise functioned as an impediment to listening comprehension. They also showed that topic familiarity played a bigger role in second language listening comprehension. In other words, two of the breakdown factors (British and speech rate) in this study could be used in other listening passages in future test development because these factors did not affect listening comprehension. These factors could be added into listening prompts to make the prompts more authentic.

Item level. At this level, an interview question focused on test takers' opinions of the test items. A dominant point of feedback was that the item types were familiar to the test takers. They also agreed that those items were appropriate to measure second language listening. Though they complained about the listening passages, test takers did not mention the difficulty of the test items. This feedback is consistent with findings from the classical item analysis (see Tables 15 and 16).

Interestingly, two test takers pointed out a relationship between test performance and the item order of the test. Test takers expect test items in the exact order as presented in the listening passage. They missed a few items because they were expecting them in a different order. Test taker G received score 13, and the test taker H received score 17.

> Test taker G: the items were not ordered by the content of the passage. I was looking at number 2 and I waited…waited…and I missed the rest of the questions. I previewed the items but I thought I can answer the question one by one while I was listening.
>
> Test taker H: When I listen to the dinosaur lecture, I missed a couple of items because of the item order. So I read the questions first and took notes and then I played the next lecture. It helped a lot. I did do well on the first lecture but it was fine after I previewed the items.

From their comments on item order we can assume that item sequence could affect test performance. Higher proficiency test takers eventually overcame this issue, but lower proficiency students could not. At the test development stage, the researcher received similar feedback from a colleague, but the operational test was already revised. However, a couple of test takers mentioned that the item order was an irrelevant factor to their performance, so the implication for this issue is that more research on item ordering is necessary to determine the severity of item ordering effects.

Cognitive awareness of second language listening. One of the goals in this study was to investigate test takers' cognitive awareness of second language listening. This area has been studied deeply in teaching and learning but not in language assessment. The assumption of this study was that test takers with higher proficiency or higher cognitive awareness could perform better on the listening test with or without breakdown factors because they could manage their listening strategy when they perceived listening obstacles.

In the interview session, test takers were asked to share their strategy uses and cognitive awareness of each lecture. Test taker I explained her listening process approach during the interview. She said she was confident in English so she did not use a particular strategy to aid the comprehension, and she received score 11. She regretted that she did not use a particular strategy when she perceived the difficulty.

> Test taker I: I thought I heard everything, but I could not answer a few questions because I don't remember the answer whether it was hurricane or tornado. I regretted that I should have taken notes.

Test taker J, who received score 17, perceived his listening proficiency as at an intermediate level. He did not solely rely on listening ability. Instead he tried to use different strategies to overcome the difficulty. One of the strategies that he used was to write down key words and focus on those key words while he was listening. He thought he might not comprehend the lectures completely, so he strategized how he would comprehend the lectures by key wording based on previewed comprehension questions. His cognitive awareness prompted him to use this strategy, and the metacognitive strategy that he used was effective to overcome his weakness in listening.

> Test taker J: I reviewed the questions first, and then I wrote down the key words that I am going to listen. When I listen to the lectures, I tried to hear these key words. I did not pay

attention to the rest of the lectures. I don't really know about the topics but I think I did okay on the test.

Test taker K took the speech rate test and received score 9. He did not share his opinions much because he did not remember how he did on the test. He was confident, but he showed some signs of metacognition deficiency and his test score was the lowest. He mentioned the topics of the lectures but was not able to explain what he understood from the test. His attitude was not fully positive, possibly due to cognitive awareness of his listening proficiency and his performance on the test. He did not give up listening to the lectures, but also did not try to overcome the comprehension breakdown because he was not able to apply cognitive knowledge and skills.

Test taker K: I don't remember. It was about dinosaur and the weather change. I cannot answer your questions (reflection questions). I don't remember. But I think I did well. I don't really take notes but I did preview the questions. But it did not really help me to answer the questions.

Test taker L, who received score 18, reported a higher metacognitive strategy use during the test: pre-directed listening. He perceived himself as a bilingual, so his confidence level in English was high. At the planning stage, he previewed the comprehension questions and tried to draw a picture of the lectures before listening to them. His schema was fully activated. He tried to answer the comprehension questions based on his background knowledge. After he finished marking the answers, he listened to the lectures. Instead of answering questions by listening, he listened to the lectures to verify his pre-selected answers. He occasionally changed an answer based on his comprehension. He used metacognitive strategies such as planning and monitoring. As a test-taking strategy, he used previewing comprehension questions and eliminating distractors.

> Test taker L: I read the questions first and marked the possible correct answers. While I was listening to the lectures, I crossed out the wrong answers and confirmed my previous answer. When I take listening tests, I always answer questions in this way.

Test taker M took the noise version of the test and received score 10. She did not review the comprehension questions prior to the listening. After she perceived difficulty understanding the lectures, she modified her test-taking strategy by adding note-taking. She must have constantly monitored herself because she knew what she missed after the test. She was using a metacognitive strategy, monitoring, but she had a production deficiency so she could not apply her knowledge and skills due to the noise. Clearly, there was a breakdown effect in this case. The breakdown effect was so strong that she failed to comprehend parts of lectures even though she had adequate metacognitive knowledge and skills.

> Test taker M: I tried to concentrate on listening but the noise was distracting. I did not take the note at the first place but I started to write what I heard. Once I start taking notes, I could concentrate more but I missed a lot.

Summary of research question 4. In the reflection session, questions were asked to collect test takers' opinion on areas of the study: the test, items, and cognitive process of second language listening. Findings showed that familiarity on test form reduced the anxiety but unfamiliarity on topics of listening passages affected test takers' performance. As one of quantitative analyses showed in Table 24, test takers also confirmed that topic familiarity could affect their performance in testing.

An unexpected order of test items could be a factor to influence test takers' performance. Participants mentioned item order hindered their comprehension. In the real world, a listener may not be able to retrieve an exact order of aural information. Unlike low proficiency listeners, higher proficiency listeners were able to overcome this factor.

Planning before taking the test could be helpful to the test takers. This factor was not significantly crucial to affect test takers' performance, but aided to overcome impediments.

Chapter 5

Discussion and Conclusions

Chapter Overview

This chapter discusses the findings of the study for each level of inference. The first level of inference addresses different aspects of validity and reliability and is based on the quantitative analysis. The iterative aspect of test specifications and low reliability coefficients are discussed in relation to how the study was planned and why the findings were not significant. The second level of inferences is derived from the breakdown effects observed. Possible reasons are proposed for the unintended effect that could represent universal samples. At the third level of inference, test takers' processes for listening comprehension test are discussed, and at the last level of inference, implications are suggested for test level, item level, and test takers' cognitive awareness.

Limitations of the study are explained, and these include design restrictions on measurement; certain limitations were also encountered when I created speech samples with breakdown factors. However, the study does point to future research investigating degrees of threshold for the breakdown factors and applying breakdown factors in other contexts where impediments could have very serious consequences (such as aviation English).

Findings and Discussions on Levels of Inferences

I do not have a crystal clear understanding of how second language learners process second language listening because they deal with many impediments to processing spoken language. Unlike other domains, listening comprehension can be disrupted by as the inherent time constraint, extraneous noise, unfamiliar accents, speech rate, and topic familiarity. The current study was conducted to answer questions about breakdown effects, validation of a test

with breakdown factors, and test takers' cognitive awareness of their listening comprehension and breakdown factors. To understand how the test takers process second language listening and what factors affect their comprehension, breakdown factors were created and implemented in a new listening test for this study. Most second language listening tests are "clean"; they introduce no impediments in the listening prompts to prevent listening comprehension breakdown in order to accurately measure test takers' listening proficiency; however, this test form is highly inauthentic, so test results may have weak validity. The current study introduced a test with breakdown factors specifically to measure their effects; I attempted to determine what kinds of impediments cause listening breakdown, what degree of impediments disrupt listening comprehension, the validity of a test with impediments, and the test takers' awareness of breakdown factors.

First level of inferences: Scoring. The research question for scoring inference investigated test takers' performance on tests with and without impediments. For this level of inference, test specifications, descriptive statistics of test scores, reliability analysis, and standard error of measurements were used to scrutinize participants' performance on the tests.

For the test development stage of this research, an iterative process was used in developing test specifications which would not only minimize confounding effects on the measurement but would also to enhance the study and test design. Because breakdown factors in a test are considered confounding variables from the perspective of measurement, the literature on listening comprehension was examined so as to avoid adding extraneous breakdown factors that would further confound test scores. This examination helped to identify types and degrees of breakdown factors suitable for our purposes. The iterative process for development of test

specifications meant that both the test blueprint and the items which it generated were created and critiqued by the researcher, the audio engineer, and several language testing experts.

From the many listening features identified in the literature as breakdown factors, three were selected for this study: British accent, fast speech rate, and background noise. Since these factors are known as possible impediments in listening comprehension, we suspected that they may cause test takers to perform less well than they expected. Hence, when these breakdown factors were sampled, possible scoring deflation and test takers' poor performance were expected. The accent was selected because most people have experienced difficulty to understand a particular accent. Learners don't usually receive education to familiarize them with other accents, so depending on one's experience and familiarity, any kind of accent could be a breakdown factor. For this study, the British accent was selected because it was one of Standard English but possibly not as familiar to some international students as the Midwest American accent.

During the test development, selection of an appropriate speech rate sample was challenging because of two issues: a) controlling the speech rate and b) optimum rate for the sample. It was very crucial to define these two issues before we selected the sample to be operationalized because they would potentially affect the scoring and validity of the test. During the recording session, the actor first attempted to read the scripts rapidly, but he was not able to produce a consistent rate of speech, mainly because of the length of the scripts. We next attempted to manipulate the speech (recorded at a "normal" rate) by audio editor software: *Audacity*. This technique proved much more manageable. We altered the speed of the audio by re-sampling and reduced the length of the selection to 90%, 85%, and 80% of the original. Three different samplings were used to verify their appropriateness, authenticity, and comprehensibility.

Our examination of the samples showed that the 90% length was appropriate for operationalization because the sound was more natural than the other two samplings.

Finally, three types of noise were tested: white, brown, and pink noise. White noise is a pure static noise with a flat frequency spectrum in linear space. Pink noise is a static noise that is decreased by 3dB per octave from the white noise spectrum, and Brown noise is a different kind of static noise that is decreased by 6dB per octave from the spectrum. Pink noise was selected mainly because it was more comfortable to listen to for a long period of time and we could thus rule out fatigue as a confounding variable. Then three types of pink noise were sampled, 5%, 15%, and 30%. 15% was deemed most appropriate for the study.

Although these breakdown factors were carefully designed, quantitative analysis did not clearly differentiate the test takers' performance in the four test groups. Unfortunately, the control group's scores were lower than those of the treatment groups though there was no group difference in TOEFL scores. The regular group's minimum score was the lowest among the four groups, and the standard deviation of the scores was larger than the other groups. The noise group also had a low mean, but their standard deviation was relatively smaller than other groups, and the score distributions of the British and speech rate groups were similar to each other.

Cronbach's alpha was computed to check the internal consistency of the test scores for examinees of the four groups. The reliability analysis indicated that the items on the test were not consistent measures of test takers' listening proficiency. In particular, the lowest coefficient value for the noise test indicated that the items were possibly measuring a different dimension (although this study did not detect what that dimension might be). The noise test's Confidence Interval for alpha was wider than other three tests. The three groups' width of Confidence Interval for alpha was similar. This inconsistency may indicate a breakdown effect, but the

evidence was not strong enough to support this interpretation because the control group's scores and the reliability coefficient value were also not good enough to compare with the noise group. With the Cronbach's alpha values for the tests, the standard error of measurement was calculated to estimate the error of interpretation for each individual's test score. The standard error of measurement for the regular and noise groups were larger than other two groups. If the error band is larger, then a person's true score may fall into a larger range, thus degrading the precision of the measurement. Though the same items were provided to all four groups, the scores for individuals from the regular group were less precise than the scores for those from the other test groups.

