

**ARGUMENT-BASED VALIDATION OF A  
HIGH-STAKES LISTENING TEST IN VIETNAM**

By

**Tran Thi Ngoc Diep**

A thesis

submitted to Victoria University of Wellington  
in fulfilment of the requirements for the degree of  
Doctor of Philosophy

Victoria University of Wellington

---2020---

## Abstract

More than a decade ago, the Vietnamese Government announced an educational reform to enhance the quality of English language education in the country. An important aspect of this reform is the introduction of the localized test of English proficiency which covers four language skills, namely listening, speaking, reading, and writing. This high-stakes English test is developed and administered by only a limited number of institutions in Vietnam. Although the validity of the test is a considerable concern for test-takers and test score users, it has remained an under-researched area. This study aims to partly address the issue by validating a listening test developed by one of the authorized institutions in Vietnam. In this thesis, the test is referred to as the Locally Created Listening Test or the LCLT.

Using the argument-based approach to validation (Kane, 1992, 2013; Chapelle, 2008), this research aims to develop a validity argument for the evaluation, generalization and explanation inferences of the LCLT. Three studies were carried out to elicit evidence to support these inferences. The first study investigated the statistical characteristics of the LCLT test scores, focusing on the evaluation and generalization inference. The second study shed light on the extent to which test items engaged the target construct. The third study examined whether test-takers' scores on the LCLT correlated well with their scores on an international English test that measured a similar construct. Both the second and third study were carried out to support the explanation inference.

These three studies did not provide enough evidence to successfully support the validity argument for the LCLT. The test was found to have major flaws that affected the validity of score interpretations. In light of the research findings, suggestions were given for the betterment of future LCLTs. At the same time, this research helped to uncover the impacts of certain text and task-related factors on the test-takers' performance. Such insights led to practical implications for the assessment of second language listening in general. The

results of this research also contributed to the theory and practice of test localization, a relatively new paradigm in language testing and assessment.

## Table of contents

<b>1. Chapter 1. Introduction .....</b>	<b>1</b>
1.1. Changes in foreign language education policies in Vietnam .....	2
1.1.1. Vietnamese six-level framework of reference for foreign languages .....	4
1.1.2. The three-level test of English proficiency for Vietnamese people.....	7
1.2. Issues in assessing listening .....	8
1.2.1. The inadequacy of relevant literature.....	9
1.2.2. The development of listening test tasks .....	9
1.2.3. Factors affecting second language listening.....	11
1.3. Validity and the argument-based approach to validation.....	12
1.4. Overall structure and aims of the research .....	15
1.5. Ethical considerations .....	18
1.6. Significance of the research .....	18
1.7. Organization of the thesis.....	20
<b>2. Chapter 2. Literature Review.....</b>	<b>21</b>
2.1. Listening in a second language .....	21
2.1.1. Bottom-up processes.....	24
2.1.2. Top-down processes.....	26
2.2. The assessment of second language listening .....	29
2.2.1. Defining the listening construct.....	29
2.2.2. Factors affecting listening task difficulty .....	37
2.2.3. Construct-irrelevant variance.....	43

2.3. The concept of validity and validation .....	45
2.3.1. Three types of validity .....	45
2.3.2. Validity as a unitary concept.....	46
2.4. The argument-based approach to validation .....	49
2.4.1. Interpretive argument.....	50
2.4.2. Validity argument .....	54
2.4.3. The argument-based approach in practice .....	58
<b>3. Chapter 3. Overall methodology and research design .....</b>	<b>61</b>
3.1. General description of the LCLT .....	61
3.2. Argument-based validity for the LCLT .....	62
3.3. Overall research design .....	66
<b>4. Chapter 4. Statistical characteristics of the LCLT scores .....</b>	<b>68</b>
4.1. Aims and methodology of the study .....	68
4.1.1. Participants.....	69
4.1.2. Data collection procedure .....	69
4.1.3. Data analysis procedure .....	70
4.2. Statistical characteristics of the LCLT scores .....	71
4.2.1. Descriptive statistics .....	71
4.2.2. The difficulty levels of LCLT items .....	72
4.2.3. Discriminating power of LCLT items.....	77
4.2.4. Evidence of construct-irrelevant variance .....	79
4.2.5. Reliability of the LCLT scores .....	83

4.3. Assessment of the evaluation and generalization inferences .....	84
4.3.1. Inappropriate difficulty levels.....	84
4.3.2. Reasonable discriminating power .....	85
4.3.3. Relevance between test items and the targeted listening construct .....	85
4.3.4. Unsatisfactory person reliability .....	86
<b>5. Chapter 5. The test-takers' cognitive processes.....</b>	<b>89</b>
5.1. Verbal report in theory and in language testing research.....	89
5.1.1. Theoretical background of verbal reports .....	89
5.1.2. Verbal reports in research on listening tests .....	92
5.1.3. The validity and reliability of verbal reports .....	94
5.2. The specifications and actual design of the LCLT.....	95
5.2.1. Descriptors for listening skill in the CEFR-V.....	96
5.2.2. The LCLT specifications .....	97
5.2.3. The actual design of the LCLT .....	99
5.3. Aims and methodology of the study .....	100
5.3.1. Participants.....	100
5.3.2. Collection of verbal reports .....	101
5.3.3. Analysis of verbal reports .....	102
5.4. Results of verbal report analysis .....	104
5.4.1. Coverage of the target construct .....	104
5.4.2. Factors affecting test-takers' performance on LCLT items.....	165
5.5. Judgement on the first warrant of the explanation inference .....	177

5.5.1. Poor coverage of the targeted listening construct .....	177
5.5.2. Presence of construct-irrelevant variance .....	179
<b>6. Chapter 6. The comparability between the LCLT and IELTS Listening .....</b>	<b>180</b>
6.1. Background of the study .....	180
6.2. Comparability studies on language tests .....	181
6.3. Methodology .....	185
6.3.1. Participants.....	186
6.3.2. Test administration.....	186
6.3.3. Expert judges .....	187
6.3.4. Data analysis .....	187
6.4. Evidence of comparability between the LCLT and IELTS.....	190
6.4.1. Results of construct analysis.....	191
6.4.2. Results of the quantitative analysis on test-takers' performance.....	197
6.5. Judgement on the second warrant of the explanation inference.....	203
<b>7. Chapter 7. Overall discussion.....</b>	<b>205</b>
7.1. Argument-based validity of the LCLT.....	205
7.2. Positive features of the LCLT .....	210
7.3. Problems with the LCLT and proposed solutions.....	212
7.3.1. Poor construct coverage.....	212
7.3.2. Inappropriate difficulty levels.....	216
7.3.3. Substantial construct-irrelevant variance.....	218
7.3.4. Unsatisfactory person reliability .....	219

7.4. Issues in the localization of English proficiency tests.....	220
7.4.1. Poor construct conceptualization .....	221
7.4.2. Substandard technical quality .....	225
7.4.3. Lack of quality control.....	226
7.4.4. Negative social consequences.....	228
7.5. Concerns about the assessment of second language listening.....	230
7.5.1. Undesirable impacts of the multiple-choice format.....	230
7.5.2 Practical considerations for developing listening tests .....	232
7.6. Critical reflection on the argument-based approach to validation .....	235
<b>8. Chapter 8. Conclusion.....</b>	<b>237</b>
8.1. Contributions of the research .....	237
8.2. Practical implications of the research.....	239
8.3. Limitations of the research.....	240
8.4. Suggestions for future research .....	242
8.5. Concluding comments.....	243
<b>References.....</b>	<b>244</b>
<b>Appendices.....</b>	<b>268</b>



## List of tables

<b>Table 1.1.</b> Correspondence between CEFR-V and CEFR levels .....	4
<b>Table 1.2.</b> Examples of CEFR-V global scales .....	5
<b>Table 1.3.</b> English proficiency requirements in Vietnam.....	6
<b>Table 2.1.</b> Cognitive attributes of L2 listening.....	33
<b>Table 2.2.</b> Overall listening comprehension in the CEFR.....	36
<b>Table 2.3.</b> Factors affecting listening task difficulty.....	38
<b>Table 2.4.</b> A taxonomy of construct-irrelevant variance.....	44
<b>Table 2.5.</b> The two-by-two representation of validity.....	47
<b>Table 3.1.</b> Content of the LCLT .....	62
<b>Table 4.1.</b> The evaluation and generalization inferences for the LCLT .....	68
<b>Table 4.2.</b> Descriptive statistics for the LCLT scores .....	71
<b>Table 4.3.</b> Difficulty measure of LCLT items.....	74
<b>Table 4.4.</b> Discrimination indices for LCLT items .....	78
<b>Table 4.5.</b> Fit statistics of LCLT items.....	80
<b>Table 4.6.</b> Reliability indices for the LCLT .....	83
<b>Table 4.7.</b> Evidence and judgements on the evaluation and generalization inferences .....	87
<b>Table 5.1.</b> General descriptors for listening in CEFR-V.....	96
<b>Table 5.2.</b> The LCLT specifications.....	98
<b>Table 5.3.</b> The actual content of the LCLT .....	99
<b>Table 5.4.</b> Background information of the participants .....	101
<b>Table 5.5.</b> Listening subskills targeted by the LCLT .....	103
<b>Table 5.6.</b> Listening subskills engaged by the LCLT items.....	105
<b>Table 5.7.</b> Items engaging the ability to understand clearly articulated details .....	107
<b>Table 5.8.</b> Items engaging the ability to understand implicit details.....	114

<b>Table 5.9.</b> Auditory input for item 26, 27 and 28.....	127
<b>Table 5.10.</b> Evidence of test-taking strategies in the verbal data .....	167
<b>Table 6.1.</b> The framework for describing listening ability.....	189
<b>Table 6.2.</b> Expert judgements on the LCLT construct .....	192
<b>Table 6.3.</b> Expert judgements on the IELTS construct .....	195
<b>Table 6.4.</b> Descriptive statistics for the LCLT and IELTS scores .....	197
<b>Table 6.5.</b> Person and item reliability for the LCLT and IELTS .....	201
<b>Table 6.6.</b> Correlation between LCLT and IELTS scores.....	202
<b>Table 7.1.</b> Summary of the inferences, warrants and relevant studies .....	206

## List of figures

<b>Figure 1.1.</b> Overall structure of the research.....	17
<b>Figure 2.1.</b> Cognitive processes in L2 listening and their interrelationship .....	23
<b>Figure 2.2.</b> Six levels of proficiency in the CEFR.....	35
<b>Figure 2.3.</b> The bridge analogy .....	50
<b>Figure 2.4:</b> Links in the extended interpretive argument.....	52
<b>Figure 2.5.</b> Five inferences in an interpretive argument .....	53
<b>Figure 2.6.</b> Toulmin's model of inference .....	55
<b>Figure 2.7.</b> Example of an evaluation inference for a listening test.....	57
<b>Figure 3.1.</b> Validation framework for the LCLT .....	65
<b>Figure 3.2.</b> Overall research design .....	67
<b>Figure 4.1.</b> LCLT score distribution .....	72
<b>Figure 4.2.</b> Item-person map for the LCLT .....	76
<b>Figure 4.3.</b> Bubble chart representing infit ZSTD of LCLT items .....	81
<b>Figure 5.1.</b> Various types of verbal report procedure .....	90
<b>Figure 6.1.</b> Comparison of the listening construct between the LCLT and IELTS .....	196
<b>Figure 6.2.</b> Item difficulty and person ability in the LCLT .....	199
<b>Figure 6.3.</b> Item difficulty and person ability in IELTS .....	200
<b>Figure 7.1.</b> Argument-based validity for the LCLT .....	209

**Abbreviations**

AUA: Assessment Use Argument

AEC: ASEAN Economic Community

CEFR: Common European Framework of Reference

CIV: Construct-Irrelevant Variance

GEPT: General English Proficiency Test

IELTS: International English Language Testing System

IUA: Interpretation/Use Argument

LCLT: Locally Created Listening Test

MCQ: Multiple-Choice Question

MOET: Ministry of Education and Training

TOEFL: Test of English as a Foreign Language

TOEIC: Test of English for International Communication

## **Acknowledgement**

I would like to express my deepest gratitude to my patient and supportive supervisors, Associate Professor Peter Gu and Associate Professor Averil Coxhead. Thank you for giving me your feedback, advice and emotional support when I needed it most.

I gratefully acknowledge the funding I received from Vietnam's Ministry of Education & Training and Victoria University of Wellington. This four-year Phd journey would not have been possible without their financial support.

My appreciation also extends to the Vietnamese university that hosted this entire research. They went to great lengths to make sure that I got the data I needed and for that, I can never thank them enough.

I am indebted to my family in Vietnam for all of their unconditional love and support. Despite the distance, they never run out of ways to show me how much they care.

Last but not least, heartfelt thanks to my partner and his family for making me feel loved and appreciated always.

## Chapter 1. Introduction

My career came to a turning point when new policies in foreign language education in Vietnam dramatically changed my routine as a university English teacher. In 2010, I was appointed to serve as a language expert for the major educational reform which involved setting new standards for English proficiency, assessing English proficiency nation-wide, and re-training those who did not meet the new standards. After taking the job, I travelled all over the country to administer English proficiency tests and provided English training courses to learners in a number of remote provinces. In 2015, following the launch of the English proficiency test for Vietnamese people, I was assigned a new role for which I had no substantial, prior experience. I became an item writer, focusing on the Listening component of the test. It was both a challenge and a source of learning opportunities that later inspired my entire PhD study.

As an item writer, I became much more concerned with the concept of validity. Each item I wrote left me questioning whether it was easy or difficult, whether it truly measured test-takers' listening ability, or whether it accidentally favoured a certain group of candidates and disadvantaged others. The actual work in designing test items made me realize how important it was to ensure the validity of the English proficiency test for Vietnamese people. Yet, there were hardly any efforts in validating this test and each of its four components (Listening, Speaking, Reading, Writing). I, therefore, decided to venture into this under-researched area to seek the truth, answer my own questions and eventually figure out ways to improve the test in the future. Within the limit of a PhD study, I chose to focus only on Listening, the component that I had the most experience with. In this thesis, this component is referred to as the Locally Created Listening Test or the LCLT.

In this research project, the validity of the LCLT was investigated from multiple perspectives. The validity evidence was elicited from not only the test-takers' scores but also the cognitive processes involved in answering test items. Another important part of this research was dedicated to the comparison between the LCLT and an external measure of listening ability. This multi-perspective investigation was guided by the argument-based framework, an increasingly influential approach to test validation over the past 2 decades.

This chapter sets the scene for the entire thesis, casting further light on the personal account of why I decided to embark on this research project. It begins with a general description of the local context that led to the development of the English proficiency test for Vietnamese people. Then, important issues in assessing listening are highlighted. What follows is a brief introduction of the argument-based approach to validation and its application in validating language tests. After that, the focus narrows down to the context, aims, significance, and ethical considerations of this research. The overall structure of the thesis is provided at the end of this chapter.

### **1.1. Changes in foreign language education policies in Vietnam**

In response to the growing need for regional and international integration, Vietnam has made considerable changes to its foreign language policy over the past decade (Dudzik & Nguyen, 2015). Initiated in 2008, the National Foreign Language 2020 Project, hereinafter referred to as Project 2020, has become known as Vietnam's ambitious move in revolutionizing language learning and teaching all over the country with a significant emphasis on English language. According to Decision No. 1400/QĐ-TTg, dated September 30, 2008, by the Prime Minister, the general goal of the project is to ensure that:

By 2020 most Vietnamese students graduating from secondary, vocational schools, colleges and universities will be able to use a foreign language confidently in their daily communication, their study and work in an integrated, multi-cultural and

multi-lingual environment, making foreign languages a comparative advantage of development for Vietnamese people in the cause of industrialization and modernization for the country.

This goal was driven by an urge to promote the competitiveness of the Vietnamese workforce to prepare for regional economic integration after the realization of the ASEAN Economic Community (AEC) in 2015. In an ASEAN context, high levels of English proficiency afford a significant advantage for Vietnam because English is a lingua franca of ASEAN and the AEC allows free movement of labourers (About AEC, n.d). However, it was reported that a large number of Vietnamese workers failed to meet global demands of language proficiency, particularly in English (Nguyen & Hamid, 2015). A report by EF Education First (2013) put Vietnam in the group of countries with low proficiency in English. This was seen as a “hindrance to international integration and the improvement of the quality in Vietnamese higher education” (Nghị & London, 2010, p. 56).

To address that language proficiency crisis, Project 2020 adopted the Common European Framework of Reference (CEFR) in 2008 with the aim to restructure the national foreign language education system and to upgrade its workforce to an international level (Nguyen & Hamid, 2015). That decision reflected a tendency of adopting international standards and introducing Western models of pedagogy to local contexts (Hu & McKay, 2012). However, the CEFR was not applied directly to the Vietnamese context but it underwent a process of adaptation. The outcome of this process was the Vietnamese six-level framework of reference for foreign languages. In this thesis, the framework will be referred to as CEFR-V so as to avoid the repetition of the long title.



### 1.1.1. Vietnamese six-level framework of reference for foreign languages

The Vietnamese six-level framework of reference for foreign languages (CEFR-V) was officially launched through Decree No. 01/2014/TT-BGDĐT, dated January 24, 2014, six years after the CEFR was introduced into the Vietnamese context. The proficiency levels in CEFR-V and its corresponding levels in the CEFR are presented in Table 1.1.

**Table 1.1.** Correspondence between CEFR-V and CEFR levels

CEFR-V levels & interpretation		CEFR levels & interpretation	
Advanced	Level 6	C2	Proficient user
	Level 5	C1	
Intermediate	Level 4	B2	Independent user
	Level 3	B1	
Elementary	Level 2	A2	Basic user
	Level 1	A1	

Similar to the CEFR, CEFR-V uses can-do statements to define learners/users' proficiency at each level. Table 1.2 provides examples of the description of overall language ability for each level in CEFR-V according to Decree No. 01/2014/TT-BGDĐT, 2014.

**Table 1.2.** Examples of CEFR-V global scales

	<b>Level</b>	<b>Description</b>
<b>Advanced</b>	Level 6	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation.
	Level 5	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions.
<b>Intermediate</b>	Level 4	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity with native speakers.
	Level 3	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken.
<b>Elementary</b>	Level 2	Can understand sentences and frequently used expressions related to areas of most immediate relevance. Can communicate with other people about simple, every day topics.
	Level 1	Can understand and use familiar everyday expressions and very basic vocabulary aimed at the satisfaction of needs of a concrete type.

Apart from the global scales, the CEFR-V also provides detailed descriptions for the four language skills: listening, speaking, reading, and writing. Despite a six-year process of adaptation to the Vietnamese context, the local community of language teachers and researchers were not satisfied with CEFR-V. This adapted framework was considered embryonic with most of the descriptors for language skills and can-do statements being closely translated from the CEFR English version (Nguyen & Hamid, 2015). However,

appraising the entire CEFR-V is outside the scope of this thesis. Since my research aims to validate the listening component of a locally created English proficiency test, it focuses only on the CEFR-V's descriptors for listening skill which will be found in Chapter 5.

Following the launch of Project 2020, fundamental changes in the language proficiency requirements for Vietnamese people were announced. Decision No.1400/QD-TTg dated September 30, 2008 stipulates that university students need to reach a certain level of proficiency as a requirement for graduation. This ruling applies also for English teachers at different levels of the national education system. According to Decree No.15/2014/TT-BGDĐT, those who wish to pursue post-graduate studies in Vietnam also need to gain proof of adequate English proficiency as both entry and exit requirements. Table 1.3 summarizes the language requirements for English teachers, undergraduates and post-graduate students in Vietnam as determined in the two documents mentioned above.

**Table 1.3.** English proficiency requirements in Vietnam

<b>Target groups</b>		<b>Proficiency levels in CEFR-V</b>
<b>Teachers</b>	English teachers at primary & secondary school	Level 4
	English teachers at high school	Level 5
	English teachers at university	Level 5
<b>Under-graduates</b>	Language-major students	Level 5 (C1-CEFR)
	Non-major students	Level 3 (B1-CEFR)
<b>Post-graduates</b>	Master	Level 3 (B1-CEFR) entry and exit requirements
	PhD	Level 3 (B1-CEFR) entry requirement Level 4 (B2-CEFR) exit requirement

It is important to note that Decision No.1400 and Decree No.15/2014 were issued without any guidance on which organizations candidates should go to if they wanted to gain proof of their English proficiency. In the absence of clear guidance, a great number of candidates were left to choose whether to take a costly international test (IELTS, TOEFL) or opt for a more affordable choice, the English proficiency tests developed by Vietnamese universities. With that being the case, Dunlea et al. (2018) emphasized that the quality of assessment became a widespread concern for both test-takers and test-score users. To address this issue, a new test was needed to help provide standardized, national measures of English proficiency (Dunlea et al., 2018). In March 2015, the Ministry of Education and Training (MOET) issued Decision No. 729/QD-BGDĐT on the localized test of English proficiency which was designed to suit the context of English language use in Vietnam.

### **1.1.2. The three-level test of English proficiency for Vietnamese people**

With the new regulations on English proficiency requirements (see Table 1.3), Decision No. 729/QD-BGDĐT stipulates that the English proficiency test for Vietnamese people targets level 3 to level 5 of CEFR-V and consists of four subtests for four language skills. It states the format that each of the subtests must follow. Unfortunately, a clear decision on which particular universities or organizations were responsible for the development and administration of this test was missing from this documentation. This lack of guidance, once again, brought about confusion and chaos in the assessment of foreign language proficiency in Vietnam after the decision took effect. Universities around the country were left to develop their own English proficiency tests in the absence of quality control measures. It was not until 2017 that the MOET issued Circular No. 23/2017/TT-BGDĐT, dated September 9, 2017, that introduced stricter regulations on the assessment of foreign language proficiency based on the CEFR-V. According to this document, the universities that offer majors in foreign languages are authorized to (a) develop and

administer foreign language proficiency tests and (b) award certificates to test-takers. Most recently, in May 2019, the MOET issued Official Document No. 538/QLCL-QLVBCC, naming eight universities as being eligible to develop and administer localized tests of language proficiency. This document also stipulates that only the certificates of foreign language proficiency awarded by these universities are valid and accepted in Vietnam.

The release of multiple documents as listed above can be understood as an effort from policy-makers to address the issue of misdirected and inconsistent assessment for language proficiency in Vietnam. However, the issue of quality management has not received much attention in those legal documents. In fact, there has not been any official guidance on validating these tests, especially when they are developed by different universities. The lack of validity evidence has become a significant concern for both test-takers and test-score users. It is also important to note that although this is a localized test which is intended for use solely in Vietnam, it claims alignment with the CEFR and aims to reach a level of comparability with international tests of English proficiency. This ambitious goal needs to be backed by evidence. Without validation research, there is no way to tell how the test would play out in terms of international benchmark of proficiency. This research project aims to partly address these issues by validating a listening test developed by one of the authorized universities.

## **1.2. Issues in assessing listening**

Listening is a vital skill in second language learning. Nunan (1998) emphasizes that more than 50% of the time that students spend functioning in a foreign language will be devoted to listening. This skill plays an active role in the language acquisition process (Brown, 2006) and a positive transfer takes place from listening to speaking (Krashen, 1981). Therefore, the assessment of listening is of paramount importance in evaluating the overall communicative ability of a second language learner. Yet, listening is, arguably, the most

difficult skill to assess (Field, 2013). The three major issues in assessing listening are (a) the lack of literature on listening and assessing listening, (b) the development of listening test tasks, and (c) the number of factors affecting second language listening. Let us look at each of these areas in turn.

### **1.2.1. The inadequacy of relevant literature**

As Buck (2001) emphasizes, listening is a very complex process and if we want to measure it, we must understand how that process works. For many years, listening has been an overlooked dimension of language acquisition (Feyton, 1991), remaining the ‘Cinderella skill’ in language learning and teaching (Nunan, 1997). The lack of adequate theories and models of listening made it one of the most elusive skills to describe (Brindley & Wigglesworth, 1997). Moreover, there has been relatively limited coverage on the assessment of listening ability in the language testing literature (Aryadoust, 2013; Brindley, 1998; Sawaki et al., 2009).

While a number of listening skill taxonomies (Buck & Tatsuoka, 1998; Goh & Aryadoust, 2015; Richards, 1983) have been proposed and used to form the basis of listening test development, little is known about the legitimacy of these taxonomies. In particular, these taxonomies attempt to arrange the subskills according to their potential cognitive load, assuming that certain subskills are more difficult than others. Such a hierarchy has not been adequately validated by research (Aryadoust, 2018) and the relationship between listening skill taxonomies and listening test tasks has not been explored very much. Research is still much needed to examine whether the skills described in a taxonomy are actually reflected in test performance (Brindley, 1998).