What can be said with some confidence is that the low reliabilities (overall) and the lowest reliability (of the noise group) point toward a detrimental effect of listening impediments on measurement quality. In short, impediments to listening reduce systematic variance. Whether they add a second dimension of measurement (e.g. the "ability to overcome impediments") is yet unknown. What is apparent is that these tests lowered reliability. And of all the tests, noise added the most noise.

With respect to Kane's model of validation, at this level of inference, I attempted to verify that the listening test was able to measure a listening construct with a breakdown factor. Though many factors were predicted and planned at the test development stage, the precision of scores was not accurate enough to claim that the construct was measured reliably.

Second level of inferences: Generalization. This research question was formulated to investigate the difference between test takers' performance and talk types, and possible breakdown effects that I could not detect from the previous analyses. To detect this phenomenon, two aspects were examined: test level and item level. At the test level, generalizability analysis

and the 4-by-4 Latin square design were used to find the differences between individual performance, talk types, and test groups. At the item level, classical item analyses were conducted to test the difficulty of items and item discrimination, factors that determine response differences between lower and upper performance groups.

First, variance components for person, items, and interaction between person and items were estimated using GENOVA. The generalizability analysis was conducted for all four tests, and the findings showed that person variance components were relatively small compared to the total variance components for all four tests. Larger person variance components are usually expected. On the other hand, item variance components were larger than the person variance components. Possible reasons for small variance components for person and items are first, that the test takers' scores were clustered with the mean of the tests and second, the low reliability of scores. The small number of test takers, items, and clustered mean could result in small variance components for person and items. The largest variance components were interaction between person and items. This indicates that the test was measuring multiple facets, and breakdown effects were captured in the residual variance components. It is challenging to detect which breakdown factor played a significant role as residuals, but the breakdown effect itself was detected. Due to small sample size and items, it was hard to separate three different breakdown factors in this study. Hence, it was difficult to detect the most significant listening impediment among these breakdown factors using GENOVA.

Second, the 4-by-4 Latin square design was used to detect which factor caused the breakdown effect in the test. The analysis of variance showed no significant difference in test takers' performance between the test types, but there was a difference in test performance between talk types. Among the four different talks, scores on the first long lecture had a

significantly larger mean than other talk types. The breakdown effect occurred in talk types, not in test types. So the intended breakdown factors did not play a significant role in the test, but their performances were affected by talk types. Between the two topics, test takers had more difficulty understanding the *dinosaur* and *global cooling*.

At the item level, the item difficulty and discrimination were investigated to compare function differences among the four tests. Some items were not performing well in four tests. The item difficulty indices for most items fell into a good range of difficulty, so the quality of items was not questioned except for the few easy items.

Item discrimination analysis was used to calculate discrimination power between the correct responses of the upper and lower groups. The discrimination indices for all four tests showed that discrimination powers were different across the four groups, even though the items were same. This indicated a possible breakdown effect at the item level. A couple of items were flagged as items with no discrimination power across the four test groups. Most items' discrimination indices fell into a good range of discrimination power, even though they functioned different across the groups.

At the test level, generalizability analysis indicated that breakdown effects exist as an interaction effect between person and item variance. This analysis could not detect which factor was causing breakdown effects, but the analysis of variance from the Latin square design showed that talk types caused the performance difference in the test. From these analyses, we could assume that topic familiarity is a stronger breakdown factor than the three intended breakdown factors. Since talk types functioned as a breakdown factor, the poor performance from the control group could be explained. For this group a clear version of the listening input was not an advantage.

At the item level, the item analyses showed that breakdown effect was not detected in the item difficulty indices, and most items functioned well in terms of difficulty and discrimination. Item difficulty indices were similar across the four groups, but discrimination powers functioned differently. The same items were provided to all four groups, so the different discrimination indices could be viewed as a breakdown effect, but the test takers' proficiency levels were not equally distributed across the four groups, so it is difficult to claim that the effect was solely due to the discrimination indices.

To sum up, I tried to determine the contributions of the test types and whether the test takers or test types contributed to the source of variance. Findings showed that talk types contributed more than test takers or test types to the source of variance. Test takers were apparently more affected by the talk types than the intended breakdown factors except for noise. Using a lengthy and unfamiliar topics as listening test input may not have been a good decision because the initial purpose of the study was to detect the breakdown effects. However, I found that this type of talk is a more severe breakdown factor than accent, speech rate or noise. Furthermore, item functioning was significantly different across the four test groups even though items were embedded in the spoken texts. From this finding, we can assume that some items were not influenced by the intended breakdown factors in the spoken texts even though items and texts were bound together. The intended breakdown factors did not hit the threshold of listening comprehension breakdown, so items were possibly unaffected by those factors.

Third level of inferences: Extrapolation. Evaluation of the extrapolation inference depends on the relationship between the universe of generalization and the target domain (Kane, 2006). In this study, test takers' cognitive awareness was a key to understanding the universe of generalization in second language listening comprehension and the breakdown effects in the

listening test. To understand the relationship between listening comprehension and breakdown effects, test takers were asked to provide responses on a cognitive questionnaire. The following four topics are discussed based on their responses on the questionnaire: self-awareness of their second language listening proficiency, listening process approach, use of metacognitive strategies, and use of listening test-taking strategies.

First, the response pattern of self-assessment was different across the test groups. A majority of test takers from the regular, speech rate, and noise groups perceived challenges to their second language listening. A Chi-Square test for independence also confirmed that this is an association between responses of test groups and self-assessment. On the other hand, test takers from the British group did not agree. They reported that second language listening was not challenging for them. Though the participants were randomly assigned to each group, more people with confidence in English happened to be assigned to the British group than to the other three groups. In addition, their perception on second language listening was closely related to their test performance. Performance of the British group was relatively higher than the performance of the other three groups. The evidence confirms that learners' self-awareness or perception affects their listening comprehension in testing situations. This phenomenon could stretch to real world settings. For instance, if learners perceive second language listening as challenging, their second language listening comprehension could be impeded when they encounter listening difficulties such as listening breakdown factors.

Furthermore, most test takers reported that they did not feel nervous when they were taking the test. The literature has shown that test anxiety affects test performance, and test takers' anxiety could increase if the test is high-stake. In this study, test takers' awareness of their listening proficiency influenced their test performance, not test anxiety. The main reason

they did not feel more anxiety was because participants were free volunteers and their test scores would not affect their academic life.

Test takers' various listening approaches were also recorded for this study. Participants were asked which approach they used and how effective it was for aiding their listening comprehension. The majority of the test takers from the four groups reported that they used a top-down listening approach. The breakdown factors and item types did not encourage test takers to switch from this approach to another such as bottom-up or parallel approaches. They did not pay attention to the details but tried to understand the context of the talks by using this approach. However, sticking with the top-down approach may create obstacles when answering detail questions which may require the bottom-up approach. Some test takers with higher scores used the top-down approach with efficient note-taking. They used this strategy to overcome the possible shortcomings of the top-down approach.

The literature shows that lower proficiency learners may rely more heavily on the bottom-up approach because they feel they must focus more on word and sentence recognition (Vandergrift, 1999). However, in this study, test takers with lower test scores also used the top-down approach because the length of the talk was too long to maintain using the bottom-up approach. The bottom-up approach requires more working memory to process the spoken text, but the load would be too heavy to process using word/sentence recognition during an almost 4-minute talk. Another possible reason these participants used only the top-down approach is that they had not learned how to use different listening approaches for different needs. Despite the effectiveness of the approach, they might use it more for its familiarity than its effectiveness.

The third topic addressed in the questionnaire was the test takers' metacognitive process of listening. To understand the test takers' listening comprehension in testing situations and real

world settings, it was important to investigate how they used the metacognitive process and what metacognitive strategies they deployed while listening. The majority of test takers from the four groups reported that they planned how to listen before they started the test. As a way of planning, they previewed the comprehension questions and tried to predict the content of the passage. The response rate varied, but most reported that they invoked their schema of background knowledge of the topic.

Along with planning and schema activation, monitoring is a distinctive feature of the metacognitive process. Monitoring sometimes helps to comprehend spoken texts because listeners adjust their interpretation based on what they monitored during the processing. The majority of respondents from the regular and British groups agreed that they monitor themselves periodically to check their comprehension, but the respondents from the speech rate and noise groups did not. Still, over 50% of all test takers agreed that they adjusted their interpretation if they realized inaccurate comprehension.

In this study, I found that most test takers used the metacognitive processes of planning and monitoring when they listened to the spoken texts, and they confirmed their effectiveness. This suggests that metacognitive processes should not be neglected in language test development. Students may engage them more on listening prompts if the test design allowed or encouraged them to do so. Commonly used listening comprehension tests do not provide test takers enough time to plan before they begin the test. However, I found that planning time should be provided so that test takers can activate their schema and engage their metacognitive skills because this step would eventually help them to comprehend second language listening better.

Test takers' responses on strategy uses were sometimes related to their listening process approach. Guessing activates test takers' schema and allows them to predict the whole meaning

by linking their background knowledge to what they do understand from the talks. Furthermore, strategies that the test takers used are commonly applied in real world settings. Most people understand the context even if they do not accurately understand a speaker's auditory information. If listeners have good background knowledge on a topic, this strategy can fully activate their schema. In testing, test takers could answer comprehension questions accurately even though they could not completely understand the talk if they have background knowledge of the topic. This study showed that most test takers from groups used a guessing strategy to overcome comprehension obstacles. Some respondents reported that they did not use a guessing strategy perhaps because they did not perceive comprehension obstacles or they were able to use another strategy. This speculation is based on their responses reporting that they used different strategies depends on the length of the talks. Some test takers from this group may able to use a complex strategy combined with guessing strategy depending on what they listen to. Which complex strategy they used was not revealed from their responses, but it is certain that the test takers who perceived differences in spoken texts used a combination of strategies that included guessing.

In both real world settings and testing situations, learners use metacognitive strategies to overcome comprehension breakdown, and each learner has a personal preferred strategy for overcoming obstacles in comprehension. More importantly, however, second language researchers have confirmed the effectiveness of metacognitive strategy use in improving second language listening (Vandergrift, 1999, O'Malley and Chamot, 1990). However, L2 listening assessments have not included the activities that could improve learners' second language listening comprehension. Metacognitive strategy embedded in test items could enhance the validity of the test and measure their L2 listening more accurately. For example, assessment

prompts could be divided into pre-listening, listening, and post-listening so that test takers could have enough time to engage and to use metacognitive strategy while they were listening to each listening prompts. Then, metacognitive strategy will help to activate test takers' listening process in order to comprehend better at listening section. The post-listening would the confirmation stage that their listening proficiency was measured properly.

Lastly, I investigated the kinds of strategies that test takers used during the test and why they used a particular strategy. Less than half of test takers took notes while they were listening to the talks. Note-taking was not considered an effective test-taking strategy. On the other hand, the majority of the respondents from all groups reported that background knowledge was crucial and helpful to answer the comprehension questions.

For this level of inference, evidence shows that test takers processed with metacognitive approaches such as planning and monitoring while they were taking the test. Planning and monitoring did activate their schema and enhance their listening comprehension. This evidence clearly indicates that metacognitive processes should be considered in testing. Without proper planning before the test, their listening schema would not be fully activated to help understanding what they hear in second language. Unlike planning, the ways that the test takers monitor their listening process were divergent across the groups. Monitoring could a key to success in listening comprehension, but if the test takers are constantly monitoring themselves, they might not be able to see the whole context of what they heard.

The test takers' metacognitive strategy uses are linked to monitoring, one of the metacognitive processes. Their strategy uses were varied across the test groups, however most test takers used metacognitive strategy, guessing to activate their schema to overcome the comprehension breakdown. Furthermore, the test takers who reported that they used guessing

strategy to overcome the breakdown, they also passively listened the talk when they could not understand the meaning of it. In a way of using strategy, they did not change their strategy whether they listened to short or long talks. On the other hand, the test takers who changed their strategies depends on the length of the talks, they did not use guessing much to overcome the comprehension breakdown. Instead, they used a different strategy to overcome the comprehension breakdown. To some test takers, the length of the talks could be a factor to encourage using various types of strategies to comprehend the lectures successfully.

Lastly, responses to our questionnaire showed differences in terms of test-taking strategy use. Most test takers did not take notes during the testing, but a few test takers did take notes to remember key words, and they believed the effectiveness of note taking. When the learners listen to a long talk, it may not be efficient and effective to take notes, but writing key words and points could help them to understand the talk. The test takers who planned how to listen before the test wrote down key words to focus on during the listening. Taking notes while listening may not be an effective strategy, but taking notes before listening could activate their schema for the topic and thus enhance their test performance. The majority of test takers believed that prior knowledge on the topic was very crucial to answer the comprehension questions, and they used their prior knowledge at the planning stage to answer some of the comprehension questions before listening. While listening they verified the correctness of their pre-listening choices. Not only does prior knowledge activate schema to overcome comprehension breakdown, it also functions as a test-taking strategy. From this study, note-taking during the test was not an effective way to take the test. If we allow test takers to take the note for planning purpose, then the note-taking strategy will be effective strategy to enhance listening comprehension.

Fourth level of inferences: Implication. The test takers' oral reflections on the test and other findings from the quantitative analyses were used to support the claim for research question four. The think aloud protocols were analyzed and categorized by topic: participants' opinions of test items and their cognitive awareness on the listening test. Findings from the quantitative analyses were also used to link and to support individuals' reflections on these topics.

Test level. At the test level, quantitative analyses failed to reveal a significant effect of the intended breakdown factors on test takers' performance, but topic familiarity was detected as an influential factor. Test takers were asked for their overall impressions of the test and the breakdown factors. They reported that they felt that the test format resembled the TOEFL, the test most widely administered to adult international students who want to test their English proficiency. Most test takers in this study are very familiar with the TOEFL because they have taken it several times to get an adequate score for admission to universities and colleges in the US. Due to their familiarity with the format of TOEFL, participants' impression of the current test was not bad because they had no difficulty adjusting to the format of items and spoken texts. This report is also reflected by questionnaire responses that they did not feel nervous while they were taking the test. To sum up, the familiarity of the test format reduced anxiety about the test, and we can conclude that the breakdown effect did not occur due to the format of the test.

Unlike the format of the test, the test takers' impression of the topics of the test varied depending on their degree of background knowledge of the topics. In turn, their degree of prior knowledge of the topics influenced their perception of the difficulty of the test. For instance, one test taker who was not familiar with the provided topics felt that his impression of the test and his awareness of his listening proficiency were affected by his lack of prior knowledge. In fact, topic familiarity not only influenced the test takers' impression and awareness of their listening

proficiency, it also emerged as a breakdown factor in listening comprehension. Indeed, the quantitative analysis confirmed that topic familiarity played a bigger role than the intended breakdown factors. It seems safe to conclude that if the test takers know enough about the topics to be confident to comprehend the content, they may not be affected by unfamiliar accents and fast speech rates. The implications for testing are clear: if testers used general topics for the spoken text input, they could add these breakdown factors to make the spoken texts more authentic, without fear of adding construct irrelevant variance.

Though the intended breakdown factors did not have a significant breakdown effect in the listening test, some test takers reported that background noise did function as an impediment, and their prior knowledge could not ease their comprehension. I also found cases, however, in which test takers learned how to manage the noise, but self-awareness of their performance did not correlate with their actual test performance. Although the quantitative analyses did not capture the breakdown effect of the noise, nevertheless test takers reported that they felt the noise was disturbing while they were listening to the talks.

The implications for test development are that some breakdown factors could be used in order to create authentic audio prompts, but that doing so runs the risk of lowering reliability. It is important to know the exact threshold of each breakdown factor so that it could make the spoken text more authentic without creating breakdown effects for most students and without substantially impacting measurement consistency. Further research is needed on these points. If I could craft this authentic spoken text, it would not be necessary to invest more time and money to create clean audio prompts, and I could use real lectures or conversations as audio prompts if these factors are clearly verified. In this study, British accent and speech rate were not perceived as breakdown factors. Audio prompts with British accent and fast speech rate could be used for

testing because these factors are not causing any breakdown effects. In other words, audio with British accent or fast speech rate could be used as listening prompts for testing. Background noise, however, may not be feasible to add to ordinary academic listening tests, but this factor could be suitable to use in tests for specific purposes such as aviation tests.

Item level. Classical item analyses showed that item difficulty and discrimination indices were varied across the four tests, indicating significant breakdown effects at the item level. A few items were marked as easy items, but most item difficulty and discrimination indices show that the quality of the items was good. Most items indices were at good range, but index values were different across the groups even though they had the same items. The comparison between item difficulty and discrimination indices showed various patterns among breakdown groups. Some items functioned not well in the British and speech rate groups that those items were not easy to higher proficiency test takers than those to lower proficiency test takers. Some items of the breakdown groups had no discriminating power regardless of item difficulty indices.

At this level of inference, test takers were asked to share their opinions on the items. Test takers did not specifically mention any particular item, but they agreed that the quality of items was good. And though they complained about the spoken texts, they did not complain about the quality of items. Instead of criticizing the quality of items, a couple of test takers mentioned the order of items in the test. They claimed that the order of items was not exactly the same as the flow of the spoken text so they had a difficult time managing the sequence of items.

Based on their feedback on item order, I found that this feature could affect test takers' cognitive awareness and listening process especially when they encounter new information because at the planning stage they expect a certain flow. The test takers planned how they were going to listen and answer the questions, but the item order mixed up their plan and their

performance was negatively affected. If I provide planning time before the test, the item order should follow the flow of the spoken text in order not to impede the participants' listening comprehension. If the test takers do not have time to plan, as in the current TOEFL format, item order may not have a significant effect on their performance. In real world settings, learners have to process an enormous amount of information, some well-ordered and some not. When information is not predictably ordered they face an unexpected situation which may induce a higher level of anxiety and comprehension breakdown. Unlike the testing situation, however, in the real world they can deploy clarification strategies to overcome the comprehension breakdown. Since the listening test cannot provide an interactive and communicative situation, the item order should be consistent with the text to avoid influencing test takers' listening process.

Cognitive awareness of second language listening. Test takers' cognitive awareness, their listening process, and their test taking strategy uses were investigated in the study. Findings of the quantitative analyses showed that cognitive awareness, metacognitive process, and strategy use varied across the test groups.

This study found a positive relationship between test takers' cognitive awareness of the test format and topic and their anxiety level. Some test takers reported that they felt challenges in English, but although their perception on English varied, their anxiety level was low during the test. However, the test takers who were familiar with the test format and test topics had lower anxiety levels and their cognitive awareness on the test was more positive than those who were not familiar with the test format and the topics.

The metacognitive processes of planning how to listen to the talks and monitoring their listening process were very important for activating their schema and correcting their interpretations while listening. Planning was crucial for the test takers, so this process should be

considered in the test development. However, the monitoring process was varied across the groups, so this process may not be necessary for all test takers, but it is helpful for some to overcome their comprehension breakdown. In this study, test takers' metacognitive process was investigated, but the process was not embedded in the test design. Metacognitive process embedded second language listening test could be developed and tested for its validity. Finding shows that planning is a crucial stage for second language listeners so providing simple information on each topic or previewing comprehension questions should be implemented to listening tests

In their responses to both the questionnaire and the self-reflection, participants reported effective metacognitive and test-taking strategy use. Higher proficiency test takers exhibited various metacognitive and test-taking strategy uses, and they were able to switch strategies depending on the length of the lectures. However, lower proficiency test takers tended to use strategies such as guessing or focusing on sentence segments. They did not apply complex strategies to overcome comprehension breakdown. According to feedback from the reflection session, when a higher proficiency student perceived a breakdown factor, she adapted to manage background noise during the listening. Initially, she could not comprehend the lecture due to the noise, but eventually she managed to control the noise.

From this study it is fair to conclude that higher proficiency test takers are able to manage their comprehension easily and employ metacognitive and test-taking strategies while lower proficiency test takers tend to use one or two metacognitive strategies such as guessing and predicting from sentences. Test takers without applicable strategies cannot manage their listening process when facing unexpected breakdown factors or item sequences, and therefore they do not overcome these impediments and their performance suffers. To improve listening

comprehension, it is very important to enhance the metacognitive process and strategy use. This study strongly indicates that I should develop items with more attention to metacognitive processes and that items embedded with metacognitive processes could measure second language listening more accurately.

Issues of reliability and validity in this test. A question was raised whether the test is valid since the tests had low reliability especially the noise test. Some researchers argue that reliability is a precondition of validity so a test with low reliability cannot be a valid test. A test with low reliability indicates high measurement errors that reflect a gap between test takers' actual performance and the test scores. The relationship between reliability and validity is controversial among researchers. Some researchers believe that reliability is a necessary but not sufficient condition for validity. Other claims that reliability is not necessary for validity. Furthermore, these concepts describe different aspect of attributes.

To review validity theorists' concept of validity, Messick (1989) claimed that the value of assessment is based on evidence or consequence of the test. Moreover, the test scores provide information of interpretation or use of assessment. Kane (2006) used the term validation instead of validity to include the development of evidence to support the proposed interpretations and uses, and to correspondence with an evaluation of the proposed interpretations and uses whether those are plausible and appropriate. Borsboom and et. al (2004) claimed that epistemological issues are related to validation and consequential issues are central to test use, and these should not be relevant to the concept of validity itself.

Based on the stance of validity theorists, reliability is a requirement for validity when reliability is defined as consistency among independent measures intended as interchangeable. Mislevy (2004) emphasized that reliability is a credibility of evidence; validity cannot be exist

without reliability. He claimed that the internal consistency reliability indices aid to evaluate the quality of evidence. However, Moss suggested that reliability could be an option rather than a requirement if the purpose of assessment is to improve teaching and learning. And inconsistency in test takers' performance does not invalidate the test but it becomes an empirical puzzle to be solved by search for more comprehensive interpretation.

To borrow the definition of validation from Kane and Borsboom, the main purpose of this study was not to apply (a strict definition of) construct validity of the second language listening test. Rather, this study focused more on the development of evidence, the process of evaluation of the test and test takers' listening process. Internal consistency of all tests was not acceptable however it is not necessary to judge this test is not a valid test. Quantitative findings were not sufficient to claim the generalizability of the test, but qualitative findings supported the test takers' listening process and their evaluation on the test.

Inconsistency of test takers' performance was somewhat expected because I included impediments that other researchers claimed possible listening comprehension hindrance. Unexpected factor, talk type, was found from the evaluation of inferences that weaken my validation claims, still evidence on listening process approach and breakdown effect on both noise and talk type made the test meaningful.

That said, and as noted before, future researchers and test developers can view these findings as a caution: impediments to listening tests can indeed lower reliability. The real issue here is not whether or not that will happen, but whether or not the test developer expects it to happen and the degree to which the test developer has alternate sources of validity evidence to counteract it.