### **1.2.2. The development of listening test tasks**

Since listening is not directly observable, listening test tasks always involve some reading, writing and/or speaking to demonstrate one’s level of comprehension (Brunfaut,

2016). With that being the case, a major threat to the validity of a listening test is the presence of factors irrelevant to the listening construct which are often called construct-irrelevant variance. Frequently used formats in listening tasks include selected-response items such as multiple-choice questions or matching tasks, which are usually presented in writing. When using this format, test developers must always be aware of the amount of reading that the test items require listeners to undertake in order to answer the questions (Green, 2017). This means that the questions must be carefully worded so that the test-takers will not waste their time trying to understand the questions while simultaneously listening to the input and figuring out the answer. Although the selected-response format is usually favoured since it allows for straightforward and objective scoring, it is very difficult to write good selected-response items, especially multiple-choice questions (Hughes, 2003). In the meantime, constructed-response item types such as note-taking, gap-filling or short answer require test takers to produce an answer rather than selecting one from a range of options. This method makes test items more cognitively demanding (Field, 2013) because of the need to manipulate one's language in order to answer the questions (Green, 2017). In assessing listening, choosing the most appropriate test method is not always straightforward. Test developers must constantly be aware of the test construct and at the same time, the construct-irrelevant factors that certain test methods may induce.

A test developer needs to make a number of other important decisions related to the input text such as the length of the text, the speech rate, the nature of information, the accent, the sociolinguistic considerations, and the grammatical and lexical demand, (Elliott & Wilson, 2013). The characteristics of a task setting should also be carefully weighed up. Issues such as the test room, the time constraint, the number of times heard, the rubric or the technology involved in administering a listening test can all have an impact on test-takers' performance (Buck, 2001). Another important job of a test developer is to ensure that task

instructions are short, simple and clear so as to lessen test-takers' anxiety (Green, 2017). In general, developing a listening test task is a laborious process in which many aspects of the task need to be taken into consideration to make sure that it reflects the targeted construct and contains minimal construct-irrelevant variance. It takes time to develop good tasks that can provide a reliable and valid indicator of test-takers' listening ability.

### **1.2.3. Factors affecting second language listening**

Test-takers' performance on a listening task can be affected by a wide range of variables. Research on second language listening has found that the characteristics of the input texts and the test tasks have observable impacts on the difficulty of listening test items (Buck, 2001; Brindley, 1998; Griffiths, 1992; Jensen et al., 1997; Nissan et al., 1995; Rost, 1990; Rubin, 1994). The factors affecting the cognitive load of a listening test items include but are not limited to background knowledge, speech rate, lexical demand, accent, response format, or the lexical overlap between the text and the test questions. The test-takers and their unique characteristics can also lead to variances in their performance on a listening task. In order to reduce the risk of bias, it is important to consider the context of the test and the characteristics of the target candidature such as their age, gender, personality, emotional state, working memory, level of education or examination preparedness (Elliott & Wilson, 2013). That said, very little is known about their relative effects on levels of comprehension (Brindley, 1998; Bloomfield et al., 2010). Test developers may adjust the cognitive load of an item by manipulating these factors; however, the effect of this strategy has not been adequately researched (Bloomfield et al., 2010). Therefore, it is difficult to set the difficulty level of a certain item and explain why some items are more difficult than others.

In summary, the elusive nature of listening has made it a very difficult skill to describe and to assess. A number of overviews and studies on second language listening from the 1990s until recently (Aryadoust, 2018; Brindley, 1998; Buck, 2001; Goh & Aryadoust,



2015; Rost, 1990) all emphasize that a great deal of work still needs to be done to guide the teaching and testing of listening. This research is carried out with the view to shedding further light on the listening process and the extent to which the targeted listening construct is reflected by multiple-choice test items.

### **1.3. Validity and the argument-based approach to validation**

In order to conduct a validation study, an appropriate framework for validation is of great importance. Over the years, the evolution of validity theories has brought about the development of parallel frameworks for validation (Xi, 2008). In the 1950s and 1960s, validity was understood to have distinct classes and divided into three types, namely content, construct and criterion-related validity (Cronbach & Meehl, 1955; Smith, 2001). The validation process, at this time, was defined as conducting studies to collect evidence of one or more of these three types of validities (Aryadoust, 2013). This approach was once referred to as the one question, three validities approach to validation (Chapelle & Voss, 2013). It was not until the 1980s that a radically different view on validity was proposed. Messick (1989) emphasized that validity should be seen as a unitary concept. In his point of view, validity is not a property of the test itself but a property of the interpretation and use of test scores. Until now, Messick's theory of validity still remains highly influential (McNamara, 2006). However, it has been challenged over the years for its lack of guidance on how validation work should be done (Xi, 2008). Since validity is now a unitary concept, any information obtained in the development and use of a test can be used as evidence for a claim of validity (Anastasi, 1986). Furthermore, it is difficult to identify the beginning and the end of the validation process (Kane, 2012). Therefore, practitioners who wish to conduct validation research still need to be guided by a more transparent and manageable framework. To address this need, the argument-based framework was proposed as an appropriate approach to validating the interpretation and use of test scores.

The argument-based framework provides a systematic way to examine the inferences that underlie score interpretation and use so that a coherent validity argument can be developed (Chapelle et al., 2010; Kane, 2006, 2012). This approach is coherent with the modern view on validity and emphasizes that validity has a great deal to do with how test scores are interpreted and how a test is used (Wolming & Wikstrom, 2010). In second language testing, in general, a number of studies have applied the argument-based approach (Chapelle et al., 2008; Johnson & Riazi, 2016; Frost et al., 2012). However, in validating second language listening tests, in particular, the attempt to use the argument-based framework is still relatively limited. This research project will add more empirical evidence to help build a comprehensive and coherent validity argument for the assessment of listening.

Central to the argument-based approach are the concepts of interpretive argument and validity argument. In the discussion on how to apply the argument-based approach in validation studies, Kane (2012) emphasizes the development of an interpretive argument as the first step in the process. In an interpretive argument, specific claims are made about the desired uses and interpretations of test scores. Kane (1992) proposes three inferences which refer to the connections between the data and the claims. These are the evaluation, generalization and extrapolation inferences. Later on, Chapelle et al. (2008) argued for the existence of the explanation inference to bridge the gap between the generalization and extrapolation inference. The final inference, utilization, was proposed by Bachman (2005). These five inferences can be briefly explained as follows:

- The evaluation inference links observed performance to observed scores. It typically makes assumptions about the appropriateness of the rubrics, the rules for combining scores or the statistical characteristics of test items.

- The generalization inference links the observed performance on a particular set of tasks to expected performance in a larger domain of tasks over several occasions and conditions of observation.
- The explanation inference links the observed performance to the test construct. It rests on the assumption that the targeted construct accounts for test-takers' observed performance.
- The extrapolation inference extends the interpretation into new performance domains. It usually links the observed performance with the expected future performance on different kinds of tasks in different contexts (including non-test, real-life contexts).
- The utilization inference focuses on the relevance between a test and its intended use. It links score-based interpretations to score-based decisions and test consequences.

Kane (1992, 2002) points out that the inferences in an interpretive argument cannot be proven. Instead, they can only be evaluated in terms of how convincing they are in light of available evidence. It is in this evaluation stage that the validity argument is developed.

The validity argument seeks evidence that the interpretive argument is coherent, clear and plausible (Kane, 2006). It is important to note that each inference of the interpretive argument requires different lines of evidence. For example, the evaluation inference can be supported by expert judgement on the rubric or the data regarding rater consistency while the generalization inference needs evidence of test reliability (Kane, 2013). Kane (2013) emphasizes that some inferences may be more questionable than the others and these inferences might require several parallel lines of evidence. The empirical validation methods used in the validity argument are many and varied, including both quantitative and qualitative analyses. These methods include but are not limited to item analysis, reliability analysis, correlational analysis, questionnaire, interviews, observation or think-aloud protocols (Xi, 2008). Kane (2006) clarifies that individual studies in a validity argument may focus on

statistical analysis or content analysis but the validity argument as a whole requires the integration of different kinds of evidence from different sources.

In general, the argument-based approach to validation is straightforward with two steps: state the claims being made (interpretive argument) and evaluate the plausibility of these claims (validity argument) (Kane, 2013). Although all the inferences in the interpretive argument deserve attention, it might be more useful to focus on the most questionable ones. Studies on these questionable inferences are potentially most informative since they help to address the weakest link in the argument (Cronbach, 1988). Moreover, if empirical studies are to be conducted to evaluate every inference, validation would be a never-ending process. Validation effort, therefore, should focus on the most questionable parts of the interpretive argument (Crooks et al., 1996).

#### **1.4. Overall structure and aims of the research**

This study aims to investigate the validity of the LCLT, a listening test developed by a university in Vietnam. This university is one of the most prestigious institutions in the country in terms of its English language program. Therefore, it has always been considered a reliable provider of proficiency tests for those who wish to obtain proof of their English proficiency levels. Since 2015, this university has developed and administered a great number of English proficiency tests, following the MOET-approved format and CEFR-V. Up until the current study, they had never sought to validate any of their English tests. This is also the case for other universities which were authorized to design and organize the same kind of test. Until recently, the number of searchable publications, Master or PhD theses that venture into the validation of localized English proficiency tests in Vietnam has been very limited. This lack of validation effort is an issue that needs to be urgently addressed to ensure that Vietnamese test-takers can really benefit from a well-developed and valid high-stakes assessment instrument.

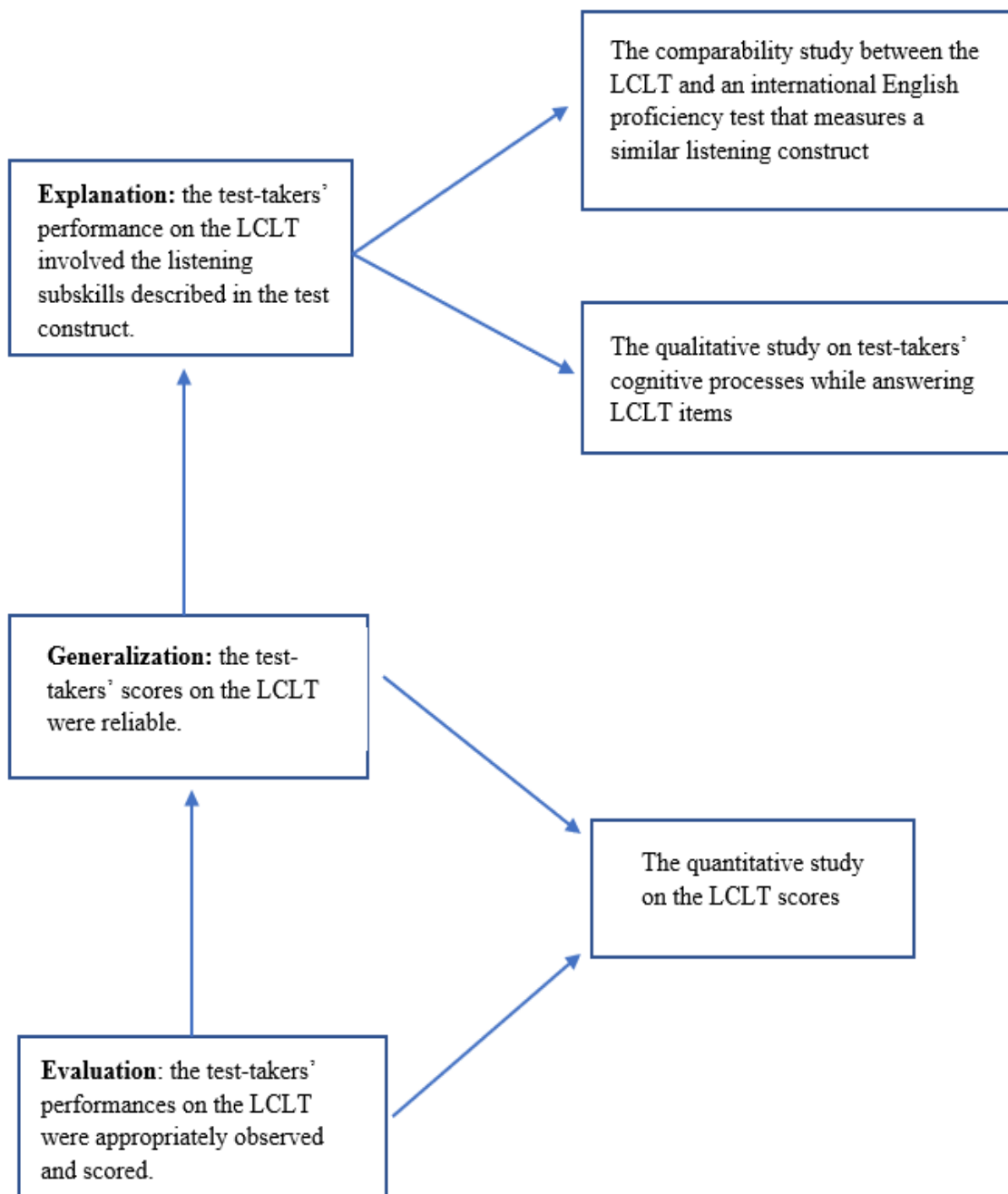
This present study investigates the validity of the LCLT, using the argument-based approach to validation. Five inferences for the LCLT can be presented as follows:

- **Evaluation inference:** the test-takers' performances on the LCLT were appropriately observed and scored.
- **Generalization inference:** the test-takers' scores on the LCLT were reliable.
- **Explanation inference:** the test-takers' performance on the LCLT involved the listening subskills described in the test construct.
- **Extrapolation inference:** the test-takers' performance on the LCLT can predict their future performance on different kinds of tasks in real-life contexts (teaching, undergraduate and post-graduate studies).
- **Utilization:** the test-takers' scores on the LCLT enables proper decision making with regard to graduation for undergraduates, admission to post-graduate studies, and employment for English teachers in Vietnam.

Within the limit of this research, the argument-based validation of the LCLT focuses only on the first three inferences. Evaluating the extrapolation and utilization inferences requires a great deal of quantitative and qualitative data from multiple sources, which will go far beyond the scope of this thesis. These inferences are, therefore saved for future research.

Three different studies are conducted to obtain evidence necessary for the assessment of the evaluation, generalization, and explanation inferences. The first study investigates the statistical characteristics of the LCLT scores, focusing on the evaluation and generalization inference. In this study, Rasch measurement is the main analytical tool. The second study is a qualitative study on the cognitive processes that test-takers undergo while answering the LCLT items. The data regarding test-takers' thinking process is collected by means of think-aloud protocols. The third study examines the extent to which test-takers' scores on the LCLT correlate with their scores on an international English proficiency test that measures a

similar listening construct. Content analysis, statistical analysis of test scores and correlational analysis are the main methods used in this study. Both the second and third studies are carried out to elicit evidence for the explanation inference. The overall structure of this research is presented in Figure 1.1.



**Figure 1.1.** Overall structure of the research

Once the three studies (see Figure 1.1) are conducted, all the necessary findings will be put together so that judgements can be made about whether the inferences are supported by evidence. These judgments make it possible to arrive at the final decision on the validity of the LCLT.

### **1.5. Ethical considerations**

Since this research uses confidential data provided by a university in Vietnam, certain ethical requirements must be strictly followed. First, this university is not to be named. Throughout the thesis, it will only be referred to as the host university. Second, confidential data such as the test questions, audio recording, and test-takers' answer sheets will not be fully disclosed. In chapter 4 and 5, certain test items, the test-takers' answers to these items, and relevant parts of the audio recording are presented but only for the sake of data analysis. Third, the real name of the localized English proficiency test in Vietnam will not be mentioned. Instead, it is referred to as the English proficiency test for Vietnamese people and the listening component is called the Locally Created Listening Test (LCLT).

The request for permission to access data in both English and Vietnamese is provided in Appendix A and B. All the information sheets and consent forms can be found in Appendix C-H. In these appendices, any information that might identify the host university, the people involved in the provision of research data, and the research participants will be hidden.

### **1.6. Significance of the research**

This research is one of the first attempts to conduct a multi-perspective investigation on the validity of a high-stakes listening test developed for Vietnamese learners of English. The insights gained from this project are, first and foremost, meaningful for the host university in the development of their future listening tests. At the same time, they call for further validation research on not only the listening component but also the reading, writing

and speaking component of the English proficiency tests that the host university has developed. It is hoped that the results of this research will also be useful for test developers from other universities in Vietnam.

By focusing on a listening test, this research contributes to the growing but still limited body of research on the assessment of second language listening. Apart from investigating the statistical characteristics of test scores, a major part of this research is dedicated to examining the extent to which the test items reflect the listening construct. Such an investigation contributes to the limited amount of research on how the multiple-choice format manipulates the way test-takers tackle listening tasks. While a number of other factors such as speech rate, vocabulary or accent are believed to affect the difficulty of items in a listening test, very little is known about the impact they have on the listeners' thinking process. This research bridges that gap with evidence of the test-takers' cognitive behaviours in response to certain text and task-related factors. From there, it helps to map some of the ways in which test developers can create good items and effectively control item difficulty.

It is hoped that this research will contribute to a deeper understanding of the argument-based approach in language test validation. The argument-based validation of the LCLT will lead to a critical reflection on both the benefits and challenges of the argument-based framework. Such insights, which are hard to obtain without practical experience, could help those who wish to use this framework to shape and guide their research in the future.

Finally, this research highlights important issues in the localization of English proficiency tests, a relatively new trend in language testing and assessment. It provides insights which will help the developers of current localized English tests reflect on the effectiveness of their work. Policy-makers and test-developers who wish to venture into test localization can also benefit from the practical suggestions on how to develop an English proficiency test that effectively reflects the local context.



## **1.7. Organization of the thesis**

The present thesis consists of seven chapters including this introductory chapter. Chapter 2 provides an overview of listening in a second language, the development of validity theories and the argument-based approach to validation. Chapter 3 presents the overarching validation framework that shapes this entire research as well as the overall methodology and research design. Chapter 4 reports on the quantitative study that investigates the statistical characteristics of the LCLT scores. In Chapter 5, the qualitative study on the test-takers' cognitive processes will be presented. Chapter 6 focuses on the comparability study between the LCLT and the listening component of an international English proficiency test. Chapter 7 summarizes the main findings and presents the judgement on the validity argument for the LCLT. This chapter also provides an overall discussion on the major research findings. The thesis ends with Chapter 8 which includes a summary of key contents, acknowledgement of limitations, and suggestions for future research.

## **Chapter 2. Literature Review**

The goal of this study is to validate a locally created listening test (LCLT) in Vietnam, following the argument-based approach. To lay out the theoretical background for the study, this chapter starts with an overview of listening in a second language, principally from a cognitive point of view. As the chapter progresses, its focus shifts to the assessment of second language listening. Another important part of the chapter is dedicated to a chronological review of validity theories and validation frameworks. In this part, the argument-based approach to validation is described in terms of its structure, its major strengths, and its application in language testing and assessment. The rest of the chapter then elucidates how the argument-based approach is used in this research to shape the validation framework for the LCLT.

### **2.1. Listening in a second language**

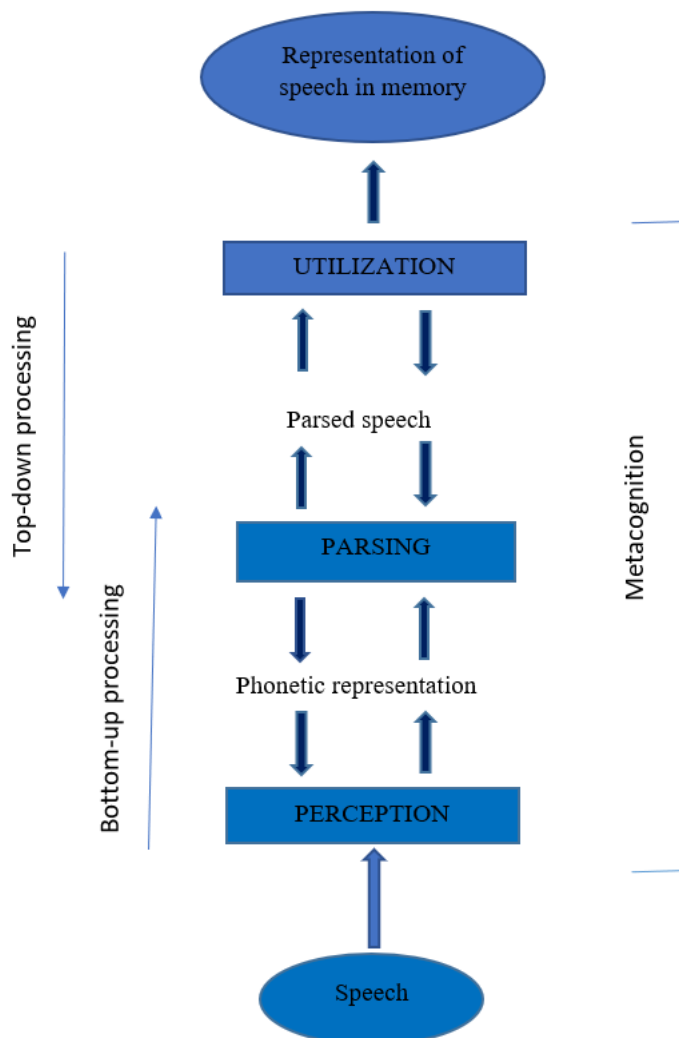
Listening is a complex process in which “the listener takes the incoming data, the acoustic signal, and interprets that, using a wide variety of information and knowledge, for a particular communicative purpose” (Buck, 2001, p. 29). In a person’s first language (L1), listening is the natural precursor to speaking and it plays a significant role in the early stages of language development (Nation & Newton, 2009). In second language (L2) learning, listening also plays a very important role since it is considered the way of learning a language (Nord, 1980) or a condition necessary for language learning to occur (Hasan, 2000; Krashen, 1981; Newmark, 1981). Outside of the language learning context, listening has a key role in a wide variety of communication settings. Nunan (1998) emphasizes that more than 50% of the time that students spend functioning in a foreign language will be devoted to listening. Despite its importance, listening is often considered the most difficult skill to learn (Hasan, 2000; Graham, 2003). One possible explanation for this might be that learners are not taught how to learn listening effectively (Vandergrift, 2007). The narrow focus on answering

questions does very little to help learners understand and control the process leading to comprehension (Kurita, 2012). In terms of assessment, listening is arguably the most difficult language skill to assess due to its elusive nature (Field, 2013). It also remains the least researched of all four language skills (Vandergrift, 2007). That said, the existing literature on L2 listening has addressed, to a certain extent, various important issues among which the most relevant to this study include the listening processes, the construct of L2 listening, and factors affecting the difficulty of listening tasks.

Much of our current understanding of the cognitive processes involved in listening comprehension has been illuminated by theoretical perspectives from cognitive psychology (Celce-Murcia et al., 2013). A classic and much quoted model of listening proposed by Anderson (1995) outlines three recurrent and overlapping phases which are perception, parsing, and utilization. Perception is the stage at which listeners match the sound they hear to words they know. It is referred to as a process in which L2 listeners decode, segment words in a stream of speech (Richards, 1983; Cutler & Clifton, 1999). The decoded words are almost simultaneously related to the wider context in which they occur according to grammar or lexical cues in a process known as parsing (Anderson, 1995). In the last stage, utilization, all the information processed at phonological, grammatical and lexical levels is related to the listeners' prior knowledge of the facts and listening context to enable them to construct the meaning and interpret the functions of the auditory input (Goh, 2014).

Cognitive processes involved in listening are complex but they can occur harmoniously as sound signals interact with listeners' prior knowledge and as different knowledge sources are referred to and connected through the process (Goh, 2014). An early and still popular distinction in the listening literature is between bottom-up and top-down processing (Aryadoust, 2013). These terms are used to mark the difference between information derived from the sounds that have been decoded (bottom-up) and information

derived from the application of prior knowledge about the target language and the world (top-down) (Field, 2004). Figure 2.1 illustrates the interrelationships between these two processes in L2 listening.



**Figure 2.1.** Cognitive processes in L2 listening and their interrelationship (Goh, 2014, p.76)

Both bottom-up and top-down processes are necessary for L2 listening and in order to achieve success, listeners need to be able to engage in both sets of processes effectively.

Vandergrift and Goh (2012) assert that these processes are directed by the listeners' metacognition which is the ability to think about the processes and manage them. The next sections deal with bottom-up and top-down listening in more detail.

### 2.1.1. Bottom-up processes

Bottom-up listening involves the use of aural stimuli, such as phonemes, syllables, sounds to construct larger units like words and the grammatical relations between words (Morley, 2001). Such processes, as described by Field (2003), are adopted by listeners as they assemble the message piece-by-piece from the speech stream, going from the part to the whole. The bottom-up approach sees language comprehension as a process of consecutive stages and processing occurs on a number of levels from low to high, in serial order (Buck, 2001). The output of each level becomes the input for the next level. More specifically, the acoustic input would be decoded into phonemes first, then phonemes are put together to identify words, then words are put together to be considered at syntactic level and next comes the stage where meaning is constructed. This way of processing aural input will be useful when listeners attempt to understand how much something costs, what the flight number is or how someone's name is spelt.

In the first stage of listening, listeners receive the auditory input and decode it, using their knowledge of not only phonemes but also syllables and suprasegmental features of speech. Specifically, listeners need to transform the phonemes, the smallest segment in which spoken language can be sequentially described, into groups of syllables (Cutler & Clifton, 1999). In a language that features lexical stress, listeners also need to go a step further by identifying which syllable carries the stress. This is really important for a language like English since it serves as an important cue to word recognition (Grosjean & Gee, 1987). For example, the stressed syllable /pju/ provides the cue for *comPUter*, or /tɪə/ the cue for *volunTEER*. Geranpayeh and Taylor (2013) emphasize that “the output of decoding is not a string of phonemes but a string divided into syllables that are marked, where appropriate, for relative stress” (p. 98).

The next level of bottom-up listening is lexical search, a process in which an auditory stimulus is matched against a number of phonetic and lexical representations (Foster, 1976). This process is assumed to involve the identification of a number of likely matches for a given stretch of signals (Field, 2013). The phonological cues such as the number of syllables, stress or tone can help the listener narrow down the options and figure out the best match. A typical example is when learners of English listen to numbers. When they hear a string of phonemes that include /f/, /l/, /f/, /t/, /i/, they will naturally think of two possibilities, either fifteen or fifty. In this case, the placement of stress and the presence (or absence) of an ending consonant are the key factors that help them identify the correct number.

At this level, word recognition is a major problem for listeners, especially non-native listeners. While in reading texts, the boundaries between written words are explicitly marked, this is not the case for listening. In fact, what listeners hear from the auditory input is not separate words but connected speech in which they must determine the boundaries between words. The fact that the standard citation forms of words are modified when they occur in connected speech makes lexical segmentation even harder for L2 listeners (Field, 2003). The modifications of pronunciation such as assimilation, elision or intrusion which take place during fast, informal speech significantly reduces comprehension for second language learners (Henrichsen, 1984). Buck (2001) emphasizes that even high-level L2 listeners can still fail to recognize words in a language that they know very well.

Beyond the recognition of single words, listeners also need to comprehend groups of words by putting them together into familiar chunks frequently encountered in the target language (Field, 2008). This process is called parsing which is aided by syntactic clues. An understanding of standard word order in the target language is essential at this stage. Buck (2001) points out that, in English, some structures are generally more difficult to process than others. For example, it takes more time to process negative statements than positive

statements, and learners will process passive statements more slowly than active ones. Apart from the syntactic cues, semantic information can also assist parsing. If an idea unit is plausible and explicitly articulated, catching the key content words may be enough to process meaning (Buck, 2001). Realistically though, messages from the auditory input are not always straightforward. In many cases, the meaning of the input has to be inferred and constructed from the application of prior knowledge about the world as well as the language. It is in these situations that top-down processing comes into play to help listeners make sense of what they hear.

### **2.1.2. Top-down processes**

Top-down listening is the process in which meaning is derived from the listener's prior knowledge and is associated with the aural stimuli (Morley, 2001). This process is crucial for meaning construction and often considered to be at a higher level than bottom-up processing. When engaged in top-down listening, L2 listeners associate aural stimuli with their schemata, situational knowledge, contextual knowledge, discourse knowledge as well as their own experiences and attitudes (Goh, 1997, 1998, 2008). Oftentimes, the surface meaning of the speakers' words is not enough to convey the significance or implications of what is said. Therefore, listeners need to employ the prior knowledge stored in their memory to successfully make sense of what they hear. This sense-making process is greatly facilitated by pragmatic knowledge, world knowledge and discourse knowledge.

As noted by Goh (2014), pragmatic knowledge is crucial during the utilization phase where the listener interprets what is said and constructs appropriate responses. The interpretation of the speaker's illocutionary intentions goes beyond literal meaning drawn from successful decoding and parsing of the input. Garcia (2004) stresses the importance of pragmatic knowledge in listening, explaining that this type of knowledge enables listeners to:

- understand a speaker's intentions;

- interpret a speaker's feelings and attitudes;
- differentiate speech act meaning such as the difference between directive and commissive;
- evaluate the intensity of a speaker's meaning such as the difference between a suggestion and a warning;
- recognize sarcasm, joking, and other facetious behaviour;
- respond appropriately. (p. 1-2)

A number of studies on L2 listeners' ability to understand pragmatic meaning found significance differences between high and low proficiency listeners (Cook & Liddicoat, 2002; Garcia, 2004; Kasper, 1984). A possible reason is that higher-proficiency learners can process both linguistic knowledge and pragmatic knowledge in listening since they have achieved a higher level of language processing automaticity (Cook & Liddicoat, 2002). Meanwhile, for lower-level proficiency learners who tend to rely more on bottom-up processing of linguistic information, the ability to comprehend pragmatic meaning can be problematic (Bardovi-Harlig, 1999). More recently, test designers have incorporated items that target pragmatic understanding in the listening sections of major standardized exams such as the IELTS or TOEFL since they fathom the crucial role of pragmatic knowledge in measuring genuine language abilities of learners (Allami & Aghajari, 2014).

World knowledge is usually referred to as schemata which are assumed to exist for most things we would want to represent in our memory including concepts, events, sequences of events, action and sequences of actions (Buck, 2001). This type of knowledge is non-linguistic but it emerges as a powerful factor affecting listening. Specifically, if the topic of the input text is familiar to the listener, it is easier for him/her to make appropriate inferences and thus achieve successful comprehension of the text. At the same time, the lack of relevant knowledge to the content of the text will make it more difficult to make inferences and hinder



comprehension. A number of studies have showed that topic familiarity does facilitate listening (Chiang & Dunkel, 1992; Markham & Latham, 1987; Schmidt-Rinehart, 1994; Sadighi & Zare, 2006). Bacon (1992) found that successful listeners make effective use of world knowledge particularly during the utilization phase.

Listening takes place in a wide variety of contexts such as work places, schools, shops or hospitals and in each of these contexts, listeners are expected to experience a different kind of discourse. Goh (2014) points out that each type of discourse has a different pattern and the knowledge about how specific discourses are structured can enhance second language listening. In listening to lectures, for example, listeners will be more likely to understand the input if they understand the role of discourse markers and how a lecture typically begins, develops and ends (Flowerdew & Miller, 2014). Likewise, the knowledge of other discourse types such as comedy or radio news can help the listeners anticipate the context of the texts. If an L2 learner knows the genre of the aural input before listening, they will become more active and able to use their prior discourse knowledge effectively.

The key process in top-down listening is inferencing (Nation & Newton, 2008) and the inferences that listeners make depend largely on their knowledge of the world, of pragmatics and discourse. Apart from that, their understanding of the context surrounding the input text and the characteristics of the speakers can also contribute to facilitating the inference-making process.

It should be pointed out that although bottom-up and top-down listening are usually described as two distinctive processes of listening, they do not take place separately and in any particular order (Buck, 2001). Instead, the overall listening process can be described as a dynamic and interactive interplay between the bottom-up and top-down processes (Field, 2004; Vandergrift, 2011). Specifically, listeners have to decode the auditory input and at the same time, make use of other resources such as their world knowledge or the knowledge

about the context in which an utterance is made (Hulstijn, 2003; Vandergrift, 2007). Moreover, they need to be able to simultaneously process speech and refer to different sources of knowledge in real time (Rost, 2005). The overall listening process in a second language is, therefore, truly complex (Vandergrift, 2011) and assessing this skill is, without a doubt, a major challenge for language testers. This research sets out to investigate the ways in which bottom-up and top-down processes interact with each other as test-takers answer comprehension questions. Further, it seeks to analyse the extent to which the design of test items affects listeners' use of these two processes.

## **2.2. The assessment of second language listening**

Over the last couple of decades, the main approach to assessing listening has been testing it as an isolated target skill (Brunfaut, 2016) and a listening test is supposed to measure a particular listening construct. The definition of such a construct is, therefore, important for the assessment of listening. This section reviews several attempts in defining the listening construct as a taxonomy of listening subskills. It also discusses the factors affecting listening task difficulty as well as construct-irrelevant variance that can contaminate the construct and prevent test takers from utilizing their listening ability.

### **2.2.1. Defining the listening construct**

The understanding of the construct, i.e. what is supposed to be measured by a particular test, is the starting point for test development (Buck, 2001). In the early 20<sup>th</sup> century, construct definition in listening tests tended to be implicit; however, as the approaches to assessing listening evolved over time, the process of construct definition became more sophisticated and explicit (Taylor & Geranpayeh, 2011). The following section discusses speculative taxonomies, research-based taxonomy, and the action-oriented description of listening subskills in the Common European Framework of Reference (CEFR).

### 2.2.1.1. Speculative taxonomy

With the view to assisting the teaching and testing of L2 listening, numerous taxonomies have been proposed to describe the hierarchy of listening subskills. The simplest taxonomy is the one proposed by Carroll (1972) on the basis of the two-stage view which divides listening into two processes:

- apprehending linguistic information (lower level processing)
- relating the information to a broader context (macro-comprehension).

A similar view is also shared by Oakeshott-Taylor (1977) and Hughes (2003) when they attempted to describe listening at both micro and macro levels.

A more comprehensive list of listening subskills is the one proposed by Richards (1983). This list, which was developed from need analysis, discourse analysis and other related research, consists of 32 sub-skills in conversational listening such as the ability to:

- discriminate among the distinctive sounds of the target language;
- recognize stress pattern of words;
- identify words in stressed and unstressed positions;
- guess the meaning of words from contexts in which they occur.

Apart from conversational listening, Richards (1983) also proposed a set of sub-skills relevant to academic listening such as the ability to:

- identify the purpose and scope of a lecture;
- identify topic of lecture and follow topic development;
- infer relationships (e.g: cause, effect, conclusion);
- detect the attitude of the speaker toward the subject matter.

Comprehensive as it was, Richards' (1983) taxonomy of listening was found to be speculative and its validity was not supported by empirical evidence. Taylor and Geranpayeh (2011) also

point out that Richards' specification of academic listening skills reflects only the knowledge and abilities needed in a traditional university lecture listening context.

In Weir's (1993) framework for testing listening, direct meaning comprehension and inferred meaning comprehension are the key components, each of which is specified as follows:

#### Direct meaning comprehension

- listening for gist
- listening for main idea(s) or important information; includes tracing the development of an argument, distinguishing the main idea(s) from supporting detail, differentiating statement from example, differentiating a proposition from its argument, distinguishing fact from opinion when clearly marked.
- listening for specifics; involves recall of important details
- determining speaker's attitude/intentions toward listener/topic (persuasion/explanation) where obvious from the text

#### Inferred meaning comprehension

- making inferences and deductions; evaluating content in terms of information clearly available from the text
- relating utterances to the social and situational context in which they are made
- recognising the communicative function of utterances
- deducing meaning of unfamiliar lexical items from context

Apart from these two principal components, Weir (1993) also laid stress on two other relevant aspects which are contributory: meaning comprehension and note-taking. While the former involves linguistic knowledge of the target language such as phonological features, grammar, syntax, lexicology and discourse, the latter involves the ability to extract important points and summarize the input text.

Beside the taxonomies proposed by Richards (1983) and Weir (1993), several other researchers in the field have offered lists of general listening subskills (Buck, 2001; Hughes, 2003; Munby, 1978; Rost, 1994). The common approach shared among these authors is to package listening as a collection of subskills, which might be fairly easy to operationalize and implement for assessment purposes (Aryadoust, 2018). However, these taxonomies are mostly intuitively derived and not supported by empirical evidence (Taylor & Geranpayeh, 2011). As pointed out by Buck & Tatsuoka (1998), they will still remain hypotheses until they are subjected to some form of empirical validation.

### **2.2.1.2. Research-based taxonomies**

Unlike speculative taxonomies, research-based taxonomies have their roots in empirical studies. The description of listening abilities proposed by Buck and his associates in the 1990s was among the most prominent works. A study by Buck et al. (1997) sought to examine L2 listening from a subskill-based approach. It looked at 30 multiple-choice items from a TOEIC test and found 14 abilities which are considered most important for test-takers' performance. For example:

- the ability to process faster input
- the ability to process lower-frequency vocabulary
- the ability to process more complex structures
- the ability to synthesize scattered information
- the ability to use word-matching strategies

In a similar study, Buck and Tatsuoka (1998) investigated a test with 35 short-answer comprehension questions, taken by 412 Japanese college students. They identified 15 listening attributes which were categorized into five major groups as shown in Table 2.1.

**Table 2.1.** Cognitive attributes of L2 listening (Buck & Tatsuoka, 1998, p. 141-142)

<b>Task identification attributes</b>	The ability to identify the task by determining what type of information to search for in order to complete the task
<b>Context attributes</b>	The ability to scan fast spoken texts, automatically and in real time
	The ability to process a large information load
	The ability to process a medium information load
	The ability to process dense information
<b>Information location attributes</b>	The ability to use previous items to help information location
	The ability to identify relevant information without any explicit marker to indicate that
<b>Information processing attributes</b>	The ability to understand and utilize heavy stress
	The ability to process fast texts automatically
	The ability to make text-based inferences
	The ability to incorporate background knowledge into text processing
	The ability to process L2 concepts which have no literal equivalence in L1
	The ability to recognize and use redundant information
	The ability to process information processed throughout a text
<b>Response construction attributes</b>	The ability to construct response quickly and effectively

Despite the empirical basis, the findings of the above-mentioned study seem to be limited in generalizability since they look at the listening abilities needed to tackle particular test tasks, namely multiple-choice questions and short answers, rather than listening in general.

A more recent study by Goh and Aryadoust (2015) examined the divisibility of listening subskills. By means of factor analysis, this study found evidence for the divisibility of the five following subskills:

- understanding and responding to the unexpected statements and/or questions;

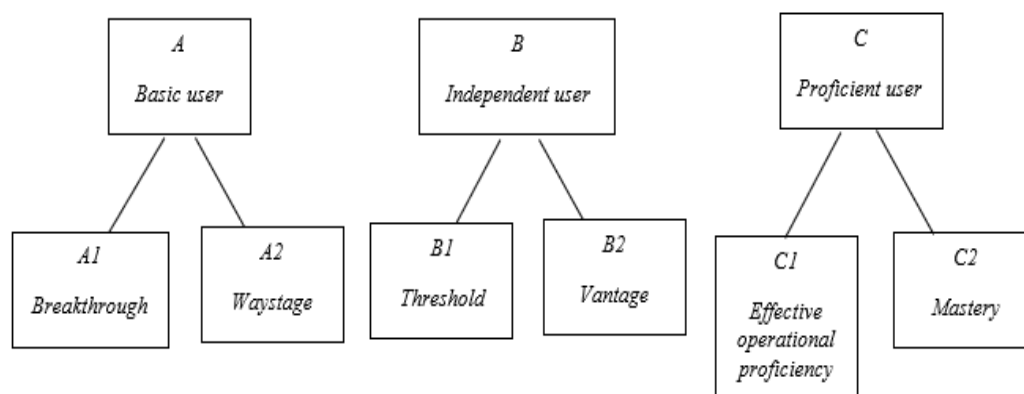
- understanding details and explicit information;
- making propositional inferences;
- making enabling inferences; and
- drawing conclusions.

Although statistical evidence showed that listening could be divided into subskills, Goh and Aryadoust (2015) still acknowledged that the identified subskills might not operate in isolation but in unison to achieve a listening goal.

Taken together, the taxonomies discussed so far have showed that listening is a multi-faceted process which consists of a large number of subcomponents (Buck, 2001). While speculative taxonomies seem detailed and systematic, they still remain hypotheses that need empirical validation. Research-based taxonomies, on the other hand, are supported by statistical evidence but the lists of subskills they offer are tentative and also need to be further validated. In both types of taxonomy, the hierarchy of listening subskills has not been successfully established. An attempt to develop such a hierarchy is found in the set of descriptors for listening in the Common European Framework of Reference (CEFR), which is the focus of the following section.

### **2.2.1.3. Descriptors for listening in the Common European Framework of Reference**

During the 1990s, another approach to identifying language proficiency, in general, and listening proficiency, in particular, emerged with ability levels expressed through illustrative descriptors (Taylor & Geranpayeh, 2011). The Common European Framework of Reference (CEFR) takes the action-oriented approach and clearly defines six levels of proficiency which can be used in learning, teaching as well as assessment. These six levels can be broadly described in Figure 2.2.



**Figure 2.2.** Six levels of proficiency in the CEFR

From Figure 2.2, it can be seen that the six levels of proficiency are respectively higher and lower interpretations of the classic division into basic, intermediate and advanced (Council of Europe, 2001). For the listening skill, in particular, there are general descriptors for listening at each level (see Table 2.2). Apart from this overall description, there are also specific descriptors for listening in a number of contexts which include:

- listening to public announcements (information, instructions, warnings, etc.);
- listening to media (radio, TV, recordings, cinema);
- listening as a member of a live audience (theatre, public meetings, public lectures, entertainments, etc.);
- listening to overheard conversations.

In each case, the language user is expected to listen:

- for gist;
- for specific information;
- for detailed understanding;
- for implications.

Taking the action-oriented approach, the CEFR attempts to describe listening ability through “can do” statements which indicate the behaviours expected for each level.



**Table 2.2.** Overall listening comprehension in the CEFR (Council of Europe, 2001, p.66)

C2	Has no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed
C1	Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details if the accent is unfamiliar.
	Can recognize a wide range of idiomatic expressions and colloquialisms, appreciating register shifts.
	Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly.
B2	Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic and vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influences the ability to understand.
	Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation.
	Can follow extended speech and complex line of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.
B1	Can understand straightforward factual information about common everyday or job-related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent.
	Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.
A2	Can understand enough to be able to meet needed of a concrete type provided speech is clearly and slowly articulated.
	Can understand phrases and expressions related to areas of most immediate priority (e.g: very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated.
A1	Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.

The rapid uptake of CEFR has been observed within and even beyond Europe (Read, 2019). In Asian countries like China, Japan, Taiwan, and Vietnam, CEFR has been used mostly for assessment purposes (Cheung, 2012; Jin et al., 2017; Neigishi & Tono, 2014; Nguyen & Hamid, 2015). In Vietnam, CEFR scale descriptors for all four language skills have been adapted to suit the local context and the Vietnamese six-level framework of reference for foreign languages is the outcome of this adaptation process (Nguyen & Hamid, 2015). The LCLT is designed based on the descriptors for listening in the CEFR-V which will be presented in Chapter 4.

While the LCLT targets listening abilities from level 3 to level 5 of the CEFR-V (equivalent to level B1 to C1 of the CEFR), there is no guarantee that the test items are able to engage the listening abilities associated with these levels. Introspective studies are therefore needed to examine the extent to which test items actually bring out the expected behaviours. The current research ventures into such an investigation by means of listeners' verbal report in which test-takers explain their thinking process as they answer the test questions.

### **2.2.2. Factors affecting listening task difficulty**

The cognitive demand of a listening task is governed by not only the listening subskill(s) it targets but also numerous other factors. A wide range of factors have been suggested to affect the processing demands of listening tasks, thus causing variances in test-takers' performance. The following list of factors affecting listening task difficulty (Table 2.3) is put together from a number different sources including Buck (2001), Buck and Tatsuoka (1998), Brindley (1998), Brunfaut (2016), Griffiths (1992), Jensen et al. (1997), Nissan et al. (1995), Rost (1990), Rubin (1994).

**Table 2.3.** Factors affecting listening task difficulty

<b>Factors affecting listening task difficulty</b>	<b>Components of the factor</b>
Characteristics of the input text	Speech rate, length of input texts, vocabulary, syntactic complexity, discourse structure, accent, explicitness, number of speakers,
Characteristics of the test tasks	Amount of context provided, response format, clarity of instructions, the amount of lexical overlap between the text and the test questions, response length, number of times listening
Characteristics of the listeners	Level of proficiency, background knowledge, working memory, motivation, metacognition

As seen from Table 2.3, factors affecting listening tasks difficulty can be divided into three groups, namely characteristics of the input texts, characteristics of the test tasks, and characteristics of the listeners. Providing an overview of all these factors is beyond the scope of this section. Instead, the focus here narrows down to the two factors that are most relevant to this particular research. Since the LCLT is a multiple-choice test, the impact of this format on the test-takers' performance and thinking process is an intriguing matter. Besides, the amount of lexical overlap between the input text and the response options of a multiple-choice item is also worth consideration. This section therefore synthesizes previous research on the use of multiple-choice questions in listening tests and how lexical overlap affects item difficulty.

#### **2.2.2.1. Multiple-choice format**

Multiple-choice questions (MCQs) have been widely used in tests of listening ability and the impacts of this format on test-takers' performance has been the focus of a number of research studies (Brindley & Slatyer, 2002). In general, MCQs, which require test-takers to

select from given response options, have been shown to be easier than items that require some writing. Berne (1992) found that subjects performed much better on MCQs than on open-ended or cloze tasks. In a study that compared four task types (MCQ, choose-the-picture, vocabulary list and Wh-questions) used in assessing comprehension of authentic texts in French, Eykyn (1992) found that, for beginning learners of French, MCQs helped comprehension and recall the most. Cheng (2004) investigated the differences in the test-takers' performance when three types of responses (traditional MCQs, multiple-choice cloze and open-ended questions) were utilized to measure L2 listening. She found that her subjects, 159 Taiwanese college students, performed best on the multiple-choice cloze and worst on open-ended questions.

More recent research casts light on the impacts of different variations of MCQs on test-takers' performance. Yanagawa & Green (2008) studied the effects of three different formats of multiple-choice questions which were:

- Full question preview (FQP): Both the item stem and response options are showed prior to listening
- Answer option preview (AOP): Answer options are displayed prior to listening but the questions are heard after the text.
- Question stem preview (QSP): Only the item stem is provided before listening.

Answer options are heard after the text.

Yanagawa and Green (2008) found that the group of students who were allowed to preview answer options (AOP) performed significantly worse than the other two groups. There was no significant difference between the students who were in FQP and QSP conditions. Along the similar lines, Hemmati and Ghaderi (2014) investigated the effects of four formats of MCQ which include FQP, AOP, QSP and NP (no preview) on the listening comprehension of EFL learners. The results showed that there was a significant difference between NP and the other

three variations of MCQs, which suggests that previewing MCQs can facilitate comprehension. Chang and Read (2013) investigated the effects of traditional written MCQs and oral MCQs on L2 listeners' performance. The results showed that lower proficiency students scored significantly higher in the written mode than in the oral mode in which no written clue was provided. Meanwhile, higher proficiency students performed slightly better in the oral mode. These findings imply that previewing MCQ might be more useful for lower proficiency listeners than for higher proficiency ones.

The research reviewed previously has showed evidence of the facilitating effects that MCQs have on L2 listening. Underwood (1989) argues that the opportunity to preview MCQs before listening helps to reduce stress of hearing something differently, refreshing listeners' memories and getting listeners ready for the tasks. Buck (1991) believes that previewing questions helps listeners understand the purpose of the task and motivates them to identify necessary information to answer the questions. Other justifications include the opportunity to plan before listening, use contextual cues and determine which metacognitive strategy to use to tackle test tasks (Littlewood, 1981; Thompson, 1995; Vandergrift, 1999).

Despite evidence of the positive impacts of MCQs on L2 learners' performance on listening tasks, some researchers still remain sceptical about this response format. Ur (1984) and Weir (1993) believe that previewing questions before listening changes the nature of the listening task, distracts test-takers, and puts a greater burden on them. Previewing answer choices might be considered less authentic and believed to inhibit processing strategies by distracting attention from the text (Yanagawa & Green, 2008). Moreover, test-takers are very likely to have difficulty holding four or more alternatives in their head while listening to the input text (Hughes, 2003).

A major problem with MCQs is that the reading involved in answering these questions could be seen as a source of construct-irrelevant variance. With the opportunity to

read the item stem and response options prior to listening, candidates can at least in part choose the answer by eliminating the options they consider incorrect. This process is different from the identification of the correct answer by listening and understanding the input text. In this situation, test-takers tend to undertake a reading task directed by the test questions (Farr et al., 1990) instead of constructing meaning in a natural way (Khalifa & Weir, 2009). It is not uncommon that test-takers are trained to approach MCQs in this fashion as part of their test-taking strategies (Gude & Duckworth, 2008). When the test-takers' performance is affected by their reading ability, construct-irrelevant variance will occur. Field (2009) pointed out that when the written MCQs are more difficult to interpret than the input text, test-takers' failure in choosing the correct answer can possibly be explained by their inadequate reading ability.