Limitations and Suggestions for Future Research

The purposes of this study were to investigate breakdown effects in the listening test and to measure students' listening proficiency. Two limitations were shown during the study. First, breakdown effects were not captured as I expected because the speech samples were not distinctive enough to measure the breakdown effects. In developing the breakdown factors, one actor recorded both American and British accent to avoid a speaker effect. However, when the actor recorded the talk with a British accent, he had some difficulty in consistently producing a native-like accent in a long talk so the speed of the speech was a bit slow. This issue was a limitation of the British accent speech sample. According to a native British speaker, if the actor could speak faster or increase speech rate, the speech sample would be appropriate (personal conversation with a testing colleague at Pearson). When I made the speech rate speech sample, the actor first tried to simply speak fast, but I could not get a consistent speech rate so I decided to use the software editor to control the rate. Thus, although I could get same speech rate for four talks, the speech rate sample might sound a bit artificial. These reasons could have influenced the insignificant breakdown effects in the test.

To productively employ breakdown factors in listening tests, the degree of breakdown of each factor should be tested to verify its effectiveness as a valid test feature for listening comprehension (including, quite obviously, ongoing monitoring of its impact on test reliability). In this study, only noise (15% of pink noise) was marked as a strong breakdown factor. This indicates that if I add 15% of pink noise to the listening prompts, I will impede test takers' listening comprehension. Testing with different percentages of pink noise would be helpful to determine the real threshold of comprehension. In this study, brown and white noises were not used due to applicability. However these noises could be feasible in some other context. On the

other hand, British accent and 90% fast speech rate could be added to make listening prompts more authentic because these factors did not impede comprehension. Research on different accents and various speech rates should be conducted to find helpful information for constructing authentic listening prompts.

The second limitation of the study would be assigning participants to each group. Each participant was randomly assigned to each group. Pre-screening should be conducted before the main test administration to compare pre-test scores and post test scores. This comparison could distinguish the group differences.

Several suggestions for future research in test development can be derived from the findings of this study. First, several findings show that topic familiarity affects test takers' listening comprehension; in fact, it was a more influential factor in comprehension breakdown than the intended breakdown factors. Thus, a test using general topics could more be appropriate for measuring breakdown effects. For instance, using simple conversations concerning general topics as listening prompts might help to more accurately measure breakdown effects from other factors.

Second, the metacognitive process should be applied in second language test development. Dominant listening tests do not consider metacognitive listening processes or metacognitive strategy use in listening test development. This study included a cognitive questionnaire to understand test takers' perspectives on listening comprehension and their cognitive processes on the listening test, and it revealed the effectiveness of metacognitive processes and strategy uses. More research, however, is needed to understand test takers' listening processes more deeply, but it is clear that a metacognitive process-embedded language

test could be developed based on what I understand from this study. The reliability and validity of such a test may surpass that of current test forms.

A third implication of this study is the need for developing a recording manual for future researchers who want to make speech samples for listening tests. This researcher worked with an audio recording engineer and an actor, but realized that recording the listening prompts was not an easy task because of all the factors to consider and control such as text types, sound, and audio manipulation skills. It was necessary to provide explicit guidance to the actor and communicate with the audio engineer record a feasible speech sample for this listening comprehension test. A training manual which addresses actor training, recording and editing of the listening prompts, and adding additional listening factors to the recordings would be helpful for future researchers.

A fourth implication of this study is the listening test development of various degrees of breakdown factors and breakdown research at the item level. In this study, three listening impediments were used to investigate breakdown effects in the listening test, and noise was the most significant breakdown factor at the test level. Research on various noise levels or accent types could contribute to authentic listening task development. At the item level, the breakdown effect was detected across all breakdown test groups. Research on item types, orders and the relationship between item statistics and breakdown effects could help to define more effective authentic listening test. In particular, items that require different listening process approaches might affect test takers' listening comprehension if a certain type of impediments is added in the test.

A fifth implication of this study is to investigate human factors affecting second language listening proficiency. Human behavior, perception and cognition could be vary and influential to

their second language listening capacity. Research on human factors could be helpful to understand unknown area of second language listening process. It will eventually help for test developers to create valid and reliable listening tests.

My final suggestion for future study concerns English tests for academic and specific purposes using noise. For English tests for academic purposes, audios from real classroom could be used as listening test. Researchers could investigate listeners' listening comprehension in authentic noise setting. For English tests for specific purposes, research on authentic and high stake testing would be needed as aviation listening. For this type of test it would be feasible to apply breakdown factors in the listening prompts that are relevant to the special purpose. Language test developers have begun to look into aviation English testing due to the unique demands of the job and the requirements set by the ICAO (ICAO, 2007). In order to be as authentic as possible and aviation listening test really should employ breakdown factors. International pilots and controllers might need to be familiar with other English accents, and they need to manage constant ambient noise. Though they use coded languages and radiotelephony, accurate listening comprehension is crucial in order to communicate effectively during flight operations such as take-off, landing, and emergency management. Listening tests for aviators that use breakdown factors would be ideal to develop very authentic listening prompts and to reliably measure their listening proficiency.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

AREA, (1999). American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Bachman, L. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L. (2004). *Statistical Analyses for Language Assessment.* Cambridge University Press.

Bonk, W. (2000). Testing ESL learners' knowledge of collocations. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 113-142). Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center.

Borsboom, D., Mellenbergh, J. & Heerden, J.V. (2004). The concept of validity. *Psychological Review*, 111 (4), 1061-1071.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics, 18*, 171-191.

Brown, A. L. (1985). *Reciprocal teaching of comprehension strategies: A natural history of one program for enhancing learning* (Tech. Rep. No. 334). Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Brown, G. & Yule, G. (1983). *Discourse analysis*. Cambridge, England: Cambridge University Press.

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

Cadierno, T. (1999). *L2 listening comprhension. Odense Working Papers in Language and Communication* (WP-18). Denmark, The Netherlands: Odense University, Institute of Language and Communication.

Call, M. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly, 19*, 765-81.

Chang, A.C. & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly, 40*(2), 375-397.

Cohen, A.D. (2006). The coming of age of research on test-taking strategy. *Language Assessment Quarterly, 3*(4), 307-331.

Conrad, (1983). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition, 7*, 59–69.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity of clinical procedures. *Candian Journal of Psychology, 13*, 102–128.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational Measurement* (2nd Ed.). Washington, D. C.: American Council on Education.

Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Eduation.

Davidson, F. & Lynch, B. (2002). *Testcraft*. New Haven and London: Yale University Press.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, *16*, 640-647.

Feyten, C.M. (1991). The power of listening ability: An overlooked dimension in language acquisition. *The Modern Language Journal, 75*(2), 173-180.

Flowerdew, J. (1994). *Academic listening: research perspectives.* Cambridge, England: Cambridge University Press.

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

Flowerdew, J. & Miller, L. (2005). *Second Language Listening: Theory and Practice.* Cambridge, England: Cambridge University Press.

Field, J. (2001). Finding one's way in the fog: Listening strategies and second language learners. *Modern English Teacher, 9*(1), 29-34.

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

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20* (2), 221 – 236.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, England: Routledge.

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

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

Greene, J. (2007). *Mixed methods in social inquiry.* San Francisco, CA: Jossey-Bass.

Henning, G. (1991). *A study of the effects of variation of short-term memroy load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance* (RR-90-18). Princeton, NJ: Educational Testing Service.

ICAO PRICE Study Group. (2007). ICAO Policy on Language Proficiency Testing, ICAO, Montreal, CANADA.

Kane, M. (1990). An argument-based approach to validation. Iowa City, IA: ACT.

Kane, M (2006).Validation. In R.L.Brennan (Ed.), *Educational measurement* (pp.18-64). Westport, CT: Praeger.

Kunnan, A. (1998). *Validation in language assessment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Li, J. (2006). *Introducing audit trails to the world of language testing* (Unpublished master thesis). University of Illinois, Urbana-Champaign.

Long, M. (2005). *Second language needs analysis.* New York, NY: Cambridge University Press.

Long, D. (1989). Second language listening comprehension: A schema theoretic perspective. *Modern Language Journal*, *73*, 32-40.

Lumley, T. & O'Sullivan, B. (2005). The impact of test taker characteristics on speaking test task performance. *Language Testing, 22*(4), 415-437.

Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics, 18*, 3-19.

Mendelsohn, D. J. (1998). Teaching listening. *Annual Review of Applied Linguistics, 18*, 81-101.

Mesbah, H. M. (2006). The impact of linear versus nonlinear listening to radio news on recall and comprehension. *Journal of Radio Studies, 13*(2), 187-200.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-104). New York, NY: Macmillan.

McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.

Mislevy, R. J. (2004). Can there be reliability without "reliability?" *Journal of Eduational and Behavioral Statistics, 29*(2), 241-244.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Reseracher, 23*(2), 5-12.

Moss, P. A. (2004). The meaning and consequences of "reliability." *Journal of Eduational and Behavioral Statistics, 29*(2), 245-249.

O' Mally, J. M., Chamot, A. & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*, 418-437.

O'Mally, J., & Chamot, A. (1990). *Learning strategies in second language acquisition.* New York, NY: Cambridge University Press.

Potter, R. & Choi, J. (2006). The effect of auditory structual complexity on attitudes, attention, arousal, and memory. *Media Psychology, 8*, 395-419.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Rost, M. (2002). *Teaching and researching listening*. Harlow, England: Longman.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal, 78*(2), 199-221.

Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal, 78*(2), 179-189.

Shang, H. (2005). An investigation of cognitive operations on L2 listening comprehension performance: An exploratory study. *International Journal of Listening,* 51-62.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory*: A Primer. Newbury Park, CA: Sage.

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

Tauroza, S. & Allison, D. (1990). Speech rates in British english. *Applied Linguistics, 11*(1), 90-105.

Ur, P. (1984). *Teaching listening comprehension*. New York, NY: Cambridge University Press.

Wagner, E. (2002). Video listening tests: A pilot study. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics,*

2(1). http://www.tc.edu/tesolalwebjournal/wagner.pdf

Vandergrift, L. (1996). Listening strategies of Core French high school students. *Canadian Modern Language Review*, *52*(2), 20-223.

Vandergrift, L (1999). Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal*, *54*, 168-176.

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

Young, M. (1997). A serial ordering of listening comprehension strategies used by advanced ESL learners in Hong Kong. *Asian Journal of English Language Teaching, 7*, 35-53.

Appendix A

Test Specifications

Academic listening comprehension test version 0.2

1. General description

The listening comprehension test is composed of two sections: one with regular listening tasks and one with breakdown listening tasks. The purpose of including breakdown tasks is to investigate impeding listening factors and to find listeners' strategies which ease their comprehension. The test measures test takers' strategic listening skills to comprehend explicit and implicit spoken information. Topics of the test will be general academic topics. We aim to measure pragmatic and sociolinguistic knowledge which include their understanding of the function or longer text, of sociocultural language setting and of academic setting.

2. Prompt Attributes

Test takers will hear approximately 2-3 minute long talks, discussion, or debates. After each talk, they will be given three to five listening comprehension questions. Each test taker can control the audio text by pause function. However the talk will be spoken only once and they will not be printed on the screen.

1) Spoken text attributes

a. Regular text

Spoken text attributes include four important features: text type, setting, and functions of speakers, and topic.

- Text type: whether the text is academic lecture, debate, discussion, report, or presentation
- Setting: where listening takes place.
- Functions of the speakers are as follows:
  a) Give instructions
  b) Describe/Define/Compare/Summarize
  c) Explain/Debate
- Other features of spoken text include:
  a) Length of oral text: each talk is approx. 2-3 minute long
  b) Speech rate – approx. 150 words per minute
  c) Variety of accent – standard target language
  d) Number of speakers – 1
  e) Gender – Male or Female

b. Breakdown text

- Text type: whether the text is academic lecture, report, or presentation
- Setting: where listening takes place.
- Functions of the speakers are as follows:
  a) Describe or explain
  b) Define/Compare/Summarize
- Other features of spoken text include:
  f) Length of oral text: each talk is approx. 2-3 minute long
  g) Speech rate – varies depends on text type

h) Variety of accent – standard target language
i) Number of speakers – 1~5
j) Gender – Male or Female

** Breakdown (impeding) factors

The breakdown task may contain one or more impeding listening factors. The task may have usual speech rate of speakers, background noises, or excessive number of same gender speakers.