Another concern related to MCQs is the issue of guessing. In a four-option MCQ, for example, a candidate can have 25% chance of guessing correctly at random (Hughes, 2003). However, Buck (2001) argues that candidates do not often make guesses at random but based on partial comprehension. Construct-irrelevant variance appears in case of random guesses, which is quite common especially when the test-takers cannot make sense of the input. Apart from guessing, the presence of distractors can be tricky as well. Alderson et al. (1995) points out that because of distractors, candidates are likely to doubt their interpretation of text and end up giving the incorrect answer.

Although MCQ has been criticized for misinterpreting natural listening conditions, it still remains the most popular format for the large-scale testing of listening skill because of its practical benefits (Yanagawa & Green, 2008). Since candidates are not required to write anything when answering MCQs, one possible source of construct-irrelevant variance is removed (Elliott & Wilson, 2013). There is no need to train raters, which would help to reduce costs for the large number of test-takers involved (Yanagawa & Green, 2008).

Another practical benefit of MCQs is that they can be manipulated by item writers to make sure that the items are of appropriate difficulty and discriminate well by modifying the item stem and/or the response options (Elliott & Wilson, 2013). That said, it is extremely difficult to write valid and reliable MCQs for tests of listening ability. Therefore, considerable effort should be devoted to preparing and validating multiple-choice tests of listening.

#### **2.2.2.2. Lexical overlap between the input text and test items**

As explained by Buck (2001), lexical overlap is when words used in the passage are found in the question or in the response options. The degree of lexical overlap between the input text and the test items can considerably affect test-takers' performance on a listening test. Freedle and Kostin (1996, 1999) found that lexical overlap between the correct option and the text, especially the necessary information, tended to make items easier for test-takers. Likewise, lexical overlap between the input text and the incorrect response options made items more difficult. A possible explanation for this might be that test-takers tend to choose the options which contain the words they can recognize from the text. In a more recent study by Brunfaut and Révész (2015), no significant association was found between passage-response lexical overlap and test-takers' performance when the passage shared lexis with only the distractors.

Since lexical overlap has not been thoroughly investigated, little is known about the way(s) in which it guides or misleads test-takers. It is my belief that this factor accounts for multiple-choice item difficulty, especially in listening tests that give test-takers only one opportunity to listen. This study sheds further light on this under-researched area from a cognitive point of view. By means of think-aloud protocols, it seeks to explore how lexical overlap manipulates the listeners' thinking and decision-making process.

### **2.2.3. Construct-irrelevant variance**

The performance of test takers can be affected by both construct-relevant factors and construct-irrelevant variance (CIV). While the former is part of the target construct and pertains to variables that the test developers intend to assess, the latter contaminates the construct and introduces bias and ‘noise’ into test scores (Aryadoust, 2013). In case of a language test, CIV can be defined as performance characteristics that have little or nothing to do with the test-takers’ language ability (Brown & Hudson, 1998). Messick (1995) divides CIV into construct-irrelevant difficulty and construct-irrelevant easiness. An example of the former is a listening test with poorly written items which are difficult to understand. In this case, test-takers might still have problems answering the item even when they are able to understand the input. On the other hand, construct-irrelevant easiness occurs when extraneous clues in the item or the format allow some individuals to respond correctly in ways irrelevant to the construct being assessed (Messick, 1995). Haladyna and Downing (2004) emphasize that CIV arises from systematic errors which are group or person specific. They further explain that construct-irrelevant easiness refers to a factor that tends to systematically increase test scores for a specific examinee or a group of examinees and construct-irrelevant difficulty does the opposite.

It should be pointed out that CIV can come from different sources. Haladyna and Downing (2004) propose a comprehensive taxonomy for the study of systematic errors associated with construct-irrelevant variance which is presented in Table 2.4.



**Table 2.4.** A taxonomy of construct-irrelevant variance (Haladyna & Downing, 2004, p. 20)

Category	Instances
<b>Uniformity and types of test preparation</b>	<ol style="list-style-type: none"> <li>1. Whether or not students get test preparation</li> <li>2. The extensiveness of test preparation</li> <li>3. Unethical test preparation</li> </ol>
<b>Test development, administration and scoring</b>	<p><i><b>Test development</b></i></p> <ol style="list-style-type: none"> <li>1. Item quality</li> <li>2. Test item format</li> <li>3. Differential item functioning</li> </ol>
	<p><i><b>Test administration</b></i></p> <ol style="list-style-type: none"> <li>1. Location of test site</li> <li>2. Altering the administration</li> <li>3. Participation and exclusion</li> <li>4. Computer-based testing</li> <li>5. Calculators in testing</li> </ol>
	<p><i><b>Test scoring</b></i></p> <ol style="list-style-type: none"> <li>1. Scoring errors</li> <li>2. Sanitizing answer sheet</li> <li>3. Test form comparability</li> <li>4. Rater severity and prompt choice</li> <li>5. Accuracy of passing scores</li> </ol>
<b>Students</b>	<ol style="list-style-type: none"> <li>1. The influence of verbal abilities on test performance</li> <li>2. Test anxiety, motivation, and fatigue</li> <li>3. Accommodations for special student populations</li> </ol>
<b>Cheating</b>	<ol style="list-style-type: none"> <li>1. Institutional</li> <li>2. Individual</li> </ol>

It is apparent from Table 2.4 that there are a considerable number of factors that can introduce CIV into a test. These factors vary from test preparation, test development, administration, scoring to student-related characteristics and cheating incidents. The

interpretation or use of scores on any test is vulnerable to these validity threats. Therefore, apart from investigating whether a test really measures the abilities it claims to measure, validation research should also look at the extent to which test scores are influenced by the knowledge and skills that are not intended to be measured. The investigation of the LCLT's validity takes CIV into consideration and seeks to identify sources of CIV that should be avoided for the betterment of future tests.

### **2.3. The concept of validity and validation**

In the realm of language testing and assessment, a predominant goal is to elicit knowledge and skills relevant to a linguistic domain through a test. Understandably, the test must be valid in a way that it produces results which truly represent the test-takers' language ability. The concept of validity, given its importance and complexity, has been an academic interest with its own life in the field of educational assessment (Chapelle, 1999). The way validity is conceptualized determines the scope and the nature of validity investigations and the methods to gather evidence (Xi, 2008). The following section provides an overview of the development of validity theories and validation frameworks from the mid-20<sup>th</sup> century to recent years in order to situate the current thesis in the field.

#### **2.3.1. Three types of validity**

In the 1950s and 1960s, validity was defined as the extent to which a test measures what it is supposed to measure (Aryadoust, 2013; Cronbach, 1971; Henning, 1987; Lado, 1961). It was understood to have distinct classes. The American Psychological Association divided validity into four specific types: content, predictive, concurrent and construct validity (Cronbach & Meehl, 1955). Content validity is concerned with the extent to which test items are relevant to and representative of the content intended to be tested. Predictive validity is considered the effectiveness of a test in predicting the test-takers' future performance. Concurrent validity is concerned with the degree of correlation between the intended test and

another test given at approximately the same time. Lastly, construct validity is the evaluation of how well a test reflects and measures the test-takers' abilities that it is supposed to measure from the beginning. Later on, predictive and concurrent validity were merged into one type of validity named criterion-related, an attempt to avoid terminological confusion (Smith, 2001).

The validation process, during this time, was defined as conducting studies to collect evidence of one or more of these three types of validity (Aryadoust, 2013). This approach did not satisfy the research and tester communities since it hinged on a narrow view that treated validity as an inherent property of a test. Davies and Elder (2005) maintained that classical validity studies, at the time, looked in different and unrelated directions and therefore lacked harmony and coherence.

### **2.3.2. Validity as a unitary concept**

From the 1970s, the three types of validity were no longer in fashion as it was noticed that the nature of validity is not that clear-cut. Instead, theorists began to lean toward the unification of different types of validity. Cronbach (1980) emphasized the unitary nature of validity by mentioning that "all validation is one". However, it was not until 1989 that Messick laid the foundation for a radically different view of validity. In his ground-breaking article, Messick (1989) extended the concept of construct validity, defining it as "an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p. 288). Messick did not see validity as a property of the test itself but as a property of the interpretation and use of test scores. In his point of view, validity should be judged by the availability of theoretical rationales or empirical evidence. He also posited that validity arose from the interaction of two variables, namely the sources of testing justification and the functions or outcomes of testing. The source of testing justification includes evidential basis for the validity of the assigned score as well as the

consequential support which helps to make sure that the test's method of assigning scores does not have any negative social impacts. The functions or outcomes of testing are comprised of the uses and the interpretation of test scores. The two-by-two matrix representation of validity is presented in Table 2.5.

**Table 2.5.** The two-by-two representation of validity (Messick, 1989, p. 17)

	<b>Test interpretations</b>	<b>Test uses</b>
<b>Evidential basis</b>	Construct Validity	Construct validity Relevance/utility
<b>Consequential basis</b>	Construct Validity Value implications	Construct validity Relevance/Utility Value implications Social consequences

As seen in this matrix, validity is a unified concept which takes on four crucial dimensions, construct validity, relevance or utility, value implications, and social consequences. Altogether, these four aspects are meant to enable a systematic appraisal of a given test's validity. Messick's theory of validity still remains highly influential throughout the historical development of validation theory (McNamara, 2006). Validation research that applies Messick's model, continues to adopt an evidence-gathering approach, meaning that there are different types of validity evidence to be collected and they all complement each other in assessing the unifying concept of construct validity (Snellings et al., 2004).

Although theoretically elegant, Messick's theory has been challenged over the past decade for its lack of guidance on how validation work should be done (Xi, 2008). Kane (2012) points out that Messick's unitary theory does not provide a clear approach to implementation of the concept. Since validity is now a unitary concept, any information obtained in the development and use of a test can be relevant to its validity (Anastasi, 1986). Hence, there are questions of where the validation process should start and how much

evidence would be enough (Kane, 2012). This uncertainty encourages the inclusion of any readily available evidence as empirical support for a claim to validity (Aryadoust, 2013). Messick's model, though theoretically sound, appears to confuse practitioners in their actual validation research. The search for a more transparent and manageable validation framework, therefore, continues.

From the beginning of 21<sup>st</sup> century, a new approach to validation known as the argument-based approach started to gain popularity. The notion of validity argument is not new in the area of educational measurement since it was developed about two decades ago when Cronbach (1988) started to think of using a consistent framework for structuring different sources of validity evidence. However, it only became a coherent framework for validation when Kane took up on this, formalizing the development and evaluation of the validity argument (Xi, 2008). Kane (2006) views validation as a "process of evaluating the plausibility of proposed interpretations and uses" (p. 17). Put simply, the argument-based approach to validation consists of two key steps. First, construct an interpretation and use argument, and second, evaluate the plausibility of the interpretation and use argument (Kane, 2013).

A major strength of the argument-based approach is that it helps to address the limitation of Messick's unitary theory since it was intended to avoid the open-endedness and ambiguity of construct validity (Kane, 2013). Chapelle et al. (2010) point out that the argument-based approach allows validation research to be carried out through a systematic process of examining the inferences in the interpretive argument rather than referring to different types of validity evidence. It is worth noting that although the argument-based approach is claimed to be more transparent, accessible and systematic, that does not mean validation has become easier. Quite to the contrary, validation will be even more demanding because it is no longer the case that any information gained through the development and use

of a test can be relevant to validity (Kane, 2013). Even though the process is made clear with two key steps as mentioned above, more work has to go into selecting and synthesizing evidence to support each inference in the interpretive argument.

Overall, the past 70 years has witnessed the evolution of the concept of validity, from three distinct types of validity to the unification of these three validities into an overarching one, construct validity. Attention has also shifted from the test itself to the interpretation and use of test scores. The expanded views on validity result in more sophisticated validation frameworks. The “one question, three validities” approach with its narrow view on validity and oversimplified validation framework has become an artefact from the past (Chapelle & Voss, 2013). The evidence-gathering approach under Messick’s unitary theory is still popular. This presentation of validity as an evidence-based judgement is still highly influential among validation researchers even in the absence of a clear guideline for what kind of evidence to collect or how much evidence is needed (Davies & Elder, 2005). These limitations are addressed by the most recently developed approach, the argument-based approach to validation. Although validity evidence still plays a crucial role in this approach, it can no longer be simply gathered. Instead, such evidence needs to be carefully selected, prioritized and synthesized so as to serve a higher purpose which is to support the evaluation of the interpretation and use argument. This approach has gained popularity in the past 10 years and received support from international testing experts. It is also the approach that constitutes the main theoretical structure of this research and thus, will be discussed further in the next section.

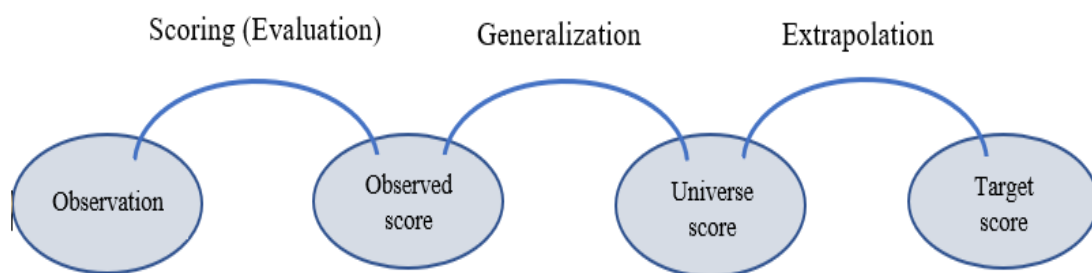
#### **2.4. The argument-based approach to validation**

The argument-based approach in language testing and assessment views validity as an argument supported by an analysis of theoretical and empirical evidence instead of a collection of evidence from separate quantitative and qualitative analyses (Kane, 1992; 2001;

2002; 2013). This approach makes use of two kinds of argument: an interpretive argument and a validity argument (Kane, 2004). To construct an interpretive argument, the desired uses and interpretations of test scores are stated. In the next stage, a validity argument is developed in order to investigate the accuracy, plausibility and feasibility of the claims made in the interpretive argument (Kane, 2006). The construction of a validity argument is based on Toulmin's (2003) argument structure. In the two sections below, the interpretive and validity arguments are explained in more detail.

#### 2.4.1. Interpretive argument

Kane (2013) posits that test scores can have multiple interpretations or uses, and it is the proposed interpretation/use that is validated, not the test itself or the test scores. With that being the case, an interpretive argument is a network of inferences and supporting assumptions leading from scores to conclusions and decisions (Kane, 1992; Kane et al., 1999). It provides an explicit statement of a proposed interpretation and a framework for developing a validity argument (Kane, 2002). Kane et al. (1999) illustrates an interpretive argument with three types of inferential bridges (Figure 2.3).



**Figure 2.3.** The bridge analogy (Kane et al., 1999, p. 9)

Kane's approach to the interpretive argument is that each of its inferences rests on assumptions that require support (Chapelle et al., 2008). The first inference which is usually referred to as the scoring or evaluation inference takes us from observed performances to an observed score (Figure 2.3). It typically makes assumptions about the appropriateness of the

scoring criteria and the methods used to generate scores from the raw data (Kane, 1992; 2013).

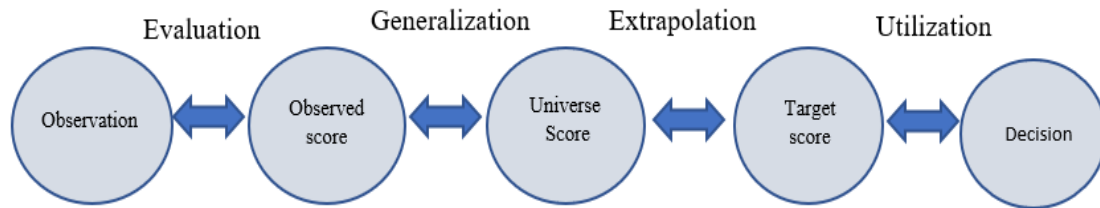
However, we usually want to go beyond the claims about test-takers' scores on a particular test at a particular point in time. Typically, we want to make claim about test-takers' future performance. This involves the generalization inference which posits that observed test scores would be consistent through multiple test administrations and parallel test versions (Aryadoust, 2013). It relies on the assumption of the representativeness of the sample and the adequacy of sample size (Kane, 2006).

The third link between the universe score and the target score involves the extrapolation inference. It has a bearing on whether test-takers' performance provides adequate evidence about their language abilities that underlie their language performance in the target domain (Xi, 2008). In language assessment, one assumption underlying extrapolation is that test task characteristics are relevant to tasks in the target language domain (Chapelle et al., 2008).

It should be noted that the claims being made vary from one case to another and the evidence needed to support these claims will vary, too (Kane, 2013). Kane (1992) also stresses that it is not possible to verify an interpretive argument in an absolute sense and the best that can be done is to show that the argument is highly plausible, given all available evidence.

As it grows in popularity, the flaws of the three-bridge interpretive argument have also been noticed and criticized. Bachman (2005) argued that the argument focuses mainly on the interpretation of test scores and should be broadened to include test use. Kane (2001, 2002, 2004) addressed this issue by extending the linkages in the interpretive argument from scores all the way to score-based decisions. This chain of inferences is illustrated in Figure 2.4.

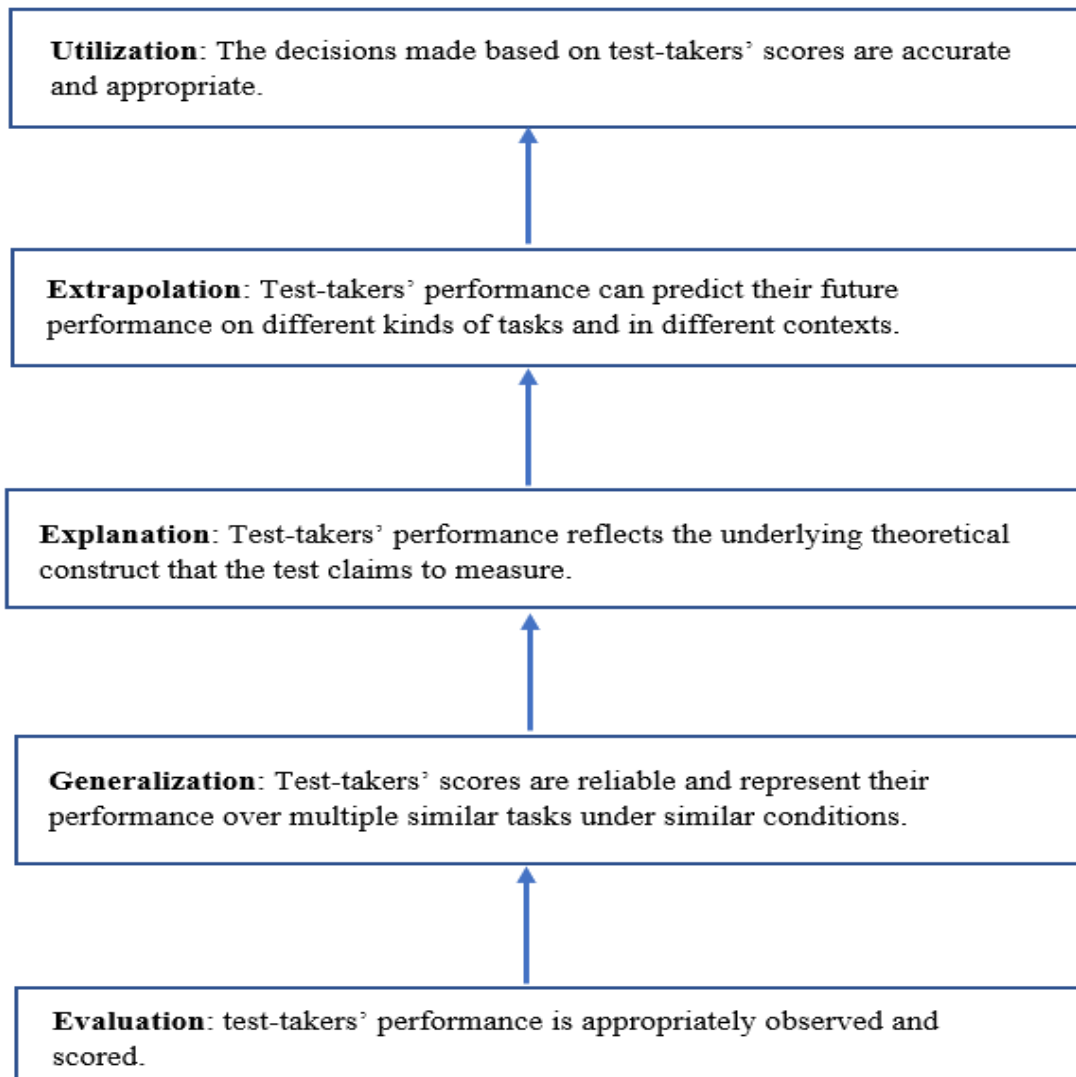




**Figure 2.4:** Links in the extended interpretive argument (Kane, 2001, 2002, 2004)

Kane (2013) reflected on the criticism from Bachman by admitting that although the term “interpretive arguments” were used to indicate statements of inferences inherent in both the interpretation and use of test scores, it might have given too much weight to interpretation and not enough to uses. Therefore, he addressed this imbalance by renaming the argument as “interpretive/use argument” (IUA).

In an attempt to operationalise Kane’s approach and develop an interpretive argument for the TOEFL test, Chapelle et al. (2008) found that the approach failed to link test scores to a theoretical construct. Specifically, Kane’s three-bridge interpretive argument did not accommodate a claim that test takers’ performance as evidenced by their scores truly reflected the targeted construct. Therefore, the explanation inference was added to bridge the gap between the generalization and extrapolation inference. It posits that the test actually measures its intended theoretical construct. Another inference was also added to link the target score to the decisions about test-takers. Chapelle et al. (2008) adopted the term “utilization” by Bachman (2005) for the TOEFL interpretive argument. As a result of the adaptations by Bachman (2005) and Chapelle et al. (2008), the interpretive argument has been extended to include up to five inferences which are evaluation, generalization, explanation, extrapolation and utilization. The full list of five inferences is provided in Figure 2.5.



**Figure 2.5.** Five inferences in an interpretive argument

Thus far, it has become apparent that the Kane's framework provides the core elements of an interpretive argument. The extension beyond the three bridges helps to cover other important aspects of score interpretation and use. However, it is important to note that the inferences discussed in this section are not a checklist. Kane (2013) emphasizes the need for flexibility in the development of an interpretive argument by asserting that the argument does not have to follow any particular pattern. Indeed, the interpretive argument should reflect the proposed interpretations and uses of a particular test. It should not be constrained to fit some pre-specified structure (Kane, 2013).

### 2.4.2. Validity argument

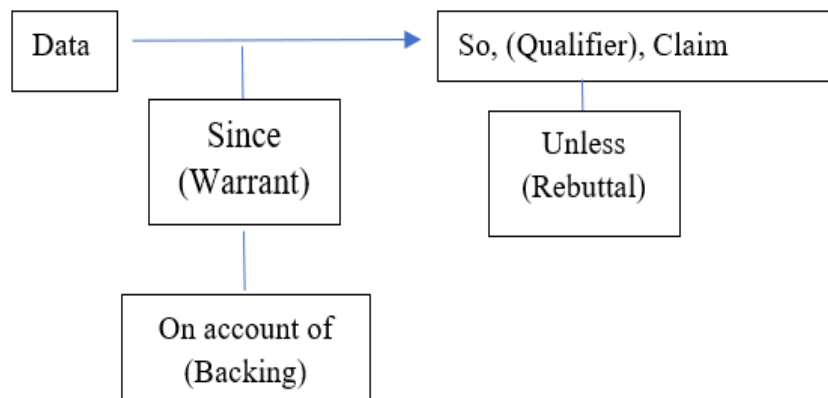
The validity argument provides an overall evaluation of the inferences in the interpretive argument (Kane, 1992, 2002). Central to the validity argument is the presentation of all the evidence relevant to the interpretations and uses of test scores as stated in the interpretive argument. Kane (1992) explained that the argument-based approach does not identify any kind of validity evidence as being generally preferable to any other kind of validity evidence. In fact, different inferences will require different kinds of evidence and analysis for their evaluation. Since the validity argument aims for a cogent presentation of evidence relevant to the proposed interpretations and uses of test scores, much attention must be paid to the way it is structured. For the construction of a validity argument, Toulmin's (2003) argument structure is adopted because it offers guidance on how to justify a proposed inference.

Developed by philosopher and lawyer Stephen E. Toulmin, the Toulmin method is the style of argumentation that takes into account six main components: data, claim, backing, warrant, rebuttal, and qualifier (Toulmin, 2003). Each of these components is defined as follows:

- A claim is “a conclusion whose merits we are seeking to establish” (Toulmin, 2003, p.90). In other words, a claim is the interpretation that we want to make on the basis of the data about what a test taker knows or can do.
- Data includes “information on which the claim is based” (Toulmin, 2003, p.90).
- A warrant is a general rule for inferring claims of certain kind from data of certain kind (Kane, 2013).
- Backings are general statements that legitimate warrants and their application (Aryadoust, 2013).

- A rebuttal indicates “circumstances in which the general authority of the warrant would have to be set aside” (Toulmin, 2003, p. 94).
- A qualifier indicates the strength of relationship expressed by the warrant (Kane, 2013). For example, adverbs such as “sometimes”, “often” or qualifiers like “most likely”, “presumably” can be used to indicate the likelihood of the claim.

The Toulmin model has been one of the most widely studied models of argumentation. This style of reasoning has gained popularity since the 1970s under various headings such as “practical reasoning”, “informal logic”, and “rhetoric” (Kane, 1992). Kane (2013) emphasized that this model of argumentation provides explicit, albeit contingent, guidance for validation. With the claim clearly stated from the beginning, the evidence required for validation is also the evidence needed to evaluate the claim. A visual representation of the Toulmin’s argumentation model can be found in Figure 2.6.

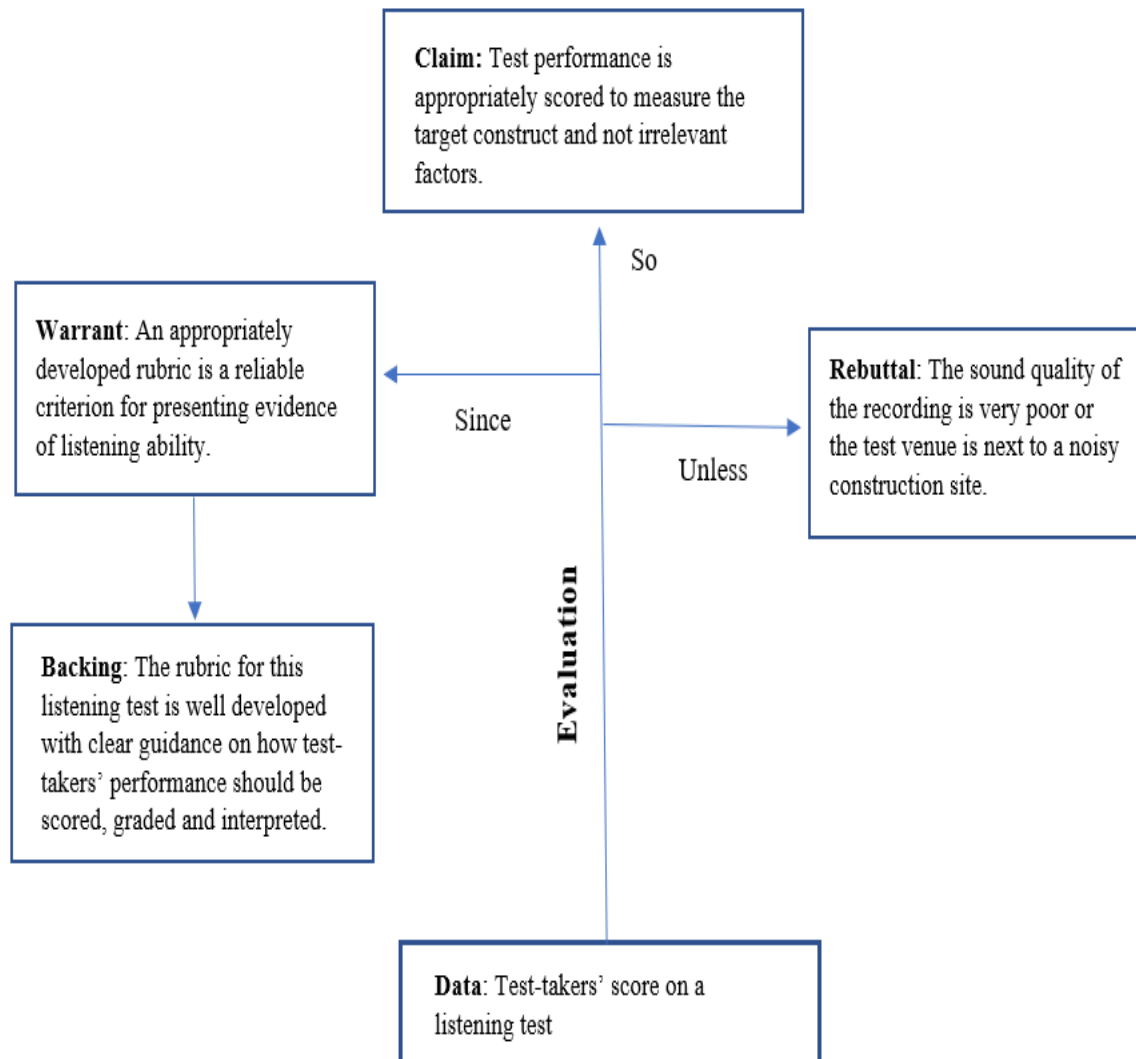


**Figure 2.6.** Toulmin's model of inference (Toulmin, 2003, p. 97)

In this model, each inference starts from data and makes a claim. For example, to use an example outside testing in the Vietnamese context, when you know that a man retires (data), you can claim that he is most likely (qualifier) 60 years old. This claim is supported by the fact that in Vietnam, the retirement age, as determined by the labour code, is 60 for men

who work in the public sector or state-owned enterprises (warrant). If you know that the man used to be the headmaster of a public secondary school (backing), then the retirement age applies. That said, he might be younger than 60 in case he decided to retire early due to health reasons (rebuttal).

The Toulmin argument framework can be used to justify an inference made about a certain test. The following example illustrates how this framework can be applied for the evaluation inference in an interpretive argument. Hypothetically, an evaluation inference is made about test-takers' scores on a listening test. That is, their test performance is appropriately scored to measure the targeted construct and not irrelevant factors. There can be several warrants to support the claim. One of them is that an appropriately developed rubric can be a reliable criterion for presenting evidence of listening ability. Backing for this warrant could include that the actual rubric is well developed with clear guidance on how test-takers' performance should be scored, graded and interpreted. Rebuttal can be the fact that the sound quality of the recordings is very poor or the test venue is next to a noisy construction site. This means that even if the rubric is very well developed, the plausibility of the claim or the evaluation inference is still highly questionable. The poor sound quality and the noisy environment are the irrelevant factors that might have undesirable impacts on the test-takers' performance. Therefore, their scores on the test might not reflect their true listening ability. A visualization of this evaluation inference is provided in Figure 2.7.



**Figure 2.7.** Example of an evaluation inference for a listening test

Toulmin (2003) emphasizes that those who make a claim have the responsibility to make it good and show that it is justifiable. The application of the Toulmin argument model helps to avoid informal, loose and contingent inferences. By using this framework, the claim-maker can specify the types of evidence needed for the evaluation of each inference in the interpretive argument.

Over the years, researchers in language testing and assessment have made great contributions in terms of the methods for collecting evidence to back a warrant in a validity argument for a certain test (Xi, 2008). Among quantitative methods, the ones which have

been frequently used are item analysis, score reliability analysis, factor analysis and correlational analysis (Bachman, 2004; Chapelle, 1999). At the same time, popular qualitative methods include the analysis of test content, the analysis of test-takers' self-report data as well as the use of questionnaires and interviews in the investigation of consequences and washback effects (Field, 2005; Green, 1998; Xi, 2008). This research uses a mix of both quantitative and qualitative methods which include item analysis, correlational analysis, content analysis, and test-takers' verbal report. The application of these methods in this research will be explained in detail in chapter 3, 4, and 5.

In a validity argument, some inferences might require more evidence than others. Kane (2013) suggests that strong claims (e.g. predictions of future performance in different contexts) would require extensive empirical support and several parallel lines of evidence. Although all of the inferences deserve some attention, it is useful to conduct studies to elicit evidence for the most questionable claims (Cronbach, 1982, 1988). If these claims are backed by sufficient evidence, their plausibility increases. The validity argument will therefore be supported. That being said, the entire validity argument will be undermined if the inferences which are in serious doubts are not found to hold up well in practice.

#### **2.4.3. The argument-based approach in practice**

The argument-based approach to validation has been widely used as the basis for validation research on major international language tests such as the Test of English as a Foreign Language (TOEFL) (Chapelle et al., 2008), the International Language Testing System (IELTS) (Aryadoust, 2013; Ebadi et al., 2014) and the Pearson Test of English (Wang et al., 2012). It has also become the foundation for Bachman's (2005) and Bachman and Palmer's (2010) assessment use argument (AUA). The past decade has witnessed an increase in the application of the argument-based approach in a number of other language testing projects (Aryadoust, 2009; Jia, 2012; Jun, 2014; Le, 2011; Li, 2015; Youn, 2015).

The growing popularity of the argument-based approach can be attributed to the advantages it offers researchers in language testing and assessment. First of all, laying out the intended interpretations of test scores provides a meaningful starting point for the validation process. Specifically, the inferences explicitly stated in the interpretive argument lay the ground for outlining essential research and interpreting validity evidence (Chapelle, Enright & Jamieson, 2010). Second, the argument-based approach provides guidance in deciding what kinds of validity evidence are needed in allocating research efforts (Bachman, 2004; Cronbach, 1988; Kane, 1992). The structure of the validity argument determines the types of evidence that need to be collected to support each inference stated in the interpretive argument. However, Kane (2013) suggests that there is no need to belabour the obvious and the validation effort should focus on the most questionable inferences. Third, the argument-based approach offers researchers a framework for synthesizing research findings so as to arrive at an overall evaluation of a validity argument (Chapelle, 2012).



## Summary

This chapter presents the theoretical framework for my entire research project which aims to develop a validity argument for a locally created listening test in Vietnam. First, I provided an overview of second language listening which laid the ground for the subsequent review of major issues in assessing this skill. These issues included the definition of the listening construct, factors affecting listening task difficulty, and construct-irrelevant variance. I also provided an account of how the concept of validity evolved and how validation frameworks changed over time. After reviewing the argument-based approach to validation, I argued that this approach had an advantage over previous validation models for its well-structured framework and clear guidance in allocating research efforts. This approach was therefore chosen to guide my research on the LCLT.

### **Chapter 3. Overall methodology and research design**

The primary objectives of this chapter are (a) to build an argument-based validation framework for the LCLT and (b) to identify the research methods needed to elicit validity evidence in accordance with that framework. In this chapter, a general description of the LCLT will be provided, followed by the justification for the inferences that this research project focuses on. The remainder of the chapter is dedicated to the overall research design that shapes the way in which evidence is collected to support each inference.

#### **3.1. General description of the LCLT**

The LCLT is a component of the English proficiency test battery that the host university developed and administered to 124 test-takers in February 2017. This test targets Levels 3 to 5 of the CEFR-V. It was designed following the MOET-approved format for a listening test. As stated in Decision No. 729/QĐ-BGDĐT, dated March 11, 2015 on the format of the three-level English proficiency tests for Vietnamese, the listening subtest consists of three parts with 35 multiple-choice questions and is up to 40 minutes long. This test targets listening subskills described for level 3 to level 5 of CEFR-V. For example, listening for details, listening for main ideas, understanding the implications of the speakers. In the first part of the test, candidates listen to eight short conversations and answer one multiple-choice question for each conversation. In the second part, there are three long conversations between native speakers of English. Candidates are asked to listen and answer four multiple-choice questions for each conversation. The final part of the test includes three monologues in which two were talks by native speakers and one was part of a lecture. Five multiple-choice questions were written for each of these monologues. Table 3.1 summarizes the actual content of the LCLT.

**Table 3.1.** Content of the LCLT

<b>Part</b>	<b>Number of questions</b>	<b>Content</b>
Part 1: 8 short conversations (Level 3)	8 multiple-choice questions (Item 1-8)	Transport, bills, leisure activity, birthday party, location, prices, holiday
Part 2: 3 long conversations (Level 3, 4, 5)	4 multiple-choice questions (Item 9-12)	Conversation 1: Movie
	4 multiple-choice questions (Item 13-16)	Conversation 2: Interviewing a musician
	4 multiple-choice questions (Item 17-20)	Conversation 3: Interviewing movie maker
Part 3: 2 talks, 1 lecture (Level 4, 5)	5 multiple-choice questions (Item 21-25)	Talk 1: Smithsonian museum
	5 multiple-choice questions (Item 26-30)	Talk 2: Hawaiian volcanoes
	5 multiple-choice questions (Item 31-35)	Talk 3: Concert of music college's students

Candidates can listen to the recording only once. They are advised to answer the questions as they listen and given time to transfer their answers to the answer sheet once the recording stops. In terms of scoring, one point is given to one correct answer and no point is subtracted for incorrect answers.

### **3.2. Argument-based validity for the LCLT**

In this present research which aims to develop a validity argument for the LCLT, the evaluation, generalization and explanation inferences are the chosen foci for two main reasons. First, since this study is the first attempt in validating this test, it makes practical sense to start with the most basic inferences. As Aryadoust (2013) emphasizes, these low-level inferences lay the groundwork for higher-level inferences (extrapolation and utilization inference) to function properly. For example, if it is found that test-takers' performance on

the LCLT does not reflect the targeted listening construct, it will not be possible to predict the test-takers' future performance in some real-life contexts which involve listening skill in English. In other words, if the explanation inference is refuted, the extrapolation inference will not hold, either. The second reason is that by focusing on the evaluation, generalization, and explanation inferences, this research sheds light on various technical issues of the LCLT. These insights will be most useful for the host university and their test developers. In light of the Toulmin's (2003) argument framework, these three inferences for the LCLT are specified as follows.

The evaluation inference rests on the claim that the test-takers' performance on the LCLT was appropriately observed and scored. This inference usually makes assumptions about the appropriateness of the scoring rubrics, task administration conditions and statistical characteristics of test items (Chapelle et al., 2008). In this particular case of the LCLT, the scoring rubrics and the test administration report were not provided by the host university, meaning only test-takers' results could be the subject of an in-depth quantitative item analysis. Such an analysis helps to examine whether the statistical characteristics of the LCLT items are appropriate for wide administration and for fair decisions to be made about test-takers' listening ability. With a focus on the statistical characteristics of the LCLT items, the evaluation inference involves the three following warrants:

- Item difficulty was appropriate for test-takers.
- Test items had reasonable discriminating power.
- Test items were relevant to the targeted listening construct.

If the item analysis reveals that the LCLT items were either too easy or too difficult for the test-takers, the inference will not be supported. Similarly, if the LCLT items have very low discriminating power, this inference will not hold, either. Evidence of construct-irrelevant

variance also contributes to weakening the claim that test-takers' performance on the LCLT was appropriately observed and scored.

The generalization inference for the LCLT involves only one warrant that the LCLT had high reliability indices. This inference will be refuted if the reliability indices are not high enough for a high-stakes test.

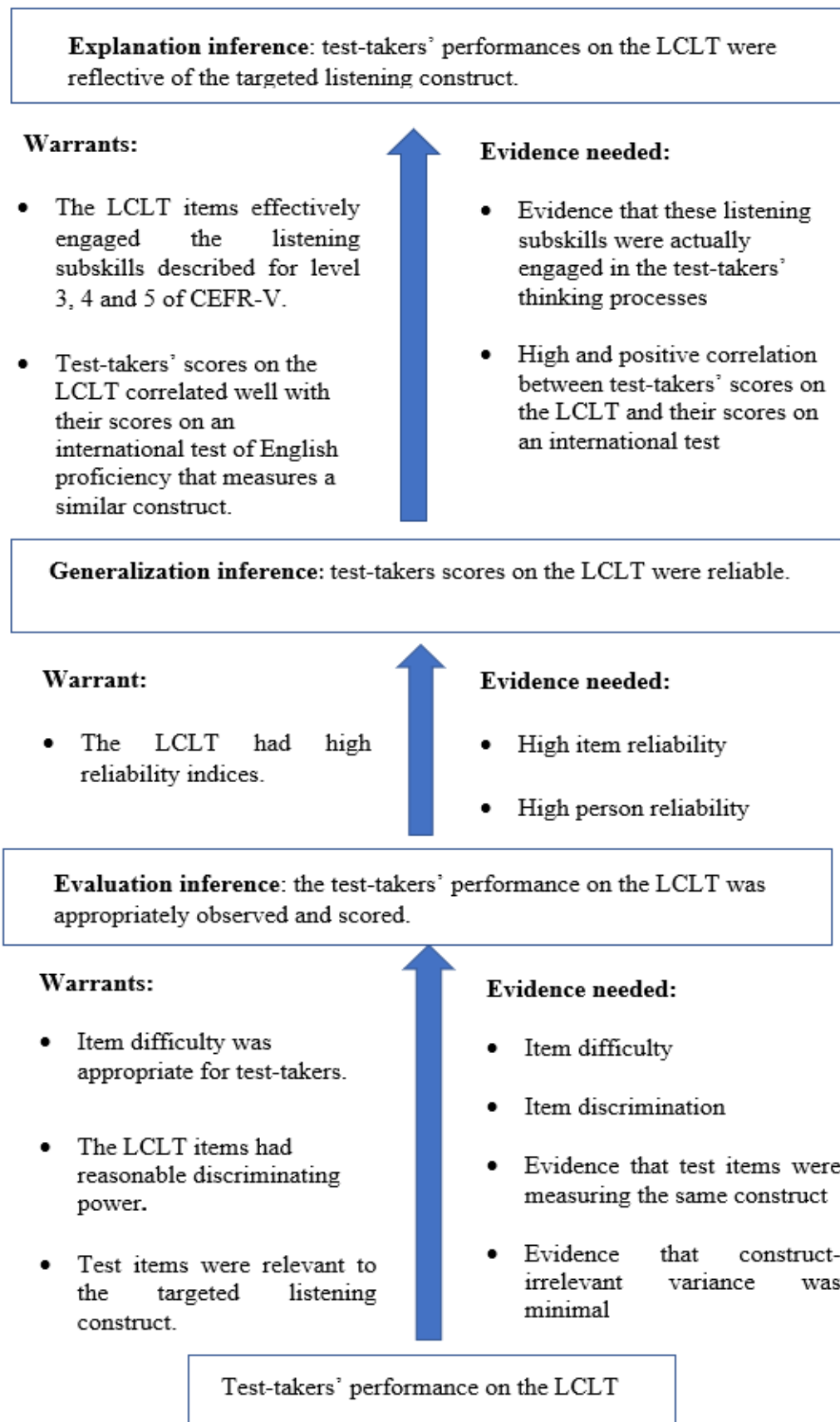
The explanation inference holds that the test-takers' performances on the LCLT were reflective of the targeted construct. Since the LCLT targets level 3 to level 5 of the CEFR-V, it is important to make sure that the test items actually engaged the listening subskills described for these three levels. Given that this test aims to achieve a certain level of comparability with an international test of English proficiency, it is also necessary to examine the extent to which scores on the LCLT correlate with scores on an international English proficiency test that measures a similar construct. The explanation inference therefore involves two crucial warrants:

- The LCLT items actually engaged the listening subskills described for level 3, 4 and 5 of CEFR-V.
- Test-takers' scores on the LCLT correlate well with scores on an international test of English proficiency that measures a similar construct.

The first warrant will be refuted if the LCLT items are found to poorly cover the listening subskills from Level 3 to Level 5 of the CEFR-V and contain construct-irrelevant variance.

The second warrant will not be supported if test-takers' scores on the LCLT do not correlate well with their scores on an international test of English proficiency that measures a similar construct.

With the evaluation, generalization, explanation inferences stated above, the validation framework for the LCLT can be summarized in Figure 3.1.



**Figure 3.1.** Validation framework for the LCLT

### 3.3. Overall research design

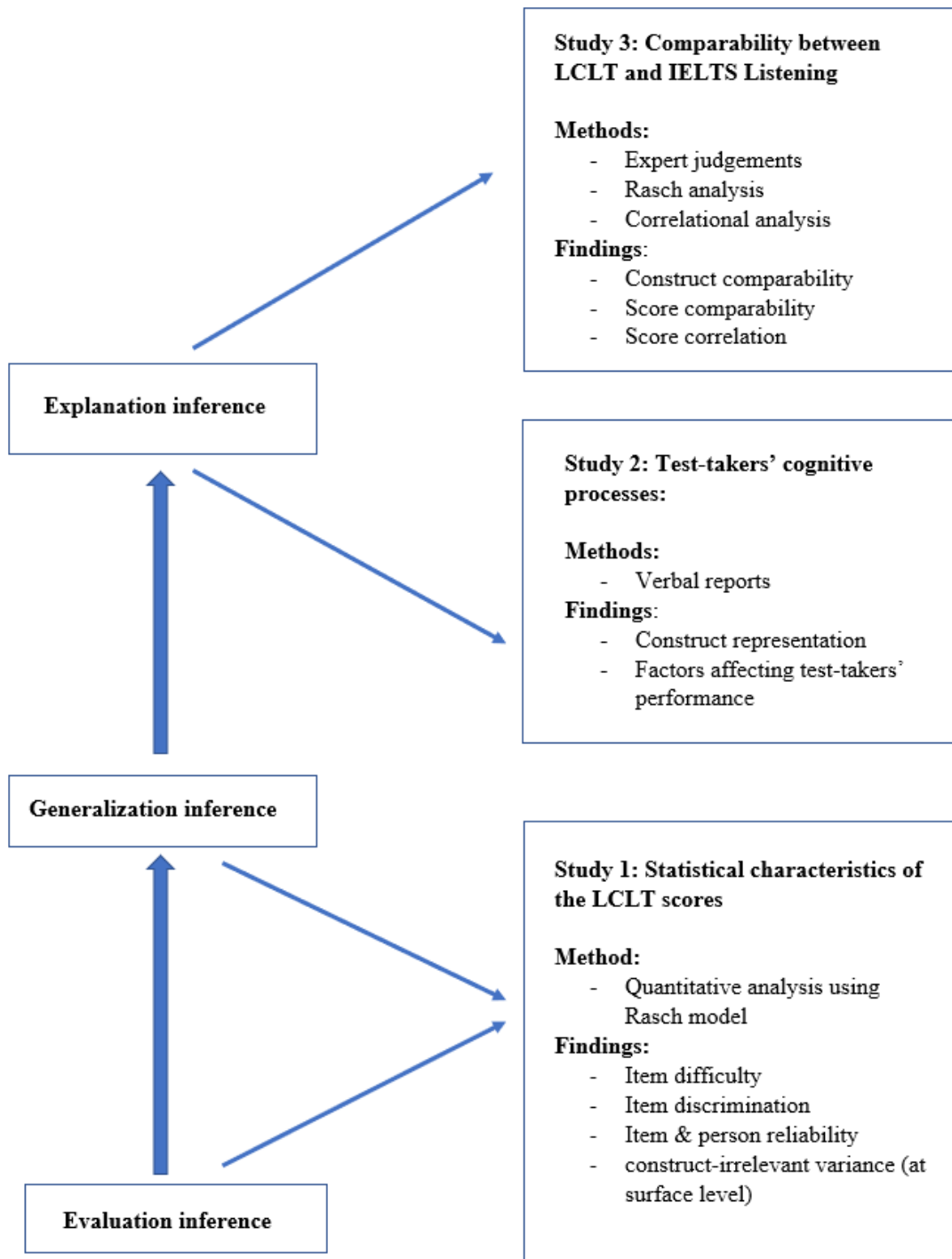
In order to elicit the evidence needed to support the evaluation, generalization and explanation inferences (Figure 3.1), three independent studies will be carried out. Each of these studies sheds light on a particular aspect of the LCLT's validity and together, they create a big picture that will make it possible to judge how well the validity argument for the test holds up in practice.

The first study which centres around the statistical characteristics of the LCLT scores elicits evidence for the evaluation and generalization inferences. In this study, the Rasch model will be used as the main analytical tool which provides insights into item difficulty, item discrimination as well as item and person reliability. Rasch analysis can also signal the presence of construct-irrelevant variance but only at the surface level.

The second study delves into the cognitive processes that test-takers underwent as they answered the LCLT items. This study provides evidence for the explanation inference by bringing to light the extent to which the test items engaged its targeted listening subskills. Verbal reports or think-aloud protocols are used to investigate the thinking behind test-takers' answers. The qualitative analysis of verbal data sheds light on not only construct representation but also the factors affecting test-takers' performance.

Apart from the investigation of test-takers' cognitive processes, the explanation inference requires further support from another study which focuses on the comparability between the LCLT and an international test that measures a similar listening construct. This third study compares the LCLT and an IELTS Listening test in two substantial aspects: the construct and the test-takers' scores. While experts are invited to give judgements on the comparability between the LCLT and IELTS constructs, Rasch analysis and correlational analysis are needed for the comparison of test-takers' scores.

The overall research design, as explained previously, is schematized in Figure 3.2.



**Figure 3.2.** Overall research design

Study 1, 2 and 3 will be reported in terms of both methodology and findings in Chapter 4, 5 and 6 respectively.