3. Question Attributes

Several types of questions will be asked for each talk:

a. recalling (identifying a fact).
b. reasoning (using logic to draw conclusions from available information).
c. problem-solving (recognizing a problem).
d. decision-making (evaluating and choosing the best solution).
e. Analysis (examining parts of a whole and their relationships, distinguish, examine, determine the cause and effect, explain the main idea)
f. Synthesis (putting parts back together to create a new whole, develop a plan, or communicate a new way)
g. Evaluation (making a judgment using a specific set of criteria)

4. Distracter attributes

Several types of distracters will appear. Those distracters are constructed using related word, repeated word, incorrect tense or inference, or same word with a different meaning.

5. Response Attributes

Test takers will choose an answer out of four choices.

6. Test form

Sections	Number of tasks	Number of items
Regular tasks	5	
Section 1: Conversation	3	7-10
Section 2: Lecture	2	
Breakdown tasks	5	
Section 1: Conversation	3	7-10
Section 2: Lecture	2	

Academic listening comprehension test version 0.5

1. General description

The listening comprehension test will be provided in four different formats. Though spoken texts and comprehension items are identical, each test will be provided with three different impediments and without an impediment.

The purpose of including breakdown factors is to investigate the function of the impeding listening factors, to find listeners' strategies, and to understand their listening process. The test measures test takers' strategic listening skills to comprehend explicit and implicit spoken information. Topics of the test will be general academic topics.

2. Prompt Attributes

Test takers will hear approximately 2-3 minute long talks and discussion. After each talk, they will be given three to five listening comprehension questions. Each test taker can control the audio text by pause function. However the talk will be spoken only once and they will not be printed on the screen.

1) Spoken text attributes

a. Regular text

Spoken text attributes include four important features: text type, setting, and functions of speakers, and topic.

- Text type: whether the text is academic lecture, discussion, report, or presentation
- Setting: where listening takes place.
- Functions of the speakers are as follows:
  d) Give instructions
  e) Describe/Define/Compare/Summarize
  f) Explain/Debate
- Other features of spoken text include:
  k) Length of oral text: each talk is approx. 2-3 minute long
  l) Speech rate – approx. 150 words per minute
  m) Variety of accent – standard target language
  n) Number of speakers – 1
  o) Gender – Male or Female

b. Breakdown text

- Text type: whether the text is academic lecture, report, or presentation
- Setting: where listening takes place.
- Functions of the speakers are as follows:
  a) Describe or explain
  b) Define/Compare/Summarize
- Other features of spoken text include:
  p) Length of oral text: each talk is approx. 2-3 minute long
  q) Speech rate – varies depends on text type
  r) Variety of accent – standard target language
  s) Number of speakers – 1~5
  t) Gender – Male or Female

** Breakdown (impeding) factors
The breakdown task may contain one or more impeding listening factors. The task may have usual speech rate of speakers, background noises, or excessive number of same gender speakers.

3. Question Attributes
Several types of questions will be asked for each talk:

h. recalling (identifying a fact).
i. reasoning (using logic to draw conclusions from available information).
j. problem-solving (recognizing a problem).
k. decision-making (evaluating and choosing the best solution).
l. Analysis (examining parts of a whole and their relationships, distinguish, examine, determine the cause and effect, explain the main idea)
m. Synthesis (putting parts back together to create a new whole, develop a plan, or communicate a new way)
n. Evaluation (making a judgment using a specific set of criteria)

4. Distracter attributes
Several types of distracters will appear. Those distracters are constructed using related word, repeated word, incorrect tense or inference, or same word with a different meaning.

5. Response Attributes
Test takers will choose an answer out of four choices.

6. Test form

Sections	Number of tasks	Number of items
Regular tasks	5	
Section 1: Discussion	3	7-10
Section 2: Lecture	2	
Breakdown tasks	5	
Section 1: Discussion	3	7-10
Section 2: Lecture	2	

Academic listening comprehension test version 0.8

1. General description

The listening comprehension test will be provided in four different formats. Though spoken texts and comprehension items are identical, each test will be provided with three different impediments and without an impediment.

The purpose of including breakdown factors is to investigate the function of the impeding listening factors, to find listeners' strategies, and to understand their listening process. The test measures test takers' strategic listening skills to comprehend explicit and implicit spoken information. Topics of the test will be general academic topics.

2. Prompt Attributes

Test takers will hear approximately 1-3 minute short and long talks. After each talk, they will be given five listening comprehension questions. The talk will be spoken only once and they will not be printed on the screen.

1) Spoken text attributes

a. Regular text

Spoken text attributes include four important features: text type, setting, and functions of speakers, and topic.

- Text type: whether the text is academic lecture or presentation
- Setting: classroom
- Functions of the speakers are as follows:
  a) Give instructions
  b) Describe/Define/Compare/Summarize

- Other features of spoken text include:
  a) Length of oral text: each talk is approx. 1-3 minute long
  b) Speech rate – approx. 150 words per minute
  c) Variety of accent – standard target language
  d) Number of speakers – 1
  e) Gender – Male

b. Breakdown text

- Text type: whether the text is academic lecture or presentation
- Setting: classroom
- Functions of the speakers are as follows:
  a) Give instructions
  b) Describe/Define/Compare/Summarize
- Other features of spoken text include:
  f) Length of oral text: each talk is approx. 2-3 minute long
  g) Speech rate – 85 or 90% shrink the original speech spectrum
  h) Variety of accent – British accent
  i) Number of speakers – 1

j) Gender – Male

** Breakdown (impeding) factors

The breakdown task may contain one or more impeding listening factors. The task may have usual speech rate of speakers, background noises, or excessive number of same gender speakers.

3. Question Attributes

Several types of questions will be asked for each talk:

a. recalling (identifying a fact).
b. reasoning (using logic to draw conclusions from available information).
c. problem-solving (recognizing a problem).

4. Distracter attributes

Several types of distracters will appear. Those distracters are constructed using related word, repeated word, incorrect tense or inference, or same word with a different meaning.

5. Response Attributes

Test takers will choose an answer out of three choices.

6. Test form

Test	Number of items
Regular Test Section 1: Presentation (Short talk) Section 2: Lecture (Long talk)	20
Breakdown Test 1 Section 1: Presentation Section 2: Lecture	20
Breakdown Test 2 Section 1: Presentation Section 2: Lecture	20
Breakdown Test 3 Section 1: Presentation Section 2: Lecture	20

Academic listening comprehension test version 1.0

1. General description

The listening comprehension test will be provided in four different formats. Though spoken texts and comprehension items are identical, each test will be provided with three different impediments and without an impediment.

The purpose of including breakdown factors is to investigate the effect of the impeding listening factors, to find listeners' strategies, and to understand their listening process. The test measures test takers' strategic listening skills to comprehend explicit and implicit spoken information.

Topics of the test will be general topics: dinosaurs and weather change

2. Prompt Attributes

Test takers will hear approximately 2-3 minute long talks and presentation. After each talk, they will be given five listening comprehension questions. Each test taker can control the audio text but each talk will be spoken only once and they will not be printed on the screen.

1) Spoken text attributes

Spoken text attributes include four important features: text type, setting, and functions of speakers, and topic.

- Text type: academic lecture and presentation
- Setting: classroom
- Functions of the speakers are as follows:
  - Describe/Define/Compare/Summarize
  - Explain/Debate
- Other features of spoken text include:
  - Length of oral text: each talk is approx. 2-3 minute long
  - Speech rate – 90% reduced from the original spectrum (85% reduced is also available)
  - Variety of accent – standard target language and British accent
  - Number of speakers – 1
  - Gender – Male
  - Background noise – 15% of pink noise added

** Breakdown (impeding) factors

Each breakdown test contains one impeding listening factors. The task may have usual fast speech rate of speakers, background noises, or British accent

3. Question Attributes

Several types of questions will be asked for each talk:

d. recalling (identifying a fact).
e. reasoning (using logic to draw conclusions from available information).
f. problem-solving (recognizing a problem).
g. decision-making (evaluating and choosing the best solution).

h. Analysis (examining parts of a whole and their relationships, distinguish, examine, determine the cause and effect, explain the main idea)

4. Distracter attributes

Several types of distracters will appear. Those distracters are constructed using related word, repeated word, incorrect tense or inference, or same word with a different meaning.

5. Response Attributes

Test takers will choose an answer out of three choices.

6. Test form

Test	Number of items
Regular Test Section 1: Presentation-topic 1 (Short talk) Section 2: Lecture –topic 1 (Long talk) Section 3: Presentation –topic 2 (short talk) Section 4: Lecture –topic 2 (Long talk)	20
Breakdown Test 1 (British accent) Section 1: Presentation-topic 1 (Short talk) Section 2: Lecture –topic 1 (Long talk) Section 3: Presentation –topic 2 (short talk) Section 4: Lecture –topic 2 (Long talk)	20
Breakdown Test 2 (Speech rate) Section 1: Presentation-topic 1 (Short talk) Section 2: Lecture –topic 1 (Long talk) Section 3: Presentation –topic 2 (short talk) Section 4: Lecture –topic 2 (Long talk)	20
Breakdown Test 3 (Noise) Section 1: Presentation-topic 1 (Short talk) Section 2: Lecture –topic 1 (Long talk) Section 3: Presentation –topic 2 (short talk) Section 4: Lecture –topic 2 (Long talk)	20

Appendix B

Needs Analysis

Listening Quizzes for Academic Purposes http://www.esl-lab.com/

Table 25

MIT Free Online Lecture 1

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Moderate	Medium	Scuba	Normal	One	Male	medium

Note: clear voice, speed, authentic (real lecture in the class), introduction of the class
http://ocw.mit.edu/OcwWeb/Athletics--Physical-Education-and-Recreation/PE-210Spring-2007/LectureNotes/index.htm

Table 26

MIT Free Online Lecture 2

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy Moderate Advanced		Biology				

http://ocw.mit.edu/OcwWeb/Biology/7-012Fall-2004/VideoLectures/index.htm

Table 27

BBC Learning English Radio 1

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy Moderate Advanced	British English	Cryptozoology	Moderate	2	Male & Female	advanced

Note: conversation, inauthentic (recorded based on a script)
http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/144_6minute/page3.shtml

Table 28

BBC Learning English Radio 2

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy Moderate Advanced	British English	Pygmy hippo	Medium	2	Male/Female	Moderate

Note: conversation, inauthentic (recorded based on a script)
http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/144_6minute/page10.shtml

Table 29

Princeton University Lecture

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy Moderate Advanced	American English	Human Right	Medium	5	Male	Moderate

Note: recording quality (so-so), forum (presentation), authentic
http://www.princeton.edu/WebMedia/lectures/
May 12, 2008 - Part 2: "Human Rights: Are They Universal? Where Do They Come From?"

Table 30

China Now Lecture Series: Can a Green Dragon Fly? China's Energy Challenges and Opportunities

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy			Slow enough to understand	1	Male	
Moderate Advanced	British English	China				Moderate

http://www3.imperial.ac.uk/media/onlinelectures

Table 31

Prions: A New Principle of Disease

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy			Slow enough to understand	2	Male/Male	
Moderate	British English	Neurology				
Advanced						Advanced

http://www3.imperial.ac.uk/media/onlinelectures

Table 32

Debate 1: The Seat

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy	American English		Slow enough to understand	2	Female/Male	Easy to understand
Moderate		The seat debate				
Advanced						

Note: Female speaker is from Japan but doesn't have any accents. Easy to understand and speech rate is slow enough.
http://www.elllo.org/english/0901/T919-Debate-BackRow.htm

Table 33

Debate 2: Fast Food

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy	American English/Singapore English		Slow enough to understand	2	Female/Male	
Moderate		Fast Food				Moderate
Advanced						

Note: Sounds like they are speaking based on scripts but authentic setting. Speaker is using practical vocabularies or phrases such as veggies, on the down side, and burger joint.

http://www.elllo.org/english/0901/T906-Clare-Fastfood.htm

Table 34

Discussion: Accident

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy Moderate Advanced	Some of them has an accent	Accident	Moderate to understand	6	Female/ Male	Moderate

Note: Six people are sharing their experience on car accident. Different accents are noticed. It could be difficult for listeners who are not familiar with car related vocabularies. Speech rate was moderate.

Table 35

News: Men More Attracted to Women in Red

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy Moderate Advanced	British English	News	Moderate to understand	1	Male	Advanced

Note: "Men more attracted to women in red" Vocabularies might be hard for lower capacity students. Speech rate was slow but the content could be difficult to understand.
http://www.breakingnewsenglish.com/0810/081030-colour.html

Table 36

Voanews: American History

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy					Males	
Moderate	American English	American History	Moderate to understand	2		
Advanced						Advanced

Note: Background knowledge might be affected to understand the audio text. Speech rate is moderate and voices are clear enough to understand. Vocabulary level is moderate. Time of audio text is around 3 minutes.
http://www.voanews.com/specialenglish/2008-10-29-voa4.cfm

Table 37

News: Foreign Student Series- Getting a US Education from Home

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy					Male	
Moderate	American English	US education	Moderate to understand	1		
Advanced						Advanced

Note: Speech rate is slow but overall speech takes 4 minutes.
http://www.voanews.com/specialenglish/2008-10-29-voa3.cfm

Table 38

Lecture: Dinosaurs

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy					Male	
Moderate		Dinosaur Mystery	Moderate to understand	1		
Advanced	British English					Advanced

Note: It is difficult to understand his English due to accent. Topic is about dinosaur adventure and linguistic difficulty is upper intermediate. It has video clip but listening difficulty would be higher if we eliminate the video clip and just use audio clip. Dinosaurs make sounds and background music could make more difficult to understand even though the speaker did not say a word. The duration of clip is 8 minutes.
http://www.youtube.com/watch?v=kYRU-dYcGDw&feature=related
http://kibishipaul.com/blog1/category/level/advanced/

Table 39

Lecture: Twister

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy					Males	
Moderate	American English/British English	Twister	Moderate to understand	4		Moderate
Advanced						

Notes: It is authentic and you can hear several voices. Vocabulary level is intermediate. It would be a good listening task without video clip. One speaker is using Southern accent.
http://revver.com/video/444279/twister-tornado-alert/
http://kibishipaul.com/blog1/2008/02/20/lesson-65-tornado-alley/

Table 40

Halloween Solar Storm Anniversary

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy					Females	
Moderate	American English	Solar system	Moderate to understand	2		
Advanced						Advanced

Note: Clear voice and moderate speech rate. Topic is very interesting and vocabulary level is moderate.
http://www.nasa.gov/multimedia/nasatv/on_demand_video.html?param=http://anon.nasa-global.edgesuite.net/anon.nasa-global/ccvideos/Halloween_solar.asx|Halloween Solar Storms Anniversary|285060main_aurora_storm_100.jpg&_id=undefined&_title=174161&_tnimage=test.gif

Table 41

Saturn's Cyclones

Difficulty	Accent	Topic	Speech rate	# of speakers	Male/female	Linguistic difficulty
Easy					Male	
Moderate	American English	Saturn's Cyclones	Moderate to understand	1		
Advanced						Advanced

Note: Speech is clear and rate is moderate but some words are spoken rapidly. Appropriate length of audio text and it could be good listening task to use.
http://www.nasa.gov/multimedia/videogallery/index.html

Appendix C

Test Form

Presentation 1 (331 words)

Source: http://www.enchantedlearning.com/subjects/dinosaurs/

Today, I am going to present about dinosaurs. Dinosaurs were one of several kinds of prehistoric reptiles that lived during the Mesozoic Era, the "Age of Reptiles." The largest dinosaurs were over 100 feet (30 m) long and up to 50 feet tall like Argentinosaurus, Seismosaurus, Ultrasauros, Brachiosaurus, and Supersaurus. The smallest dinosaurs, like Compsognathus, were about the size of a chicken. Most dinosaurs were in-between. It's very difficult to figure out how the dinosaurs sounded, how they behaved, how they mated, what color they were, or even how to tell whether a fossil was male or female.

Some walked on two legs (they were bipedal), some walked on four (they were quadrupedal). Some could do both. Some were speedy like Velociraptor, and some were slow and lumbering like Ankylosaurus. Some were armor-plated, some had horns, crests, spikes, or frills. Some had thick, bumpy skin, and some even had primitive feathers.

The dinosaurs dominated the Earth for over 165 million years during the Mesozoic Era, but mysteriously went extinct 65 million years ago. Paleontologists study their fossil remains to learn about the amazing prehistoric world of dinosaurs. The dinosaurs went extinct about 65 million years ago, at the end of the Cretaceous period, which was a time of high volcanic and tectonic activity. There are a lot of theories why the extinction occurred. The most widely accepted theory is that an asteroid impact caused major climatic changes to which the dinosaurs couldn't adopt.

Dinosaurs probably live on today as the birds. All that's left of the dinosaurs are fossils and the birds. Dinosaur fossils can be found all over the world. Although dinosaurs' fossils have been known since at least 1818, the term dinosaur (deinos means terrifying; sauros means lizard) was coined by the English anatomist Sir Richard Owen in 1842. The only three dinosaurs known at the time were Megalosaurus, Iguanodon, and Hylaeosaurus, very large dinosaurs. The oldest known dinosaur is Eoraptor, a meat-eater from about 228 million years ago.

Q1. Dinosaurs lived during the Mesozoic Era. The Mesozoic Era is also known as…(recall question)

a) Age of ultrasauros
b) Age of Deinos and sauros
c) Age of Reptiles*

Q2. Many theories have explained the extinction of Dinosaurs. What is the most supportive theory? (Recognize the problem)

a) Due to volcano eruptions
b) Due to an asteroid impact*
c) Due to the ice age

Q3. Tyrannosaurs are bipedal dinosaurs, Brachiosaurus are _____ dinosaurs. (reasoning)

a) Quadrupedal *
b) Hylaeosaurus
c) Cretaceous

Q4. Which statement is true? (evaluation)

a) Compsognathus is 100 feet long and up to 50 feet tall
b) Supersaurus is 30 m long and 15 m tall*
c) Velociraptor is slow and lumbering

Q5. What can we learn from this lecture? (Synthesis)

a) All dinosaurs are now remained as birds
b) Dinosaurs lived in different eras and they had different shapes though they were born from eggs
c) It's very difficult to figure out dinosaurs' sound, or gender (whether they are females or males) through fossils *

Lecture 1 (630 words)

Source: http://dinosaurs.about.com/od/dinosaurevolution/a/bigdinos.htm

One of the things that makes dinosaurs so appealing is their sheer size: plant eaters like Diplodocus and Brachiosaurus weighed well over 50 tons, and a well-toned T. Rex tipped the scales at 7 or 8 tons. From the fossil evidence, it's clear that dinosaurs were more massive--species by species, individual by individual--than any other group of animals that ever lived, including modern mammals.

The giant size of dinosaurs demands an explanation and one that's compatible with other theories--for example, it's impossible to discuss dino gigantism without paying close attention to the whole cold-blooded/warm-blooded debate.

So what's the current state of thinking on plus-sized dinos? Here are a few more-or-less interrelated theories.

Theory #1: Dino size was fueled by vegetation.

During the Mesozoic Era--which stretched from the beginning of the Triassic Era, 250 million years ago, to the extinction of the dinosaurs at the end of the Cretaceous Era, 65 million years ago--atmospheric levels of carbon dioxide were much higher than they are today. If you've been following the global warming debate, you'll know that increased carbon dioxide is directly correlated with temperature--meaning the global climate was much warmer millions of years ago than it is today.

This combination of high levels of carbon dioxide and high temperatures meant that the prehistoric world was matted with all kinds of vegetation--plants, trees, mosses, etc. Like kids at an all-day dessert buffet, dinosaurs may have evolved to giant sizes simply because there was a surplus of nourishment. This would also explain why the predatory dinosaurs got so big; a 50-pound carnivore wouldn't have had much of a chance against a ten-ton plant eater.

Theory #2: Hugeness in dinosaurs was a form of self-defense.

If Theory #1 strikes you as a bit simplistic, your instincts are correct: the mere availability of huge amounts of vegetation doesn't entail the evolution of giant creatures who can swallow it down to the last shoot. Evolution works along multiple paths, and the drawbacks of gigantism

(such as slow speed and limited population size) can easily outweigh its benefits in terms of food-gathering.
Some paleontologists think gigantism conferred an evolutionary advantage on the dinosaurs who had it: specifically, a jumbo-sized herbivore would have been virtually immune to attacks by predators. (This theory also lends some credence to the idea that T. Rex scavenged for its food--say, by happening on the body of an Apatosaurus that died of disease or old age--rather than actively hunting it down.)
Theory #3: Dino gigantism was a byproduct of cold-bloodedness.

Many paleontologists who study giant herbivores believe that these dinosaurs were cold-blooded, for two compelling reasons: first, based on our current models of metabolism, a warm-blooded Diplodocus would have cooked itself from the inside, like a potato, and promptly expired; and second, no land-dwelling, warm-blooded mammals living today even approach the size of the large, herbivorous dinos (elephants top out at a couple of tons, max).
Here's where the gigantism comes in. If it evolved to a large-enough size, scientists believe, a cold-blooded plant-eater could have achieved "homeothermy"--that is, the ability to maintain its own temperature. This is because a house-sized, cold-blooded creature would warm up in the sun and cool down at night very slowly, giving it a fairly constant average temperature.
The problem is, these speculations about cold-blooded herbivores run counter to the current vogue for warm-blooded dinosaurs. Although it's not impossible that a warm-blooded T. Rex could have coexisted alongside a cold-blooded Brachiosaurus, evolutionary biologists would be much happier if all dinosaurs had uniform metabolisms--even if these were "intermediate" metabolisms that haven't yet been modeled.
We have talked about three different theories about dinosaurs. What's your opinion on these theories? We will discuss more about the theories next time.
Q1. What was mainly discussed in this lecture?

a) To discuss one strong theory about the Dinosaurs' extinction
b) To discuss possible theories about blood types of Dinosaurs
c) To discuss acceptable theories about Dinosaurs' size

Q2. Big dinosaurs consumed vegetables during the Mesozoic Era. What was the main factor that made the environment to provide enough plants for Dinosaurs?

a) Higher levels of carbon dioxide *
b) Higher levels of oxygen
c) Higher levels of hydrogen

Q3. What is the proper meaning of "homeothermy" in this lecture?

a) Dinosaurs were able to control their body temperature through external means
b) Some large dinosaurs were able to maintain their own body temperature *
c) Some large dinosaurs were able to switch their temperature through internal and external means

Q4. Which dinosaur is a plant eater?

a) T-rex
b) Diplodocus*
c) Carnivore

Q5. Which statement is true?

a) Dino gigantism was a byproduct of cold-bloodedness*
b) Hugeness in dinosaurs was a form of attack

c) Dino size was fueled by meat

Presentation 2 (479 words)

Source: http://www.canadafreepress.com/index.php/article/1921

I will talk about climate change and global warming based on articles that we read last time. There is a claim that storms and severe weather will increase with global warming. Most major storms and severe weather, including tornadoes, occur in the middle latitudes between approximately 30 and 65 degrees of latitude. Fronts are the battle zone between different air masses and as they move they are labeled warm or cold. If you are warm and the temperature drops, a Cold Front has passed; if you are cold and the temperature rises, a Warm Front has passed. It's the cold air that dictates what happens because it is more dense and heavier than the warm air. It pushes the warm air out of the way or allows the warm air to move in behind.