## Chapter 4. Statistical characteristics of the LCLT scores

This chapter reports on the quantitative study that investigated the statistical characteristics of the LCLT scores. Its findings provide evidence necessary for the evaluation and generalization inferences. The quantitative analyses in this study were carried out using SPSS and the Rasch model, which is a useful, flexible statistical tool that has been widely used in the realm of test validation. This chapter begins with the aims and methodology of this study. Next, the findings are presented. In light of these findings, an overall assessment of the evaluation and generalization inference is provided at the end of the chapter.

### 4.1. Aims and methodology of the study

This study seeks evidence for the evaluation and generalization inferences. Table 4.1 summarizes these two inferences and their relevant warrants.

**Table 4.1.** The evaluation and generalization inferences for the LCLT

Inference	Warrant
<b>Evaluation:</b> The test-takers' performance on the LCLT was appropriately observed and scored.	Item difficulty was appropriate for test-takers.
	Test items had reasonable discriminating power.
	Test items were relevant to the targeted listening construct.
<b>Generalization:</b> Test-takers' scores on the LCLT were reliable	The LCLT had high item and person reliability indices.

In order to elicit evidence to support the five warrants in Table 4.1, this study involved two groups of participants and data collection was carried out in two steps. Details of the participants and data collection procedures are provided in the following sections.

#### **4.1.1. Participants**

Two separate groups of participants were involved in this study. The first group consisted of 124 anonymous test-takers who took the LCLT at the beginning of 2017. In terms of participation, this group of test-takers were not directly involved in my study. Since their personal information was not made available to me, it was not possible to provide a detailed description of these participants.

The second group of participants included 76 English majors from the host university who volunteered to take the LCLT in May 2017. This was a mock test event organized for the sake of this study only and the results of the test did not have any effect on the students' grades at the university. All the test-takers are female and Vietnamese is their first language. At the time of this study, these students, aged 18-20, were studying at the English Department of the host university. Before they decided to participate in the study, all of them were informed of (a) the purpose and design of this study, (b) their rights and responsibilities as participants, and (c) relevant ethical issues (see Appendix C and D for the information sheet and the consent form).

#### **4.1.2. Data collection procedure**

In the first stage this study, the host university granted me the access to the following data:

- the original test booklet;
- 124 answer sheets (with test-takers' candidate numbers, their responses to each individual test item and the total scores);
- the answer keys;
- the audio recording of the listening text.

In the second stage, the original test booklet and the recording of the listening text were reused in the mock test administered to the group of 76 English majors. Each student was given a candidate number and advised to write only this number on their answer sheets. The time allowed for this test was 30 minutes. All the test booklets and answer sheets were collected before the students left the test venue. These students' performance on the LCLT was then marked according to the answer keys which were previously provided by the host university. Their responses to each individual test item as well as their total scores were recorded for analysis. The test-takers' results were kept confidential and could only be revealed to those who would like their results back by emailing me a request with their candidate numbers.

Collecting test results from this extra group of 76 students helped me achieve a larger sample size of 200 cases. This is the minimum sample size that a high-stakes, multiple-choice test should have to achieve stable and accurate item parameters (Hambleton & Cook, 1983). Moreover, the merge between 124 random test-takers and 76 English majors resulted in a more representative group of participants with supposedly greater variability in English proficiency levels. Since the LCLT is a multiple-level test, such variability is crucial in examining whether the test was able to separate different levels of proficiency among test-takers.

#### **4.1.3. Data analysis procedure**

In this study, statistical analysis was carried out on a data set that included the test results of 200 candidates. This data set consisted of the test-takers' overall scores and their responses to 35 items in the LCLT. The test-takers' responses to each item were coded dichotomously into "1" for a correct answer and "0" for an incorrect answer.

Once the data set was created, descriptive statistics were established with SPSS (Version 25.0) so as to provide information of test-takers' general performance and the normality of score distribution. Then, in-depth analyses of the LCLT scores were carried out. At this stage, the dichotomously transformed data were analysed using the Rasch model and

the computer software called Winsteps (Version 4.0.1). The features of Rasch analysis that were most useful for this study were the item-person estimates, discrimination index, fit statistics, and reliability indices. The analysis results are reported in the following section.

#### 4.2. Statistical characteristics of the LCLT scores

In this section, descriptive statistics are presented first so as to provide general information of the test-takers' performance and score distribution. The main focus is the results of Rasch analysis which cover item difficulty, item discrimination, construct-irrelevant variance and reliability.

##### 4.2.1. Descriptive statistics

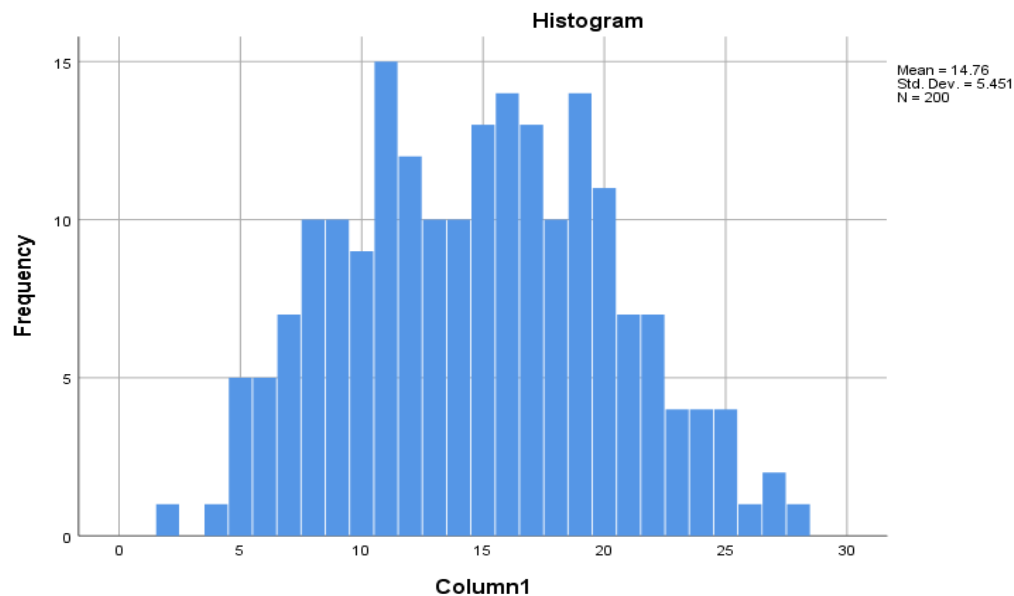
Table 4.2 provides a summary of descriptive statistics for the LCLT scores. As can be seen, the range of scores was relatively wide, from 2 to 28.

**Table 4.2.** Descriptive statistics for the LCLT scores

LCLT (N = 200, k = 35)	
Min	2
Max	28
Mean	14.76
Mode	11
Median	15
Standard deviation	5.451
Skewness	.092
Kurtosis	-.654

It was clear from Table 4.2 that the values of mean, mode and median were all smaller than half of the perfect score. The most popular score was 11 which was relatively low. The

standard deviation was fairly large (5.451), indicating that scores were spreading out rather than clustering around the mean and that test-takers' performance did vary from each other. To provide further details about the characteristics of test scores, a visualization of score distribution is presented in Figure 4.1.



**Figure 4.1.** LCLT score distribution

As seen in Figure 4.1, the LCLT scores had fairly acceptable normality in distribution. The histogram was positively skewed as the distribution was seen to move slightly to the left of the central line of the curve with skewness value of .092. Also, the score distribution was shown to be rather flat which reflected its numerical kurtosis value of -.654. The distribution did not have extreme scores (neither zero or perfect score). In general, the values of skewness and kurtosis both fell into the acceptable range from -2 to 2, which helped to confirm that the score set of the LCLT had a reasonably normal distribution.

#### 4.2.2. The difficulty levels of LCLT items

One of the most important results of Rasch analysis is the insights into the difficulty levels of LCLT items. Table 4.3 provides the difficulty measures as well as the items'

positions along the difficulty hierarchy. The most difficult items are at the top of the table and the easiest are at the bottom.

Test items, as shown in Table 4.3 (next page) covered a fairly wide range of difficulty levels, from -2.06 logit to 1.68 logit. The most difficult item was item 10 and the easiest one was item 2. A mismatch between the intended and the observed item difficulty was noted. In the description of the LCLT (see Table 3.1), the targeted proficiency levels for each part of the test was specified as follows:

- Part 1(item 1-8): level 3
- Part 2 (item 9-20): level 3, 4, 5
- Part 3 (item 20-35): level 4, 5

From the test developer's perspective, test items were supposed to spread along a difficulty continuum with Part 1 at the easy end and Part 3 at the difficult end. However, that perspective on item difficulty did not appear to hold up well in practice. The locations of many items on the item difficulty hierarchy were very different from their intended levels of difficulty, especially items in Part 1 and Part 3 of the test. For example, item 3 and 5 which were supposed to be easy for test-takers actually landed in the middle of the difficulty order.

**Table 4.3.** Difficulty measure of LCLT items

Entry number (Item)	Total score	Total count	Measure (logit)	Part of the test
10	27	200	1.68	Part 2
29	28	200	1.63	Part 3
22	31	200	1.50	Part 3
34	39	200	1.20	Part 3
11	40	200	1.17	Part 2
20	41	200	1.13	Part 2
17	45	200	1.00	Part 2
28	52	200	.79	Part 3
18	53	200	.76	Part 2
24	53	200	.76	Part 3
9	55	200	.70	Part 2
13	60	200	.56	Part 2
19	62	200	.51	Part 2
16	68	200	.35	Part 2
3	78	200	.11	Part 1
35	78	200	.11	Part 3
33	80	200	.06	Part 3
26	84	200	-.04	Part 3
25	87	200	-.11	Part 3
5	91	200	-.20	Part 1
31	91	200	-.20	Part 3
23	104	200	-.50	Part 3
27	104	200	-.50	Part 3
21	105	200	-.52	Part 3
14	106	200	-.55	Part 2
32	106	200	-.55	Part 3
6	110	200	-.64	Part 1
8	111	200	-.66	Part 1
12	115	200	-.76	Part 2
15	116	200	-.78	Part 2
30	121	200	-.90	Part 3
7	134	200	-1.23	Part 1
1	157	200	-1.89	Part 1
4	158	200	-1.93	Part 1
2	162	200	-2.06	Part 1

A closer inspection of these items' locations on the item-person map (**Error! Reference source not found.**) shows that they were placed beyond the ability levels of many test-takers. At the same time, some items in Part 3 (e.g. item 21, 23, 27, 30, 32) which were expected to be challenging turned out to rank quite low in the difficulty hierarchy. Items in Part 2 of the test were scattered along the item hierarchy, which was reasonable since this part of the test was supposed to cover a wide range of ability from level 3 to level 5. However, the erratic observed item difficulty in Part 1 and Part 3 signalled that the intended construct was not well represented by the test items.



TABLE 1.0 VSTEP \_ HUE \_ HANU\_ LISTENING RESULTS. ZOU413WS.TXT Jan 6 2019 4: 5  
 INPUT: 200 PERSON 35 ITEM REPORTED: 200 PERSON 35 ITEM 2 CATS WINSTEPS 4.0.1

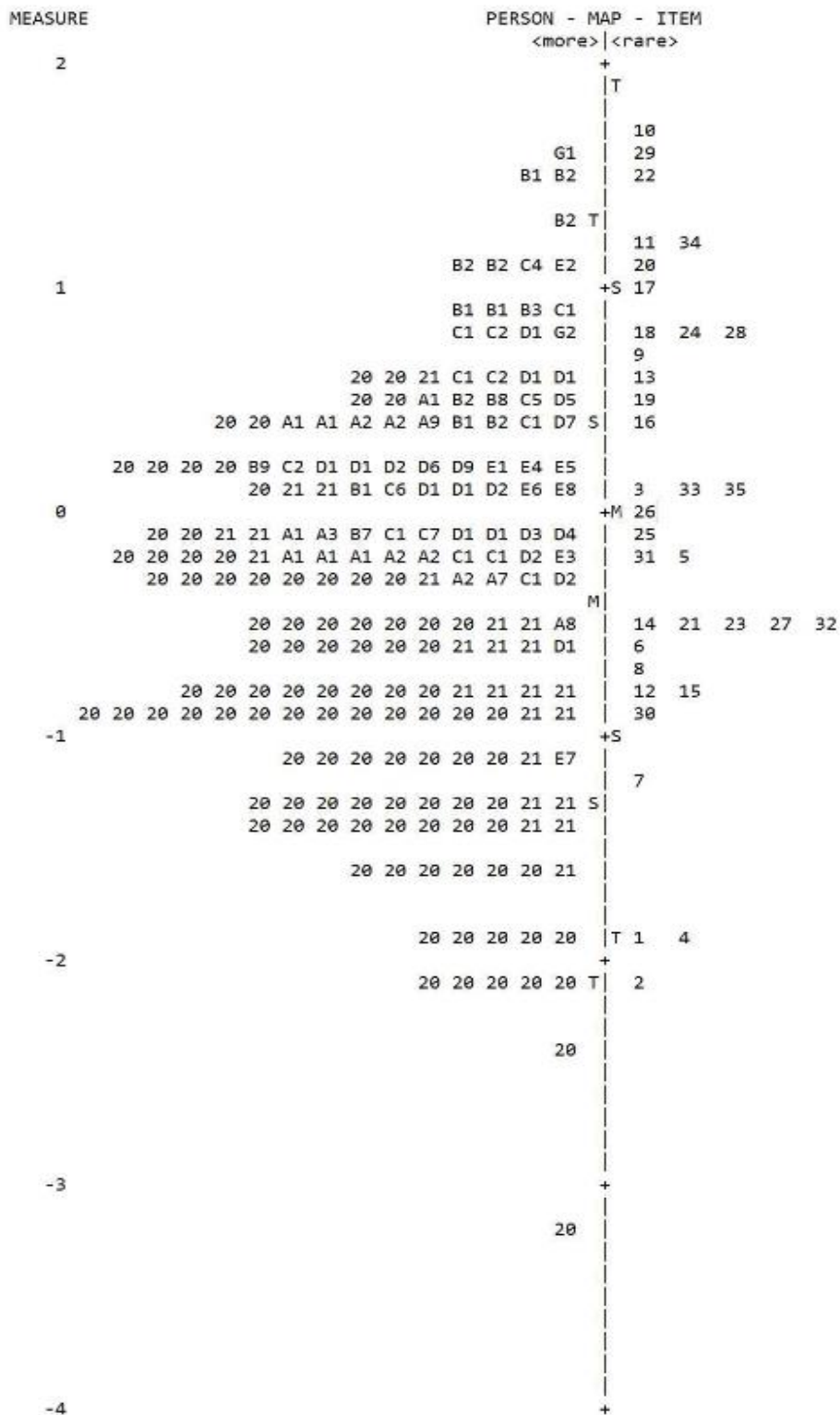


Figure 4.2. Item-person map for the LCLT

Test-takers, as shown in **Error! Reference source not found.**, covered a wide range of ability levels from -3.21 logits to 1.64 logits. The majority of both test takers and items landed on the area above -1 logit. Item 10, which was the most difficult item, was located on top of the map and did not have any test-takers to match. Item 29 and 22 can also be considered very difficult items, because of their locations on the map and the number of test-takers they could be matched with (one for item 29 and two for item 22). The next 28 items from -.90 to 1.17 logit were relatively well-matched with the test-takers placed in this area.

Below -1 logit was the area where 48 test-takers were positioned. The most striking result observed from **Error! Reference source not found.** was that these candidates who made up almost one quarter of the total number of test-takers were matched with only 4 items (item 7, 1, 4 and 2). With much fewer items at their levels on the scale, the ability of these candidates was not as reliably measured as those located between -1 to 2 logit.

#### 4.2.3. Discriminating power of LCLT items

In the Rasch model, item discrimination refers to the extent to which an item can differentiate between examinees having abilities below the item location and those with abilities above the item location (Baker & Kim 2017). Item discrimination can be ranked according to Baker's (2001) interpretations of values for discrimination as follows:

0	None
0.01 – 0.34	Very low
0.35 – 0.64	Low
0.65 – 1.34	Moderate
1.35 – 1.69	High
1.70 and above	Very high
+ infinity	Perfect

Table 4.4 presents the discrimination index for each item of the LCLT and the corresponding interpretation.

**Table 4.4.** Discrimination indices for LCLT items

Entry number (Item)	Discrimination	Interpretation
10	.90	Moderate
29	.82	Moderate
22	.82	Moderate
34	.67	Moderate
11	.64	Low
20	.91	Moderate
17	.74	Moderate
28	.91	Moderate
18	.73	Moderate
24	.87	Moderate
9	.62	Low
13	.81	Moderate
19	.89	Moderate
16	1.01	Moderate
3	1.42	High
35	.67	Moderate
33	1.27	Moderate
26	1.26	Moderate
25	1.38	High
5	1.82	Very high
31	1.16	Moderate
23	.96	Moderate
27	1.11	Moderate
21	1.39	High
14	.76	Moderate
32	1.02	Moderate
6	1.00	Moderate
8	1.52	High
12	.92	Moderate
15	1.49	High
30	1.42	High
7	1.49	High
1	1.19	Moderate
4	1.17	Moderate
2	1.25	Moderate.

In Table 4.4, items were placed in the same difficulty hierarchy as in Table 4.3 so that a general comparison between item difficulty and item discrimination can be made. As seen from the table, the discriminating indices of most items belonged to the moderate range. Specifically, there were 25 items in this range, equivalent to 71.4% of the test. Item 9 and 11 were the ones with the lowest discrimination indices (.62 and .64 respectively), which suggested that they did not work well in separating test-takers' ability levels. Seven items (item 3, 25, 21, 8, 15, 30, 7) had high discriminating power and they represented 20% of the test. Item 5 was the only item with a very high discrimination index, which means that among all items of the test, it worked best in separating test-takers' listening ability. With regard to item difficulty, it was clear from Table 4.4 that higher difficulty measures did not determine higher discrimination indices. Located on the upper half of the table (from item 16 and above) were items with high difficulty measures; however, their discrimination indices were only moderate or low. On the other hand, in the lower half of the table (from item 3 and below), items with lower difficulty measures were found to discriminate test-takers much better.

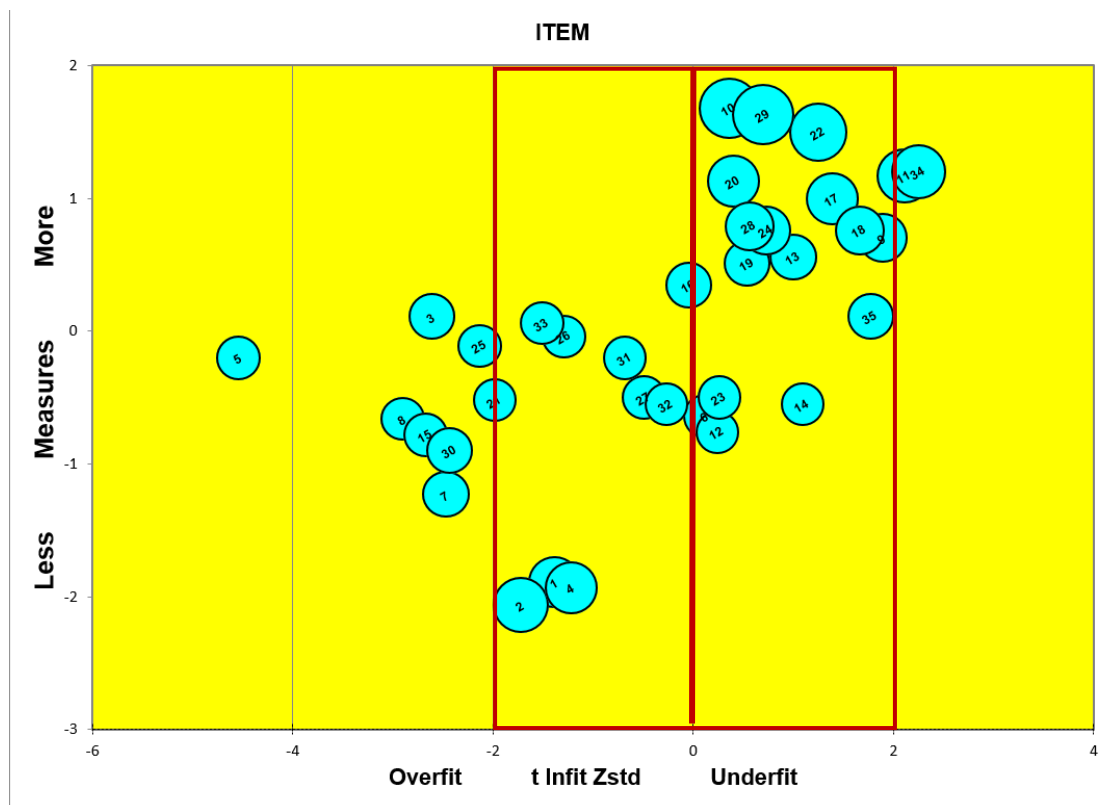
#### **4.2.4. Evidence of construct-irrelevant variance**

In the investigation of construct-irrelevant variance, the analysis of item fit has an important role to play. Table 4.5 presents both infit and outfit statistics for item in the LCLT. While the infit statistics are sensitive to test-takers' responses to items targeted their ability, outfit statistics are highly sensitive to lucky guesses and careless mistakes (Baghaei & Amrahi, 2011). Since this study did not focus on cases of lucky guesses or careless mistakes, only infit statistics will be discussed in detail.

**Table 4.5.** Fit statistics of LCLT items

Entry number (Item)	Infit		Outfit	
	MNSQ	ZSTD	MNSQ	ZSTD
10	1.05	.4	1.37	1.6
29	1.09	.7	1.66	2.6
22	1.16	1.3	1.36	1.7
34	1.25	2.3	1.66	3.4
11	1.23	2.1	1.37	2.0
20	1.04	.4	1.18	1.1
17	1.13	1.4	1.39	2.4
28	1.04	.6	1.09	.8
18	1.14	1.7	1.23	1.8
24	1.06	.7	1.11	.9
9	1.15	1.9	1.40	3.0
13	1.07	1.0	1.17	1.5
19	1.04	.5	1.12	1.1
16	1.00	.0	.99	-.1
3	.86	-2.6	.87	-1.6
35	1.10	1.8	1.18	2.2
33	.92	-1.5	.90	-1.3
26	.93	-1.3	.89	-1.5
25	.89	-2.1	.89	-1.5
5	.78	-4.5	.76	-3.8
31	.96	-.7	.94	-.9
23	1.01	.3	1.01	.1
27	.97	-.5	.96	-.5
21	.90	-2.0	.88	-1.9
14	1.06	1.1	1.09	1.4
32	.99	-.3	1.02	.3
6	1.01	.1	.99	-.2
8	.85	-2.9	.85	-2.4
12	1.01	.2	1.05	.7
15	.86	-2.7	.82	-2.7
30	.87	-2.4	.85	-2.1
7	.84	-2.5	.74	-2.9
1	.87	-1.4	.77	-1.6
4	.88	-1.2	.77	-1.6
2	.82	-1.7	.62	-2.6

As shown in Table 4.5, all infit MNSQ values fell within the range between 0.5 and 1.5; however, three of them (item 5, 34 and 11) did not meet the stringent fit range between 0.8 and 1.2 for high-stakes multiple-choice questions (Linacre, 2012). Item 5 had low infit MNSQ (.78), which suggests that test-takers' responses to this item were too predictable. On the other hand, item 34 and 11 had high MNSQ values (1.25 and 1.23 respectively), which implies that test-takers' responses to this item was very unpredictable. This result indicates the presence of construct- irrelevant variance. More insights into item fit and the presence of construct-irrelevant variance are provided in the bubble chart representing infit ZSTD (Figure 4.3).



**Figure 4.3.** Bubble chart representing infit ZSTD of LCLT items

In Figure 4.3, difficulty estimates are read vertically on the logit scale and items are placed in ascending order, from the easiest one (item 2) to the most difficult one (item 10). Fit values are represented horizontally with acceptable values from -2 to +2 (the red

rectangular). The vertical red line starting from 0 is the theoretical straight line that represents the unidimensional construct (Bond & Fox, 2001). Items which deviate unacceptably from this line, beyond the range between -2 to +2 do not fit the unidimensional construct. The size of each bubble depicts measurement error: the bigger the circle, the larger the error. As can be seen, the most difficult items (item 10, 29, 22, 11, 34) tended to have larger errors. It was likely that the correct answers for these items were sometimes achieved by lucky guesses or test-taking strategies rather than the candidates' listening ability.