Overall, Earth's atmosphere is in two air masses with a dome of cold polar air over each pole and over-running warm subtropical air separated by the Polar Front. The temperature difference across the Front is variable but quite dramatic most of the time. It is this difference that creates pressure differences and very strong winds.
At the surface waves develop and spiral into low pressure systems known as mid-latitude cyclones. They migrate along the Front like a wave moving through the ocean. In winter they bring snow and are called blizzards; in summer they bring heavy rain, occasionally with severe thunderstorms and tornadoes.

The Front moves seasonally as the cold dome expands and contracts with the changing sun angle. As the dome moves through latitudes the seasons change, marked by these low pressure storm systems.

In the US, the most extreme temperature contrast across the Front occurs when cold air pushes well south and meets with warm moist air coming off the Gulf of Mexico. This pattern creates a general zone running from the Texas panhandle northeast through the Ohio valley and in to southwest Ontario. This zone is known as Tornado Alley. It's a wide zone that varies with the season and conditions.

Storms driven largely by latent heat, and that includes thunderstorms, are expected to become stronger as the air becomes warmer and contains more moisture. Global warming does cause just such a tendency.
A researcher, Henson also claims that global warming will result in greater warming in polar air than in tropical air. This means the temperature difference across the Polar Front will decrease and, as a result, the strength of the major mechanism for storm creation will decrease.

This influx of warm moist air is needed to meet with the cold air that pushes far south, as it has all this winter. It will continue to do so the Earth continues to cool, as it has generally since 1998. The dilemma then is that storms will most likely increase in frequency and severity, but it will be because of global cooling, not warming.

Q1. According to this lecture, what create pressure differences and strong winds?

a) Temperature differences across the Front*
b) A warm air mass with a cold polar
c) The Front with the changing sun angle

Q2. The "Tornado Alley" refers to

a) A general zone running from the Texas through the Gulf of Mexico
b) The most extreme temperature contrast occurring across the Front in US*
c) Temperature difference across the Polar Front

Q3. According to this lecture, which claim is more plausible in storm creation?

a) Global warming causes more destructive tornadoes and cyclones.
b) Global cooling causes more frequent and severe storms.*
c) Greater warming in polar air creates more storms.

Q4. In this talk, the speaker mentioned a research who argued the cause of global warming. What was his claim?

a) He claimed that global warming is unstoppable.
b) He claimed that global warming will result in greater warming in polar air.*
c) He claimed that the Polar Front will increase and storm creation will increase.

Q5. What is the cause of frequent storms?

a) Global cooling*
b) Global warming
c) Global alley

Lecture 2 (796 words)

Source: http://www.buzzle.com/articles/typhoon-vs-hurricane-vs-tornado.html

Many of you told me that you are confused about what typhoons, hurricanes and tornadoes are. Today, I will clarify all your doubts with these storm systems, which will highlight the essential distinction between them. All though these storm systems differ from each other, in one thing they are the same. They all wreck havoc in all the regions, that they pass through. Before we go over and get into these storm systems, let me explain the definitions of these terms. First, Typhoons, hurricanes and tropical cyclones are three different region specific names for the same kind of storm system. In other words, they are one and the same phenomenon, although varying in intensity according to the place of origin and conditions. Hurricane is a peculiar storm system, which has a warm low pressure center, with an army of thunderstorms around it. They have their origin near the equator, around 10 degrees away from it, in the sea. Tornadoes are storm systems that form on land due to pressure differences. They are characterized by large rotating air columns, which are like funnels that are connected with clouds at the top and land below. They move with phenomenal speeds, touching 300 mph on land before dissipating. Hurricane is the same as a typhoon, the only difference being the name. The same phenomenon is known as a typhoon and a hurricane in different regions, according to the local language.

Let's discuss the origin mechanism and occurrence areas of typhoons, hurricanes and tornadoes. A hurricane and a typhoon are the same thing. Just like the rose by any other name, would smell as sweet, a typhoon by any other name, would still be as deadly! In meteorological terminology, it's called a tropical cyclone. All storms like tropical cyclones are created due to severe

differences in air pressure, caused due to temperature differences!

A tropical cyclone, like a hurricane or a tornado, is a type of storm system, which has closed wind circulation around a central low pressure area, having their origin in the Ocean. The circulation is further fueled by heat released through moist air, which condenses as it rises. Warm air is lighter and therefore exerts low air pressure. Tropical cyclones are sustained by warm cores, which maintain the low pressure at the center. They are created in the tropics every year and wreak havoc as they approach the inland coastal area. A recent example is hurricane Katrina, which was one of the most destructive tropical cyclones to ravage the North American coast land.

A tornado is a type of violent storm which occurs on land, again created by wind, moving from surrounding high pressure areas into a low pressure center. It is a fiercely rotating air column, simultaneously in contact with land at the bottom and a sometimes funnel like cumuliform cloud at the top. Tornadoes are one of the deadliest and most destructive storm systems, particularly occurring all over the world, but more common in North America. So, the main difference in hurricane and tornado comparison is, the fact that hurricanes originate in tropical seas, while tornadoes are created inland. Hurricanes are extremely powerful over the sea, causing tides and torrential rains around, but weaken and die out as they more over land, causing major damage in coastal areas. There is another point of similarity between hurricanes and typhoons, due to both being vortex based systems. Due to the 'Coriolis effect', hurricanes and typhoons in the northern hemisphere rotate counterclockwise and clockwise in the southern hemisphere. The central part or the eye of a typhoon could be as big as 370 kilometers! The hurricane season for the North Atlantic Ocean starts from 1st of June and continues up till November 30.
The average wind speed of most tornadoes ranges from 40 - 110 miles per hour, but some of the most powerful ones attain wind speeds in excess of 300 miles per hour! There are three different types of tornadoes, which are landspout, multiple vortex tornado and waterspout. A tornado warning system is created with the use of weather radar and a chain of storm spotters. Tornado safety is in early warnings and announcement and in the construction of underground storm shelters. Tornadoes occur in a region in USA called 'Tornado Alley'. Tornado season begins in spring as it is a time of transition in temperatures. Hope this distinction has cleared your ideas about all the three terms. How much ever we develop our technology, we will never be able to tame nature completely. An early warning system is the only mechanism which can be our savior from nature's wrath. Please read articles in the syllabus and prepare for the final exam.
Q1. Which statement is true?

a) Hurricanes and cyclones are same storm systems but not typhoon.
b) Tornadoes are storm systems that form on land due to air pressure.
c) Katrina was one of the most destructive tropical cyclones in the North American coast land*

Q2. Which one is <u>not</u> a type of tornadoes?

a) Multiple vortex tornado

b) Waterspout
c) Earthspout*

Q3. Hurricanes and typhoons are very similar due to _____ system.

a) Vortex *
b) Tornado
c) Torrential rain

Q4. According to this lecture, what is the best example of the 'Coriolis effect'?

a) Two hurricanes are rotating at the same direction.
b) Hurricanes in the northern hemisphere rotate counterclockwise and clockwise in the southern hemisphere *
c) Typhoons rotate clockwise in the northern hemisphere

Q5. What can you learn from this lecture?

a) Typhoons, Hurricanes, and Cyclones are storm systems which have same origin and mechanism.
b) These storm systems occur on different area but surrounding pressure areas are same.
c) Hurricanes can be very powerful over the sea but weaken out as they moved over land.*

Appendix D

Test Booklet

[English Listening Proficiency Test]

Please fill out your information.

Email: _____

Gender: Male ( ), Female ( )

Status: Undergraduate ( ), Graduate ( ), other ( )
Major: _____
Nationality: _____
First Language: _____
TOEFL score: _____ PBT ( ), CBT ( ), iBT ( )
Years lived in U.S: _____

Participant Consent Form

Hello. We would like to invite you to participate in a research project that examines test development for second language listening. This research is being carried out by Youngshin Chi for a doctoral dissertation under supervision of Dr. Fred Davidson in the department of Linguistics at University of Illinois at Urbana-Champaign. The purpose of this research is to develop and to validate a new listening comprehension test for English language learners, and to investigate learners' awareness of listening comprehension. If you agree to participate in this research, you will need 30 minutes to take listening test and to answer cognitive questionnaire which will be conducted in a reserved computer lab.

Your participation in this project is voluntary. This means that you can decide whether you want to do this project or not. Your choice to participate or not will not impact your grades or status at the university. If you want to stop doing the project at any time, you can stop. Your test scores and all the other information from this research will be kept private and secure. All information will be kept in a locked file cabinet and only people (Youngshin Chi and Dr. Fred Davidson) who work on this research will be able to look at them. A pseudonym will be used in any analysis of the data in the final research paper and discussion with the dissertation committee. Youngshin will use your information and results for her doctoral dissertation, poster or conference presentation, and journal article.

After this session, we will have an interview with participants who are interested in sharing their own opinions on the test and test taking strategy. The interview will take about 30 minutes. If you are interested in participating an interview session, please leave your contact information here.

[ ] yes, I am interested in participating an interview session.

The only possibility of risk involved would be slight fatigue.

You will be given a copy of this consent form. If you have any questions about the research or the result, please feel free to contact Youngshin Chi by e-mail at ychi1@illinois.edu and Dr. Fred Davidson by e-mail at fgd@illinois.edu.

Print name: _____ Date : / / /

Signature:

Questions: Youngshin Chi, Graduate student, Dept. Educational Psychology

University of Illinois-Urbana Champaign, Phone: 217-390-1449 E-mail: ychi1@illinois.edu

****

If you have any questions about your rights as a research participant please contact Anne Robertson, Bureau of Educational Research, 217-333-3023, or arobrtsn@ad.illinois.edu or the Institutional Review Board at 217-333-2670 or irb@illinois.edu.

Instruction: You will listen to a short presentation about dinosaurs. Please click the sound button when you are ready. Audio will be played once.

Q1. Dinosaurs lived during the Mesozoic Era. The Mesozoic Era is also known as_____.

d) Age of ultrasauros
e) Age of Deinos and sauros
f) Age of Reptiles

Q2. Tyrannosaurs are bipedal dinosaurs, Brachiosaurus are _____ dinosaurs.

d) Quadrupedal
e) Hylaeosaurus
f) Cretaceous

Q3. Many theories have explained the extinction of Dinosaurs. What is the most supportive theory?

d) Due to volcano eruptions
e) Due to an asteroid impact
f) Due to the ice age

Q4. Which statement is true?

d) Compsognathus is 100 feet long and up to 50 feet tall
e) Supersaurus is 30 m long and 15 m tall
f) Velociraptor is slow and lumbering

Q5. What is the main idea of this presentation?

d) All dinosaurs are now remained as birds
e) Dinosaurs lived in different eras and they had similar shapes though they were born from eggs
f) It's very difficult to figure out dinosaurs' sound, or gender (whether they are females or males) through fossils

Instruction: You will listen to a lecture about dinosaurs. Please click the sound button when you are ready. Audio will be played once.

Q1. Which dinosaur is a plant eater?

d) T-rex
e) Diplodocus
f) Carnivore

Q2. Big dinosaurs consumed vegetables during the Mesozoic Era. What was the main factor that made the environment to provide enough plants for Dinosaurs?

d) Higher levels of carbon dioxide
e) Higher levels of oxygen
f) Higher levels of hydrogen

Q3. What is the proper meaning of "homeothermy" in this lecture?

d) Dinosaurs were able to control their body temperature through external means
e) Some large dinosaurs were able to maintain their own body temperature

f) Some large dinosaurs were able to switch their temperature through internal and external means

Q4. Which statement is true?

a) Dino gigantism was a byproduct of cold-bloodedness
b) Hugeness in dinosaurs was a form of attack
c) Dino size was fueled by meat

Q5. What was mainly discussed in this lecture?

d) To discuss one strong theory about the Dinosaurs' extinction
e) To discuss possible theories about blood types of Dinosaurs
f) To discuss acceptable theories about Dinosaurs' size

Instruction: You will listen to a short presentation about global warming. Please click the sound button when you are ready. Audio will be played once.

Q1. What is the cause of frequent storms?

d) Global cooling
e) Global warming
f) Global alley

Q2. The "Tornado Alley" refers to

d) A general zone running from the Texas
e) The zone where tornadoes occurring across the Front in US
f) Temperature difference across the Polar Front

Q3. According to this lecture, what create pressure differences and strong winds?

d) Temperature differences across the Front
e) A warm air mass with a cold polar
f) The Front with the changing sun angle

Q4. According to this lecture, which claim is more plausible in storm creation?

d) Global warming causes more destructive tornadoes and cyclones.
e) Global cooling causes more frequent and severe storms.
f) Greater warming in polar air creates more storms.

Q5. In this presentation, the speaker mentioned a researcher who argued the cause of global warming. What was his claim?

d) He claimed that global warming is unstoppable.
e) He claimed that global warming will result in greater warming in polar air.
f) He claimed that the Polar Front will increase and storm creation will increase.

Instruction: You will listen to a lecture about storm system. Please click the sound button when you are ready. Audio will be played once.

Q1. Which one is <u>not</u> a type of tornadoes?

d) Multiple vortex tornado
e) Waterspout

f) Earthspout

Q2. Hurricanes and typhoons are very similar due to _____ system.

d) Vortex
e) Tornado
f) Torrential rain

Q3. Which statement is true?

d) Hurricanes and cyclones are same storm systems but not typhoon.
e) Tornadoes are storm systems that form on land due to air pressure.
f) Katrina was one of the most destructive tropical cyclones in the North American coast land

Q4. According to this lecture, what is the best example of the 'Coriolis effect'?

d) Two hurricanes are rotating at the same direction.
e) Hurricanes in the northern hemisphere rotate counterclockwise and clockwise in the southern hemisphere.
f) Typhoons rotate clockwise in the northern hemisphere.

Q5. What can you learn from this lecture?

d) Typhoons, Hurricanes, and Cyclones are storm systems which have same origin and mechanism.
e) These storm systems occur on different area but surrounding pressure areas are same.
f) Hurricanes can be very powerful over the sea but weaken out as they moved over land.

Appendix E

Cognitive Questionnaire

The purpose of this questionnaire is to understand your cognitive awareness in second language listening. Please check the appropriate answer.

1. Before I start to listen, I have a plan in my mind for how I am going to listen.
   Disagree Agree Don't remember

2. I feel that listening comprehension in English is a challenge for me
   Disagree Agree Don't remember

3. I translate key words as I listen
   Disagree Agree Don't remember

4. I try to get back on track when I lose concentration
   Disagree Agree Don't remember

5. As I listen I quickly adjust my interpretation if I realize that it is not correct.
   Disagree Agree Don't remember

6. I don't feel nervous when I listen to the listening passages.
   Disagree Agree Don't remember

7. I translate word by word as I listen
   Disagree Agree Don't remember

8. When I guess the meaning of a word, I think back to everything else that I have heard, to see if my guess makes sense.
   Disagree Agree Don't remember

9. As I listen, I periodically ask myself if I understand everything that I heard.
   Disagree Agree Don't remember

10. When I listen to the listening passage, if I don't understand something, I guess what the word or phrase might mean based on the context.
    Disagree Agree Don't remember

11. When I listen to the passage, if I don't understand something, I find myself thinking about the segment and passively listening.
    Disagree Agree Don't remember

12. When I listen to the passage, if I don't understand something, I just give up trying to comprehend the passage.
    Disagree Agree Don't remember

13. When I listen to the passage, I tend to focus on understanding the meaning of each sentence rather than the overall meaning of the text.
    Disagree Agree Don't remember

14. Before I listen to the passage, I try to predict the content of the passage by reviewing comprehension questions.
    Disagree Agree Don't remember

15. Note-taking is very helpful when I listen to the passage.
    Disagree Agree Don't remember

16. I read comprehension questions first before I listen to the passages.
    Disagree Agree Don't remember

17. Background knowledge of the topics was helpful to answer the comprehension questions.
    Disagree Agree Don't remember

18. When I listen to the passage, I try to memorize as much as I can in order to answer questions.
    Disagree Agree Don't remember

19. I use different listening strategies when I listen to long or short talks
    Disagree Agree Don't remember

Appendix F

Reflection Questions

1. What was your overall impression on this test?
2. You were given (breakdown type) test, were you bothered by this factor when you listen to the lecture?
3. If so, how did you overcome this obstacle?
4. Could you share your test taking strategy?
5. What element should be modified in this test?

Appendix G

Item Statistics

1. Item difficulty analysis across four groups

Item Statistics

Group		Mean	Std. Deviation	N
British	i2	.7917	.41485	24
	i3	.9583	.20412	24
	i4	.6250	.49454	24
	i5	.6250	.49454	24
	i6	.6667	.48154	24
	i7	.9583	.20412	24
	i8	.7500	.44233	24
	i9	.9167	.28233	24
	i10	.9167	.28233	24
	i11	.2500	.44233	24
	i12	.4167	.50361	24
	i13	.6667	.48154	24
	i14	.8333	.38069	24
	i15	.8333	.38069	24
	i16	.7917	.41485	24
	i17	.7083	.46431	24
	i18	.7500	.44233	24

	i19	.9167	.28233	24
	i20	.6250	.49454	24
Noise	i2	.8750	.33783	24
	i3	.9583	.20412	24
	i4	.5000	.51075	24
	i5	.5833	.50361	24
	i6	.7500	.44233	24
	i7	.9583	.20412	24
	i8	.7083	.46431	24
	i9	.9583	.20412	24
	i10	.8333	.38069	24
	i11	.2083	.41485	24
	i12	.6667	.48154	24
	i13	.7500	.44233	24
	i14	.4167	.50361	24
	i15	.5000	.51075	24
	i16	.7917	.41485	24
	i17	.7917	.41485	24
	i18	.6250	.49454	24
	i19	.8750	.33783	24
	i20	.5833	.50361	24
	i1	.9583	.20412	24
Regular	i2	.8333	.38069	24

	i3	.6250	.49454	24
	i4	.4167	.50361	24
	i5	.7917	.41485	24
	i6	.6667	.48154	24
	i8	.7500	.44233	24
	i9	.9583	.20412	24
	i10	.8333	.38069	24
	i11	.4167	.50361	24
	i12	.5417	.50898	24
	i13	.6250	.49454	24
	i14	.6250	.49454	24
	i15	.5417	.50898	24
	i16	.6667	.48154	24
	i17	.7917	.41485	24
	i18	.5833	.50361	24
	i19	.9583	.20412	24
	i20	.5000	.51075	24
	i1	.8750	.33783	24
Speech rate	i2	.8333	.38069	24
	i3	.8750	.33783	24
	i4	.3750	.49454	24
	i5	.6250	.49454	24
	i6	.7083	.46431	24

	i8	.8333	.38069	24
	i9	.9583	.20412	24
	i10	.9583	.20412	24
	i11	.3333	.48154	24
	i12	.5833	.50361	24
	i13	.7500	.44233	24
	i14	.7083	.46431	24
	i15	.6667	.48154	24
	i16	.8333	.38069	24
	i17	.7917	.41485	24
	i18	.6667	.48154	24
	i19	.9167	.28233	24
	i20	.5833	.50361	24
	i1	.9583	.20412	24

2. Item discrimination analysis

Group		Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
B	i2	13.2083	5.563	.446	.481
	i3	13.0417	6.737	-.161	.561
	i4	13.3750	5.549	.350	.493
	i5	13.3750	6.332	.013	.563

	i6	13.3333	6.232	.060	.553
	i7	13.0417	6.216	.345	.519
	i8	13.2500	5.065	.677	.429
	i9	13.0833	6.167	.258	.521
	i10	13.0833	6.341	.133	.536
	i11	13.7500	6.543	-.058	.571
	i12	13.5833	6.167	.075	.551
	i13	13.3333	5.971	.172	.531
	i14	13.1667	6.058	.217	.523
	i15	13.1667	5.884	.314	.507
	i16	13.2083	6.607	-.080	.572
	i17	13.2917	5.607	.358	.493
	i18	13.2500	5.761	.307	.505
	i19	13.0833	6.254	.195	.529
	i20	13.3750	6.505	-.056	.576
N	i2	13.4167	3.384	.157	.024
	i3	13.3333	3.449	.268	.021
	i4	13.7917	3.389	.023	.071
	i5	13.7083	3.694	-.131	.150
	i6	13.5417	3.042	.296	-.064[a]
	i7	13.3333	3.884	-.288	.137
	i8	13.5833	3.384	.055	.055
	i9	13.3333	3.797	-.182	.116

	i10	13.4583	3.303	.178	.008
	i11	14.0833	3.819	-.183	.157
	i12	13.6250	3.636	-.095	.129
	i13	13.5417	3.563	-.039	.100
	i14	13.8750	2.984	.262	-.065[a]
	i15	13.7917	2.694	.441	-.183[a]
	i16	13.5000	3.652	-.082	.116
	i17	13.5000	3.043	.330	-.072[a]
	i18	13.6667	3.884	-.223	.192
	i19	13.4167	3.819	-.181	.141
	i20	13.7083	4.129	-.336	.245
	i1	13.3333	3.362	.387	-.006[a]
R	i2	12.1667	5.971	.452	.417
	i3	12.3750	6.245	.189	.462
	i4	12.5833	6.167	.214	.456
	i5	12.2083	7.216	-.193	.532
	i6	12.3333	6.580	.059	.490
	i8	12.2500	6.283	.216	.457
	i9	12.0417	6.476	.422	.450
	i10	12.1667	6.406	.211	.461
	i11	12.5833	6.254	.178	.464
	i12	12.4583	6.433	.102	.482
	i13	12.3750	7.375	-.247	.553

	i14	12.3750	5.984	.301	.436
	i15	12.4583	5.563	.472	.392
	i16	12.3333	6.841	-.046	.512
	i17	12.2083	5.998	.387	.425
	i18	12.4167	5.906	.326	.429
	i19	12.0417	6.476	.422	.450
	i20	12.5000	7.304	-.220	.551
	i1	12.1250	6.549	.170	.469
SR	i2	13.1250	5.332	.322	.443
	i3	13.0833	5.210	.465	.423
	i4	13.5833	5.123	.301	.439
	i5	13.3333	5.449	.151	.476
	i6	13.2500	4.891	.455	.403
	i8	13.1250	5.071	.482	.412
	i9	13.0000	6.000	.000	.494
	i10	13.0000	5.739	.267	.468
	i11	13.6250	6.071	-.110	.533
	i12	13.3750	5.114	.296	.440
	i13	13.2083	5.998	-.070	.521
	i14	13.2500	5.761	.029	.502
	i15	13.2917	5.085	.334	.432
	i16	13.1250	6.375	-.249	.544
	i17	13.1667	5.536	.171	.471

	i18	13.2917	5.781	.013	.507
	i19	13.0417	5.433	.402	.442
	i20	13.3750	6.418	-.247	.565
	i1	13.0000	5.565	.451	.449