Misfitting items can be clearly identified from Figure 4.3. The items that require immediate attention are the underfitting ones with ZSTD larger than 2. Item 11 and item 34 are underfitting and the presence of construct-irrelevant variance in these items is further evidenced by their locations on this bubble chart. Item 9 should also be taken into consideration since its location suggests that this item might have been contaminated by factor(s) irrelevant to the construct. Even though the infit ZSTD values for item 9 which was 1.9 is still within the acceptable range, Figure 4.3 shows that this item did not fit the model's expectation perfectly. This item therefore requires further analysis so as to identify the construct-irrelevant factor(s) it may carry. At the same time, item 5, 3, 8, 15, 30, 7, 25, 21 were found to overfit the model which means test-takers' responses to these items were too predictable. They all had their ZSTD values lower or equal to -2.

Thus far, item fit analysis revealed that construct-irrelevant variance was present in only three items. This result indicates that the LCLT were not seriously contaminated by construct-irrelevant variance. Nevertheless, further analysis is still needed to identify the construct-irrelevant factors that these underfitting items may induce. Eight overfitting items were spotted from the bubble chart (Figure 4.3). Although these items were relevant to the intended construct, they might not be very useful measures of test-takers' ability since the response patterns to these items were too predictable.

#### 4.2.5. Reliability of the LCLT scores

The Rasch model measure reliability for both test-takers and test items with two reliability indices known as person reliability and item reliability (Bond & Fox, 2007). These reliability indices for the LCLT are presented in Table 4.6.

**Table 4.6.** Reliability indices for the LCLT

PERSON									
	200	INPUT	200	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD	
MEAN	14.8	35.0	-.41	.41	1.00	-.1	1.06	.0	
P.SD	5.4	.0	.84	.06	.20	1.2	.39	1.3	
REAL RMSE	.42	TRUE SD	.73	SEPARATION	1.75	PERSON RELIABILITY	.75		
ITEM									
	35	INPUT	35	MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD	
MEAN	84.3	200.0	.00	.17	1.00	-.3	1.06	.0	
P.SD	37.1	.0	.97	.02	.12	1.7	.25	1.9	
REAL RMSE	.17	TRUE SD	.96	SEPARATION	5.55	ITEM RELIABILITY	.97		

As can be seen, the LCLT had very high item reliability (.97), which suggests that the hierarchy of difficulty for LCLT items (see Table 4.3) would be very likely to remain the same if the test was given to a different group of test-takers with similar ability levels. In the meantime, the person reliability index for the LCLT was .75 which was acceptable. This result indicates that we can be relatively confident in the order of person ability according to their scores on the LCLT (see **Error! Reference source not found.**). However, Linacre (2012) emphasizes that person reliability should be 0.8 for a high-stakes test. Therefore, although the LCLT is shown to have acceptable person reliability index, this value is still not good enough since it is a high-stakes test. It should also be noted that if the person reliability of a test is less than 0.8, that test needs more items (Linacre, 2012).



### **4.3. Assessment of the evaluation and generalization inferences**

This section discusses the research findings in relation to the four warrants of the evaluation and generalization inferences. It will clarify how the evidence elicited from the quantitative analysis of test scores supports or refutes each warrant.

#### **4.3.1. Inappropriate difficulty levels**

*Warrant: Item difficulty was appropriate for test-takers.*

When descriptive statistics were reported for the performance of 200 test-takers on this LCLT, it was noted that the scores ranged from 2 to only 28 out of 35. There was only one test-taker who scored 28 and that person belonged to the group of English majors. The fact that the highest score was still 7 points away from the perfect score suggested that this test was very difficult even for English majors of the host university. The results of Rasch analysis shed further light on the relation between the ability of test-takers and the difficulty of the items in this test. Three items (item 10, 22 and 29) were found to be extremely difficult for this group of test-takers. Meanwhile, 28 other items had a relatively good match with the majority of test-takers whose ability levels were above -1 logit on the item-person map. That left the remaining 48 test-takers, almost a quarter of the test-taking population, with only 4 items targeting their ability levels. This being the case, there was not enough items to distinguish between these test-takers and thus, not enough information to provide an accurate ability estimate (Bond & Fox, 2001). The shortage of test items with difficulty levels suitable for this group of low-scoring test-takers raised concerns about the precision of test scores as a measure of these test-takers' listening ability.

Thus far, both descriptive statistics and Rasch analysis have showed that the LCLT was very difficult for this group of test-takers. It contained items with high difficulty levels which corresponded with higher-proficiency test-takers and tended to neglect the group of lower-proficiency ones. The test was found to be very difficult even for English majors of the

host university. Taken together, these findings did not support this warrant which holds that item difficulty was appropriate for test-takers.

#### **4.3.2. Reasonable discriminating power**

*Warrant: The LCLT items had reasonable discriminating power*

The discrimination indices of the LCLT items ranged from low to very high. Only two items had low discriminating power, 25 had their discrimination indices in the moderate range, 7 had high discriminating power and only 1 was found to have very high discrimination index. Taken together, the LCLT items had moderate discriminating power. This result, though quite modest, was still sufficient to support this warrant.

#### **4.3.3. Relevance between test items and the targeted listening construct**

*Warrant: Test items were relevant to the targeted listening construct.*

From the item fit analysis, item 11, 9 and 34 were identified as the ones threatened by construct-irrelevant variance. Some might say that these three items should be omitted since they were not relevant to the construct and thus not useful measures of listening ability. However, as Bond and Fox (2007) strongly emphasize, the identification of misfitting items should be used to assist in the detection of problematic items and person performances, not just to decide which item should be deleted from the test. Further investigation on these items is needed to shed light on what particular construct-irrelevant factors were present and how they can be avoided in the future.

Apart from these three misfitting items, all other items on the test were found to be relevant to the targeted listening construct. Nevertheless, eight items (item 3, 5, 7, 8, 15, 21, 25, 30) were identified as overfitting which means they were relevant but too predictable to be useful measures of test-takers' ability. These items should also be carefully considered before a decision is made about whether they should be omitted because deleting overfitting items could rob the test of its best items (Bond & Fox, 2007).

Overall, with three items that were found to misfit the Rasch model, construct-irrelevant variance did not seem to be a major threat to this test. This warrant is supported because the majority of test items was found to be relevant and useful for measuring the targeted listening construct.

#### **4.3.4. Unsatisfactory person reliability**

*Warrant: The LCLT had high item and person reliability indices.*

While the LCLT item reliability was high (.97), its person reliability index (.75) was not high enough for a high-stakes test of language proficiency. This result highlights the need for more items in the LCLT since 35 items did not seem enough to cover the entire construct and provide reliable indicators of test-takers' ability. This warrant is therefore not supported.

### **Summary**

Table 4.7 summarizes the evidence and the decisions made for each warrant in the evaluation and generalization inference.

**Table 4.7.** Evidence and judgements on the evaluation and generalization inferences

Inference	Warrant	Evidence	Decision
Evaluation	Item difficulty was appropriate for test-takers.	<ul style="list-style-type: none"> <li>- Maximum score was only 28 out of 35.</li> <li>- Test items tended to focus more on higher proficiency levels.</li> <li>- The listening ability of 48 low-scorers were not adequately represented by test items.</li> </ul>	Item difficulty was not appropriate for test-takers. <b>This warrant is not supported.</b>
	Test items had reasonable discriminating power.	<ul style="list-style-type: none"> <li>- 25 items had moderate discrimination indices.</li> <li>- 7 items had high discrimination indices.</li> <li>- 1 item had very high discrimination index</li> </ul>	<b>This warrant is supported.</b> However, discriminating power should still be improved for this multiple-level listening test.
	Test items were relevant to the targeted listening construct.	<ul style="list-style-type: none"> <li>- 32 items were relevant to the listening construct (they were not found to contain construct-irrelevant variance)</li> <li>- Only 3 items were threatened by construct-irrelevant variance.</li> </ul>	Construct- irrelevant variance was not a major problem with the LCLT. <b>This warrant is supported.</b>
Generalization	Test-taker's scores on the LCLT was reliable.	<ul style="list-style-type: none"> <li>- Person reliability was .75</li> </ul>	Not high enough for a high-stakes test. <b>This warrant is not supported.</b>

As seen in Table 4.7, while the first warrant for the evaluation inference were refuted, the other two were supported. Specifically, while the test was found to be very difficult for the test-takers, it had reasonable discriminating power and contained little construct-

irrelevant variance. The evaluation inference therefore was supported but only to a certain extent. With discernible evidence that item difficulty was not appropriate for test-takers, there is still a concern about the accuracy of assessment. As the only warrant for the generalization inference was not fully supported, this inference did not hold eventually.

This study, at a surface level, brought to light some issues which are worthy of attention and further investigation. The high level of difficulty is a major concern but it is not yet clear which factors contributed to the cognitive demand of the LCLT items. Although construct-irrelevant variance did not seem to seriously affect the test, this finding needs to be confirmed by qualitative evidence. The sources of construct-irrelevant variance should also be identified if they are to be avoided for future LCLTs. These issues will be addressed in Chapter 5 which reports on the qualitative study on the test-takers' cognitive processes.

## **Chapter 5. The test-takers' cognitive processes**

In the validity argument for the LCLT, the explanation inference requires two different lines of inquiry and a considerable amount of empirical evidence. This chapter reports on only the qualitative study that elicits evidence for the first warrant which holds that the LCLT items actually engaged the listening subskills described for level 3, 4 and 5 of CEFR-V. In this study, verbal report was used to gain insights into the test-takers' cognitive processes as they were tackling test tasks. This chapter provides, first of all, an overview of verbal report as a research methodology. Following this review, the test specifications and the actual content of the LCLT are presented. Another important part is dedicated to the description of the participants and how their verbal reports were collected. The findings of this study are then reported to lay the ground for a decision to be made about how well the LCLT items actually engaged the targeted listening subskills.

### **5.1. Verbal report in theory and in language testing research**

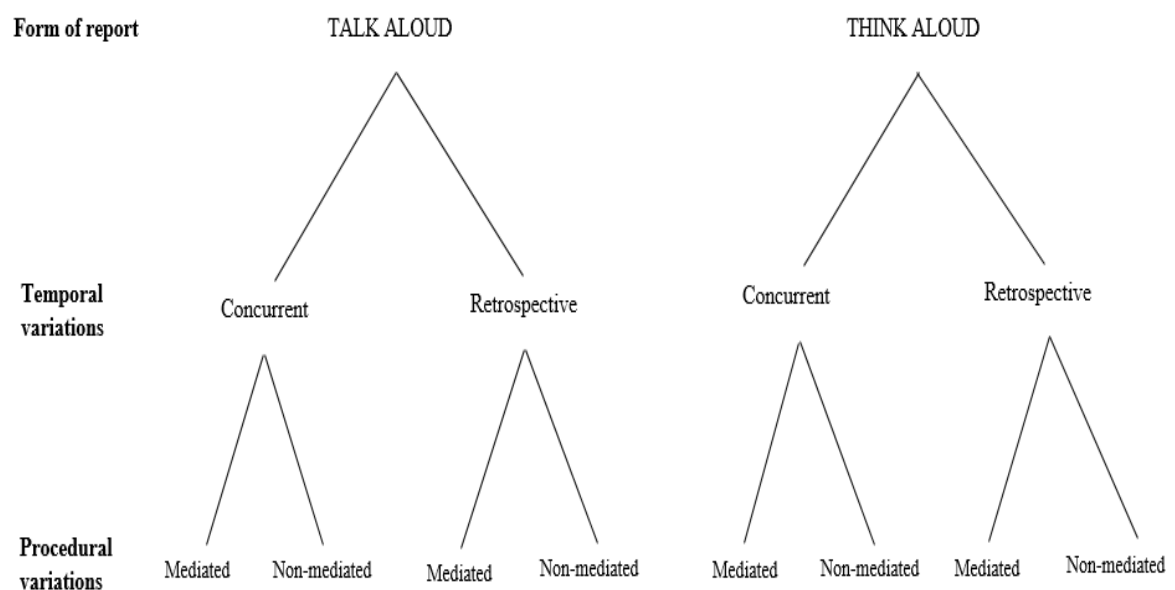
This section presents the theoretical background of verbal report and reviews relevant studies in language testing that used verbal report as the main research method.

#### **5.1.1. Theoretical background of verbal reports**

As a research methodology, verbal report is a way of gathering data from an individual under special conditions in which the person is asked to talk aloud or think aloud while or after completing a given task (Green, 1998). Verbal reporting offers individuals an opportunity to vocalize what is going through their minds as they are performing a task or solving a problem. It allows researchers to observe the similarity or differences in the way each individual approaches a problem (Gass & Mackey, 2000). The collection of verbal reports has become vital in a wide range of research areas such as psychology, counselling, treatment of autism, speech pathology, neurology, and drug and alcohol addiction (Bowles,

2010). In linguistics, this method has been used with increased frequency to investigate many aspects of first and second language acquisition.

Verbal reports can be gathered in different ways, under different conditions, depending on the research purposes and research questions (Green, 1998). It is highly important that researchers are aware of the differences between various types of verbal reports so that they could decide on the appropriate procedure for their research. Green (1998) offered a useful overview of popular categories of verbal reports (Figure 5.1).



**Figure 5.1.** Various types of verbal report procedure

As seen in Figure 5.1, verbal protocols can be divided into talk-aloud and think-aloud. Although these terms are often used interchangeably, it is sometimes important to distinguish between the two. Ericsson and Simon (1993) clarify that in a talk-aloud protocol, individuals are asked to say out loud what they are saying silently to themselves. In this case, the report will include information that is already encoded in verbal form (Green, 1998). A typical example is a task in which subjects rehearse auditorily presented phonemic information (Ericsson & Simon, 1993, p.266) as follows:

Auditory presented digits	05545550
Protocol	0454550
	04555450
	04545550

On the other hand, a think-aloud protocol requires participants to “utter thoughts that may have been held in memory in some other form (e.g., visually)” (Ericsson & Simon, 1993, p. 222). Green (1998) points out that while carrying out a task, subjects are likely to attend to not only verbal information but also non-verbal visual, auditory or tactile information. Such non-verbal information can be transformed and then verbalized in a think-aloud protocol. Therefore, the think-aloud method appears to have advantages over talk-aloud for the breadth and depth of information gained from verbal reports. However, the decision on which form of verbal report to use depends on the nature of the research. Ericsson and Simon (1993) note that individuals may not always distinguish between the two so clear and precise instructions are a must.

When it comes to temporal variations, verbal reports fall into two categories: concurrent and retrospective. Concurrent reports are generated at the same time as the individual is tackling the task while retrospective reports are gathered after the individual has completed the task. Between these two types of verbal reports, concurrent reports will be more complete and accurate since participants who think aloud during a task are not subjected to memory decay (Bowles, 2010). For retrospective reports, the interval between task completion and start of verbal report is important since the longer the delay, the more likely it is for memory to fade. In research that uses retrospective reports, immediate retrospection can be used to minimize the effect of memory on the verbal reports. Ericsson and Simon (1993) state that during performance on a task, a subset of the sequence of thoughts is stored in long-term memory in the form of a retrievable trace of episodic memory. They posit that “if the subjects are asked to report on their last previous cognitive processes,



considerable episodic memory can be retrieved from information and cues in short-term memory” (Ericsson & Simon, 1993, p.149). Thus, immediate retrospection can still be expected to reflect closely the thought process involved in task performance (Wuthe inter, 1998).

Verbal reports can be mediated or non-mediated. In the first case, the participant may be asked questions about the task so that s/he can explain the thinking process or justify the answer. The questions usually target the reason(s) why the participant uses a particular approach to solve the problem, or the knowledge and skills s/he employs to complete the given task. Gass and Mackey (2000) discuss stimulated recall as a special variation of mediated verbal report. This method creates conditions in which participants are provided with stimuli such as an audio or video recording of themselves carrying out a task, their own observation field notes or even transcription of conversations. As they hear and see these stimuli, participants were asked to recall their thought processes during the original event. This method will be useful in cases when concurrent verbal reports are not possible such as with speaking tasks (Xi, 2008). It is also one good way to overcome the memory effect associated with retrospective verbal reports. On the other hand, in non-mediated verbal reports, participants are asked to think aloud by themselves and are prompted only when they seem to struggle. The prompts should be as non-intrusive as possible and may include encouragement like ‘keep talking’ for instance (Green, 1998).

### **5.1.2. Verbal reports in research on listening tests**

When it comes to listening tests, verbal protocols can be used for various purposes such as investigating the cognitive processes involved in listening, identifying factors affecting item difficulty, comparing different stimulus materials and different test formats (Green, 1998). Buck (1991) used immediate retrospective verbal protocol to collect a considerable amount of useful information on processes involved in listening. This method

was also employed by Wu (1998) to investigate the effects of the multiple-choice format on the performance of Chinese EFL students as they were sitting a listening test. Taking a similar approach, Field (2005) conducted a study on the cognitive validity of the lecture-based questions in the IELTS Listening paper. He then continued to use immediate retrospective verbal report in another study on the effects of single and double play upon listening test outcomes and cognitive processing (Field, 2015). Along the same line, Shin (2006) relied on retrospective reports to elicit evidence of construct validity of listening test items in the Michigan English Language Assessment Battery. Apart from this type of verbal report, stimulated recall has proved to be useful for research on listening assessment as well. In a study that examined the cognitive validity of the Aptis Listening test, stimulated recall was used to unveil the cognitive processes that test-takers underwent to answer items on the test (Holzknecht & Eberharter et al., 2017). Winke and Lim (2014) used this method in combination with eye-tracking to investigate the extent to which testwiseness and test anxiety affected performance on the IELTS Listening test.

Thus far, it is evident that verbal report has become the method of choice for many researchers who ventured into validating listening tests. Due to the nature of listening skill, only retrospective verbal report is possible. So as to minimize the effect of memory, most researchers choose to use either immediate retrospective verbal reports or stimulated recall. The use of verbal report is not limited to only cognitive processes but extended to the investigation of other aspects such as the effects of test format, testwiseness and test anxiety on test-takers' performance.

Despite its usefulness in research on cognitive processes, verbal report is not uncontroversial (Green, 1998). Field (2005) points out a number of drawbacks it has as a method of researching language skill performance. He emphasizes that reading and listening skill can only be investigated indirectly. In case of listening, it is clearly impossible for

participants to engage in concurrent verbal report. Thus, retrospective report is the only possibility; however, it carries the memory effect. Moreover, the language barrier may prevent non-native participants from reporting as fully as they are expected to.

### **5.1.3. The validity and reliability of verbal reports**

The major concerns about verbal reports arise from the validity and reliability of not only the technique itself but also the data. The validity of verbal reports depends largely on the extent to which the reports correspond with the actual thinking process while a task is being carried out. Ericsson and Simon (1993), after reviewing a large number of studies, come to a conclusion that if the technique is used appropriately, verbal protocol analysis can be valid and useful. They argue that as long as the instructions were clear, the act of thinking aloud does not interfere with and change a person's cognitive processes while performing the task. However, there has been no possible way to prove that verbalized information truly reflects the cognitive process involved in completing a certain task. Validity of the technique can only be maximized by a well-prepared procedure in which appropriate and clear instructions are used to guide the production of verbal reports (Güss, 2018). Apart from that, a delay between completing a task and producing the verbal reports can result in information loss from memory, the addition of new information after the task has been completed or the alteration of the original information. Most of these problems can be avoided by concurrent verbal reports; however, they are considerable concerns for retrospective reports.

The reliability of this technique refers to consistency, the ability to collect the same data at a different time (Güss, 2018). In order to elicit evidence of the reliability of the technique, one subject can be asked to work on the same tasks several times then the verbal reports will be compared. In fact, there are only a few studies on the reliability and consistency of verbal reports.

Once the reports have been collected, they will be coded and it is this coding stage that contains certain threats to the validity and reliability of the encoded data. The validity of coding addresses the issue of whether or not a code for a given protocol truly captures the cognitive processes as understood from the verbalized information (Green, 1998).

Meanwhile, the issue of reliability revolves around the possibility that the same data might be coded in the same category by two independent coders or by the same individual coding the set of protocol twice. The reliability of encoded data might be affected by transcriber and coder variables as well. As recommended by Güss (2018), to minimize problems related to transcribing, this job should be done by native speakers of the participants' language. As far as coders are concerned, Green (1998) lays stress on the importance of establishing inter-coder reliability and intra-coder reliability. It was also found that retrospective reports are less reliable than concurrent reports because of failures to recall the correct thought process for the previous decision (Fidler, 1983). Therefore, concurrent verbalizations should probably be the preferred method so as to ensure the reliability of data.

Generally speaking, verbal report is a labour-intensive research method. The preparation, collection, coding and analysis of data are much more time-consuming compared to other methodologies. However, the breadth and depth of information acquired about the cognitive processes that subjects are engaged in during task performance is well worth the time and effort. In this particular study, the limitations of verbal reports are hoped to be compensated for by the rich insights into the test-takers' thought processes.

## **5.2. The specifications and actual design of the LCLT**

This section provides important insights into the underlying construct of the LCLT and how this construct was operationalized for the purpose of test development. First, the test specifications, which lay the groundwork for the development of the LCLT, are provided.

Another focus of this section is the way in which the test items and tasks were presented to test-takers in the real test event.

### 5.2.1. Descriptors for listening skill in the CEFR-V

The descriptors for listening in the CEFR-V are very much similar to the original descriptors that can be found in the CEFR. Since the LCLT targets level 3 to level 5 of the CEFR-V (which are equivalent to B1 to C1 in the CEFR), it will be most relevant, at this stage, to provide the descriptors for these levels (Table 5.1).

**Table 5.1.** General descriptors for listening in CEFR-V

Level	Descriptors
Level 3 (B1- CEFR)	Can understand straightforward factual information about common every day or job-related topics, provided speech is clearly articulated in a generally familiar accent.
	Can understand the main points of standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives, provided speech is clearly articulated in a generally familiar accent
Level 4 (B2 – CEFR)	Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic and vocational life.
	Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics, including technical discussions in his/her field of specialization.
	Can follow extended speech and complex lines of argument provided the topic is reasonably familiar and the direction of the talk is sign-posted by explicit markers.
Level 5 (C1 – CEFR)	Can understand enough to follow extended speech on abstract and complex topics even when the talk is not clearly structured and the relationships between ideas are not signalled explicitly.
	Can understand necessary information of public announcements

Descriptors for listening are then provided for four certain contexts:

- Listening to conversations between native speakers;
- Listening to lectures and presentations;
- Listening to announcements and instructions;
- Listening to audio media and recordings.

These descriptors can be found in Appendix I. Together with the overall descriptors in Table 5.1, they are taken as the target construct and form the basis for the development of the LCLT.

### **5.2.2. The LCLT specifications**

Once the construct of the LCLT has been identified, it is operationalized in terms of which particular subskill that each test item should target. The LCLT, with 35 multiple-choice questions (MCQ), is expected to measure a total number of 16 listening subskills which represent listening ability from level 3 to level 5 in the CEFR-V. The LCLT specifications provide details of how these subskills should be covered by the test items (Table 5.2). It is clear from these specifications that Part 1 focuses mainly on level 3 with 7 out of 8 items targeting this level. In Part 2, the majority of items (8 out of 12) represents level 4. The main target of Part 3 is level 5 with 12 out of 15 items set at this level. With this being the case, the LCLT items are expected to be arranged in order of increasing difficulty. As seen in Table 5.2, the LCLT covers 3 text types, including:

- announcements, instructions
- conversations (short and long)
- monologues (talks/lectures)

**Table 5.2.** The LCLT specifications

Part	The input text	Listening subskills targeted
Part 1	<ul style="list-style-type: none"> <li>• 08 Short and clear conversations, announcements, instructions.</li> <li>• Length: 30-60 words</li> <li>• Familiar topics related to everyday life</li> <li>• Level of input: level 3</li> <li>• Number of items: 08</li> </ul>	<ul style="list-style-type: none"> <li>• Understanding the main points of short, clear announcements, instructions or conversations (Level 3, targeted by 2 MCQs)</li> <li>• Understanding details of discussion around him/her provided speech is clearly articulated (Level 3, targeted by 4 MCQs)</li> <li>• Understanding enough to follow detailed directions or instructions (Level 3, targeted by 1 MCQ)</li> <li>• Listening and making simple inferences about the context (Level 4, targeted by 1 MCQ).</li> </ul>
Part 2	<ul style="list-style-type: none"> <li>• 03 conversations between native speakers or fluent speakers of English.</li> <li>• Length: 2 minutes (each)</li> <li>• Topic: social, occupational, educational</li> <li>• Level of input: level 3, 4, 5</li> <li>• Number of items: 12</li> </ul>	<ul style="list-style-type: none"> <li>• Understanding the main points of clear, standard speech (Level 3, targeted by 1 MCQ).</li> <li>• Understanding the main ideas of propositionally and linguistically complex speech on familiar topics (Level 4, targeted by 1 MCQ).</li> <li>• Understanding the main ideas of a discussion/conversation even when content is not straightforward and clearly structured (Level 5, targeted by 1 MCQ).</li> <li>• Understanding details of discussion around him/her when they are not clearly articulated (Level 4, targeted by 6 MCQs)</li> <li>• Listening and making inferences about the speaker's attitude and opinion through his/her tone and language use (Level 4, targeted by 1 MCQ).</li> <li>• Understanding discourse function and the purpose of the speaker (Level 5, targeted by 2 MCQs)</li> </ul>
Part 3	<p>3 talks/lectures by native speakers or fluent speakers of English</p> <p>Length: 2-3 minutes (each)</p> <p>Topic: social, educational, occupational</p> <p>Level of input: Level 4, 5</p> <p>Number of items: 15</p>	<ul style="list-style-type: none"> <li>• Understanding new words in context (Level 4, targeted by 1 MCQ)</li> <li>• Understanding the main points and structure of propositionally and linguistically complex speech including technical discussions (Level 4, targeted by 2 MCQs)</li> <li>• Understanding the purpose of the speaker even when it is not clearly and directly articulated (Level 5, targeted by 2 MCQs)</li> <li>• Understanding details in propositionally and linguistically complex speech (Level 5, targeted by 7 items)</li> <li>• Making inferences about the speaker's attitude and opinion when they are only implied and not signalled explicitly (Level 5, targeted by 1 MCQs)</li> <li>• Understanding idiomatic expressions (Level 5, targeted by 1 MCQ)</li> </ul>

### 5.2.3. The actual design of the LCLT

The design of the LCLT follows the overall format as determined in the test specifications with 3 parts and 35 MCQs. Table 5.3 provides an overview of the actual test content.

**Table 5.3.** The actual content of the LCLT

<b>Part</b>	<b>Content</b>	<b>Item</b>
Part 1 (8 short conversations)	Everyday topics: transport, shopping, holiday, birthday invitation	Item 1 to item 8 (one item for each conversation)
Part 2 (3 long conversations between native speakers)	Conversation 1: Three people discussing a movie	Item 9 to item 12
	Conversation 2: Interviewing a singer in a famous band	Item 13 to item 16
	Conversation 3: Interviewing an engineer	Item 17 to item 20
Part 3 (3 long talks given by native speakers)	Talk 1: The Smithsonian museum	Item 21 to item 25
	Talk 2: Volcanoes in the US	Item 26 to 30
	Talk 3: A students' concert	Item 31 to 35

The 35 MCQs in the LCLT were written in the traditional format. Each question has an item stem and four response options. Among the four alternatives, one is the correct answer and the other three are distractors. It is important to note that this test offers full question preview in which both item stem and response options are showed prior to listening.

All instructions are clearly given in both written and spoken forms. In Part 2 and 3 of the LCLT where long conversations and talks are included, a brief description of the context



is given before each conversation or talk. For example, prior to listening to a long conversation in Part 2, candidates can hear and read:

*“You will hear an interview with an engineer called Roger Moffet, whose working life has changed dramatically over the past ten years.”*

Before listening to a lecture in part 3, this short introduction is given in both written and spoken form:

*“You will hear a lecture about volcanoes in the US.”*

The introduction of the context helps to prepare and familiarize test-takers with the input and the test tasks.

### **5.3. Aims and methodology of the study**

This study focuses on the warrant which holds that the LCLT items actually engaged the listening subskills described for level 3, 4, and 5 of the CEFR-V. Backing for this warrant is evidence that the 16 listening subskills listed in the test specifications were adequately engaged by the LCLT items. Such evidence was elicited by analysing test-takers’ verbal reports. This section provides details of the participants as well as the procedure for collecting and analysing their think-aloud protocols.

#### **5.3.1. Participants**

This study recruited a total number of 10 participants who were English majors at the host university. At the time of the study, they were in their third and final year. These participants never took the LCLT before. The information sheet and consent form are provided in Appendix E and F. Given that confidentiality and anonymity are of paramount importance in my research, participants were given names other than their real names. The list of participants’ pseudo names and their background information is provided in Table 5.4.

**Table 5.4.** Background information of the participants

<b>Name</b>	<b>Age</b>	<b>Gender</b>	<b>First language</b>
Chi	21	Female	Vietnamese
Trang	21	Female	Vietnamese
Thu	22	Female	Vietnamese
Vy	20	Female	Vietnamese
Mai	20	Female	Vietnamese
Nam	19	Female	Vietnamese
Nga	20	Female	Vietnamese
Van	21	Female	Vietnamese
An	20	Female	Vietnamese
Minh	20	Female	Vietnamese

In this study, the participants were not selected according to any set of criteria. Since the nature of the study is exploratory, any subject will be useful for the insights they can provide about what is going through their minds as they are tackling test tasks. In this case, it is more useful to have participants who are interested in the study and willing to give in-depth information of their cognitive processes (Charters, 2003). This best describes the 10 participants who voluntarily chose to be part of this study.

It is important to note that this group of participants was not familiar with verbal reports. Therefore, each of them was given a 15-minute training session before they started the think-aloud process. In this session, I gave the participant an example of a test task and demonstrated thinking aloud as I was working on the task. After that, each participant had the opportunity to practice until they were confident and ready to take the LCLT and start reporting their thinking process.

### **5.3.2. Collection of verbal reports**

To collect verbal reports for this study, each subject was scheduled for a one-on-one meeting with me. Each meeting started with a 15-minute training session as previously mentioned. The participant then started listening and answered test items at the same time. Whenever an option was circled, the recording was immediately paused so that the participant could start explaining why she came up with that answer. This way, the effect of

memory was minimized and the information obtained through the verbal reports could closely reflect the actual cognitive processes that the participants went through. In case the responses were briefer than had been expected, further questions were asked so that the participants could explain their thinking process in greater detail. These questions were more encouraging than intrusive and only asked when necessary. To ensure that all the information was precisely conveyed, the language used for all the think-aloud sessions was Vietnamese.

With a view to conducting an in-depth investigation of the cognitive processes that test-takers underwent as they answered the test items, time was not controlled in this study. Each participant was encouraged to take their time and report as much information as they could. Indeed, the participants were asked to take a test but in a non-test condition, which certainly made the whole test-taking experience much less authentic. However, this loss in authenticity was compensated for by the depth of information that was obtained from each participant.

Once collected, the verbal reports were transcribed with NVivo 12. Each transcription included both the participant's report and questions from me. Since all the verbal reports were in Vietnamese, they were translated into English after being transcribed. I was responsible for both the transcription and translation of the verbal data.

### **5.3.3. Analysis of verbal reports**

To prepare for the analysis of the verbal reports, a coding scheme was developed by considering all the listening subskills that the LCLT targeted. This process was straightforward since a list of pre-defined listening subskills had already been provided in the test specifications (Table 5.2). For the purpose of developing a coding scheme, the listening subskills targeted by the LCLT were singled out and presented in Table 5.5.

**Table 5.5.** Listening subskills targeted by the LCLT

Part	Targeted listening subskills
Part 1	<ul style="list-style-type: none"> <li>• Understanding the main points of short, clear announcements or conversations</li> <li>• Understanding details of discussion around him/her provided speech is clearly articulated</li> <li>• Understanding and following detailed directions or instructions</li> <li>• Listening and making simple inferences about the context</li> </ul>
Part 2	<ul style="list-style-type: none"> <li>• Understanding the main points of clear, standard speech</li> <li>• Understanding the main ideas of propositionally and linguistically complex speech on familiar topics</li> <li>• Understanding the main ideas of a discussion/conversation even when content is not straightforward and clearly structured</li> <li>• Understanding details of a discussion around him/her when they are not clearly articulated</li> <li>• Listening and making inferences about the speaker's attitude and opinion through his/her tone and language use</li> <li>• Understanding discourse function and the purpose of the speaker</li> </ul>
Part 3	<ul style="list-style-type: none"> <li>• Understanding new words in context</li> <li>• Understanding the main points and structure of propositionally and linguistically complex speech including technical discussions</li> <li>• Understanding the purpose of the speaker even when it is not clearly and directly articulated</li> <li>• Understanding details in propositionally and linguistically complex speech</li> <li>• Making inferences about the speaker's attitude and opinion when they are only implied and not signalled explicitly</li> <li>• Understanding idiomatic expressions</li> </ul>

In the first step of data analysis, each subskill was coded. The coding scheme therefore consisted of 16 encoded categories. In the next step, each verbal report was segmented. Then, each segment was matched to the encoded category that corresponded to the cognitive process represented by that segment. This process was carried out with NVivo 12. Once the data had been coded, the analysis results were generated in terms of two important aspects, the particular listening subskills reported by the participants and the frequency of these subskills in the verbal data.

#### **5.4. Results of verbal report analysis**

The analysis of verbal data brought to light the extent to which the LCLT items covered the targeted listening construct. It also revealed which factors actually affected the participants' performance and in what way. These substantial findings are reported in this section.

##### **5.4.1. Coverage of the target construct**

As evidenced by the think-aloud protocols, the LCLT provided very limited coverage of the intended construct. Among the 16 listening subskills targeted by the test, only five of them were identified in all the participants' verbal reports. The test construct was clearly under-represented because the majority of the targeted listening subskills were not engaged in answering the test items. This finding raises concerns about the validity of the LCLT and the plausibility of the explanation inference. Table 5.6 highlights the listening subskills that were actually engaged in by the LCLT items and the frequency at which they were referred to in the participants' responses.

**Table 5.6.** Listening subskills engaged by the LCLT items

<b>Part</b>	<b>Listening subskills</b>	<b>No. of participants reporting</b>	<b>No. of references</b>
<b>Part 1</b>	Understanding the main points of short, clear announcements or conversations	0	0
	<b>Understanding details of discussion around him/her provided speech is clearly articulated</b>	10	128
	Understanding and following detailed directions or instructions	0	0
	<b>Listening and making simple inferences about the context</b>	10	32
	Understanding the main points of clear, standard speech	0	0
<b>Part 2</b>	Understanding the main ideas of propositionally and linguistically complex speech on familiar topics	0	0
	Understanding the main ideas of a discussion/conversation even when content is not straightforward and clearly structured.	0	0
	<b>Understanding details of a discussion around him/her when they are not clearly articulated</b>	10	163
	<b>Listening and making inferences about the speaker's attitude and opinion through his/her tone and language use</b>	10	62
	<b>Understanding discourse function and the purpose of the speaker</b>	9	9
<b>Part 3</b>	Listening and understand new words in context	0	0
	Listening and identify the main points and structure of propositionally and linguistically complex speech including technical discussions	0	0
	Listen and understand the purpose of the speaker even when it is not clearly and directly articulated	0	0
	Listen and make inferences about the speaker's attitudes and opinions when they are only implied and not signalled explicitly	0	0
	Listen and recognize details in propositionally and linguistically complex speech	0	0
	Listen and understand idiomatic expressions	0	0

As can be seen from Table 5.6, there was no evidence of important subskills which comprise the major part of the construct. Surprisingly, the ability to identify main idea(s) of a conversation or talk was completely neglected. While this ability was mentioned 5 times in the test specifications (Table 5.2), it was not found in the verbal data. All the subskills that were deemed most cognitively demanding and supposed to be targeted by items in Part 3 of the LCLT were also absent from the verbal reports.

Among the five subskills that were evidenced from the participants' responses, the top-rated subskill was "understanding details of a discussion around him/her when they are not clearly articulated". It was reported by all the 10 participants and referred to 163 times in the verbal data. The second ranked subskill was "understanding details of discussion around him/her provided speech is clearly articulated". This subskill was present in all the participants' verbal reports with 128 references. In the third place was the ability to listen and make inferences about the speaker's attitude and opinion through his/her tone and language use. It was identified in all 10 participant's verbal reports with 62 references. The ability to listen and make simple inferences about the context was also reported by all the participants but far less popular with only 32 references. Finally, the ability to recognize discourse function was referred to only nine times by nine participants.

It was clear from this result that the LCLT items engaged mainly the ability to identify details and the ability to make inferences from what is heard. The five listening subskills reported by the participants will be described in the following subsections. For each subskill, certain test items will be subjected to in-depth analysis. These are the items for which the subjects' responses provided the richest insights into their thinking processes. For each test item, the question and the multiple-choice options are given first with the correct answer in bold. Then, the relevant content from the input text is provided. After that, the most relevant parts of the subjects' verbal reports are quoted. My observations are then presented

to supplement the verbal data. Excerpts from the protocols are in italics. Words taken from the input or the test questions are in quotation marks.

#### 5.4.1.1. Understanding clearly articulated details

In the verbal data, there was extensive evidence of the ability to identify details of a discussion when speech was clearly articulated. This ability was referred to 128 times in the participants' responses to certain items in part 1, 2 and 3 of the LCLT (Table 5.7).

**Table 5.7.** Items engaging the ability to understand clearly articulated details

Part	Item
Part 1	Item 1, 2, 4, 6, 7
Part 2	Item 13, 15
Part 3	Item 22, 24, 25, 26, 27, 30

The participants' reports on how they answered item 1, 2, 7, 15, 22 will be subjected to in-depth analysis as follows.

#### **Item 1**

How does the man travel to Liverpool?

- A. By train
- B. By bus**
- C. By car
- D. By plane

Audio script:

- Excuse me, I'd like to go to Liverpool on Friday.
- Well, you can go by bus or train.
- Is the train expensive?
- Yes, the bus is much cheaper. It's only 20 pounds.
- Right, that's better for me. Can I have a ticket, please?



- Yes, certainly.

Number of correct answers: 10/10

The general impression on the participants' verbalizations was that they all felt confident about their answers and reported very quickly what was going through their minds. The following reports from the two participants named Chi and Trang are typical of the responses for this item.

*(1) Chi: I choose B, by bus. In the conversation, the woman gave the man two options, train and bus. She also said that train was more expensive and bus was much cheaper. So, the man chose to go by bus because he said it was better for him.*

*(2) Trang: My answer is B, by bus because when the woman gave the man two options, he asked whether the train was expensive. She said yes and added that bus was much cheaper. He said it was a better choice. So, he chose to go by bus and after that he bought the ticket.*

Through the examples of Chi and Trang, it was apparent that they were both very certain about the answer and able to explain themselves. These participants recalled the content of the conversation without pauses or hesitation. In their explanation, they mainly used the words from the audio script. The particular detail necessary for answering item 1 was well understood at word level and the participants did not have to go any further than retelling what they heard. Bottom-up processing was clearly evidenced in all the 10 responses to this item.

### **Item 2**

Which bill has just arrived?

- A. Water
- B. Gas
- C. Electricity
- D. Phone

Audio script:

- Is that the electricity bill?
- No, it's the water bill.
- Is it very big?
- Not as bad as last time.
- Oh, good.

Number of correct answers: 10/10

For this item, the participants gave responses in which bottom-up processing was key and the detail needed for the answer was easily identified. For example:

*(3) Vy: It was clear, it was the water bill. He thought it was an electricity bill but it turned out to be water.*

*(4) Chi: The information is clear from the conversation. Although they mentioned electricity from the beginning but it was just as question from the first person. And then the second person confirmed that it was not the electricity bill, it was the water bill.*

Vy and Chi both emphasized with certainty how clear the information was. It could be inferred that the details needed to answer the question were totally understandable at word level. These participants justified their answers by retelling or summarizing the relevant part of the conversation.

#### **Item 7**

When will they go on holiday?

- A. June**
- B. July
- C. August
- D. September

Audio script:

- Do you want to come on holiday with me in the summer?
- Sure! I can go in June, July or September.
- Well, June will be best for me. July is too hot and I have to work in September.
- Ok, where should we go?

Number of correct answers: 10/10

The participants' responses to this item included recalling key words and details from the conversation. There was no evidence of high-level thinking that went beyond the word level. For example:

*(5) Thu: Two people were planning their summer holiday together. The person who was asked said that he could go in June, July and September. However, the person who asked said that July was too hot and he had to go to work in September so they could only go in June.*

*(6) Nam: I choose June because he suggested going on summer holiday and there were three options, June, July or September. But July was too hot, September, he had to work. So, June.*

Thu and Nam were able to quickly choose the correct answer and give their explanation. They did not seem to have any trouble understanding the input. It was clear that bottom-up processing was sufficient to answer this item.

### **Item 15**

This item was written for an interview in part 2 of the LCLT. This is an interview with a singer called Nick Parker who plays in a band called Krispy with his sister Mel. The question reads:

The band Krispy started after

- A. Nick began studying at music school.
- B. two other musicians heard Nick and Mel playing**

C. Nick had enough money.

D. Nick and Mel advertised for the band members.

Audio script:

- Your band, Krispy, has two guys and two girls in it. How was it formed?

- Mel and I were playing in a concert at our college and there were two students from music school in the audience. They came to see us after the show and asked if we'd like to form a band with them. We weren't sure at first because we were much younger than them but we agreed to try it out and it was brilliant.

Number of correct answers: 10/10

The cognitive processes involved in answering item 15 are much similar to those for item 1 and 7. When the input was clear and understandable to the listeners, they made their decisions with ease, matching the straightforward information from the input text with the most suitable option. The participants were shown to be in total control and very certain about their answers like An and Nam in their responses below.

*(7) An: I choose B because he mentioned that 2 musicians came to see Nick and Mel perform and invited Nick and Mel to join them. So, the band was formed after that and it was very successful.*

*(8) Nam: Two other people came to watch Nick and Mel play. After that, they wanted to form a band with Nick and Mel.*

### **Item 22**

This item was written for a monologue in part 3 of the LCLT. This was a talk to a group of visitors to the Smithsonian museums in America. The question reads:

How many items are on exhibit in the Smithsonian museums?

A. Several

B. Sixty thousand

**C. Millions and millions**

D. Sixteen million

Audio script:

I hope you've enjoyed your visit so far in Washington DC. Today we're going on a tour of the Smithsonian. The Smithsonian is actually several museums, each with a different focus, situated together on a mall. These museums in total have more than sixty million items on exhibit.

Number of correct answers: 2/10

Unlike all other items that engage the ability to understand clearly articulated details, the participants' performance on item 22 was unexpectedly poor. An and Minh were the only two participants who answered the item correctly while all the other ones opted for the same incorrect answer (option D).

**Data for the correct answers:**

*(9) An: I heard sixteen million but I also heard that there were more than sixteen million. So, I think if I choose sixteen million, it will not be correct. Millions and millions is better because there were more than sixteen million. My answer is C.*

An picked up an important linguistic cue which was the number of exhibits in the museums. However, she misheard the exact number and thought it was sixteen instead of sixty million. It was most likely that An mistook "sixty" for "sixteen" because she wrongly identified the stressed syllable. Fortunately, the phrase "more than" saved her from making the wrong decision. With this clue, An was certain that the answer was C (millions and millions).

*(10) Minh: I choose D. She said sixty million. Million does not mean several so A is not correct. I didn't hear the word thousand so sixty thousand is not correct, either. D is also not correct because she said sixty not sixteen million. My answer is C, millions and millions.*

Unlike An, Minh recognized the correct number, sixty million. Therefore, she was able to fix the answer from the beginning. However, Minh also considered the other options and justified why they were not chosen.

**Data for the incorrect answers:**

(11) *Chi: I'm sure it's sixteen million because she said so. I choose D.*

(12) *Thu: She said more than sixteen million exhibits in these museums.*

(13) *Nga: They mentioned the number of exhibits and it is sixteen million.*

(14) *Trang: It's very clear. She said sixteen million.*

The responses from Chi, Thu, Nga and Trang were typical of those who came up with the incorrect answer. The obvious reason was that they all mistook “sixty” for “sixteen”. It was clear that word recognition was a problem when the word mentioned in the input and the one included in one of the response options have very similar pronunciation and differ mainly in the word stress. Moreover, when there was partial lexical overlap between the key word from the input text and one of the distractors, most participants quickly fell into that trap.

For item 22, the targeted detail was clearly articulated in the input and the participants did not report any problem understanding the information. That being said, given the participants' scores, it was one of the most difficult items of the test. The main reason was that it targeted the ability to recognize the exact number from a stream of speech, especially when this number had tricky pronunciation. Out of the 10 participants, only one of them was able to recognize the right number (sixty) while all the other ones mistook it for sixteen. Item 22 is an example showing that an item that targets low-level cognitive processes can still be very challenging for the listeners when it focuses on one particular tricky language item.

The participants' responses to items 1, 2, 7, 15 and 22 illustrated their ability to identify details of discussion around them when speech was clearly articulated. What stood out from their verbal reports was the dominance of bottom-up processing. Since the input was

clear and understandable to the subjects, they tended to figure out the answers based on the surface meanings of the words they could hear. Most of the items that engaged this subskill were found to be very easy for this group of participants with 10 out of 10 correct answers. The only exception was item 22 in which the participants' ability to recognize a particular number was put to the test. The problem that the subjects had in identifying the stressed syllable and distinguishing between "sixty" and "sixteen" prevented them from coming up with the correct answer. As determined in the test specifications, the ability to understand details of a discussion when they are clearly articulated should be targeted by only 4 items in part 1 of the LCLT. However, this subskill was still spotted in Part 2 and 3 where more sophisticated listening abilities were expected.

#### **5.4.1.2. Understanding implicit details**

This subskill ranked first in terms of coverage with a total number of 163 references. Although it was the main focus for Part 2, the subskill was also engaged by items in Part 1 and 3 of the LCLT (Table 5.8).

**Table 5.8.** Items engaging the ability to understand implicit details

<b>Part</b>	<b>Item</b>
Part 1	Item 3, 5, 8
Part 2	Item 10, 11, 12, 14, 16
Part 3	Item 23, 28, 29, 31, 32, 33, 34, 35

The participants' responses to items 10, 11, 16, 28 and 35 provide the most useful insights into not only their cognitive processes but also the impact of certain text and task-related factors on the test-takers' performance. These items will be discussed in detail in the following section.

**Item 10**

Context: You will hear three people Wendy, Mrs. Turner and Adrian discussing a film they have just seen at the cinema.

What do Wendy and her mum disagree about?

- A. Whether or not Wendy covered her eyes.
- B. Whether or not they will see the film again.
- C. Whether or not the film was P.G rated.
- D. Whether or not the film was frightening.**

Auditory input:

**Wendy:** Well, you always like these mindless special effect movies. I mean, where was the acting?

**Mrs. Turner:** The dinosaurs were good. The last film I saw with dinosaurs, you could see they were made of rubber but these ones look real. It's amazing.

**Adrian:** It's all done with computer graphics. I will say they look much better on the big screen than they did on my mate's pirate video. But at least I could smoke at my mate's house.

**Wendy:** Well, if everyone was smoking in the cinema, you wouldn't be able to see the picture, would you? Anyway, that's one movie I won't be bothering to take out on video.

**Adrian:** You just didn't like it because you were frightened.

**Wendy:** No, I wasn't. What are you talking about?

**Adrian:** I saw you covering your eyes.

**Wendy:** I wasn't covering my eyes. I was laughing.

**Mrs. Turner:** Well, I couldn't look sometimes. I mean it was only a PG film but some of the scenes were, well, pretty...

**Wendy:** You mean hilarious. Well, at least it wasn't too long.



Number of correct answers: 1/10

**Data for the correct answer:**

*(15) Nga: The answer is D, whether or not the film was frightening because I could hear the word “frightened” in the conversation.*

Nga’s correct answer was clearly not based on her understanding of the input. In fact, she was able to recognize only the word “frightened” from the conversation and simply matched it with option D (whether or not the film was frightening) since this option provided partial lexical overlap with what she could hear. Nga’s successful attempt with item 10 had more to do with luck than her actual comprehension of the text and the overall context of the conversation.

**Data for the incorrect answers:**

Option A – whether or not Wendy covered her eyes

*(16) Nam: I think there are two possible answers. They said the film was frightening and they also argued whether she covered her eyes or not. I think the answer is A because they disagreed on whether she covered her eyes.*

*(17) An: I think A and C are possible because they mentioned whether Wendy covered her eyes and P.G rated. I am in favour of A since they talked more about it. They did not talk a lot about P.G rated and it did not sound like they were arguing with each other about it.*

In their responses, Nam and An made no references to the number of speakers and the distinction between their voices. These two participants did not seem to be aware of the fact that more than two interlocutors were involved in the conversation. The focus of their attention was on the link between what they heard and the response options. Nam was in two minds between A and D because she could hear that the film was frightening and “they also argued about whether she covered her eyes or not”. Meanwhile, An was considering A and C because “they mentioned whether Wendy covered her eyes and P.G rated”. In the end, both

Nam and An chose option A since they were convinced by the disagreement on whether Wendy covered her eyes or not. However, this was the argument between Wendy and Adrian which obviously served as a distraction. The inability to distinguish between speakers was clearly a contributing factor in Nam and An's unsuccessful attempt with this item.

Option C – whether or not the film is P.G rated.

*(18) Thu: In this part, they mentioned that the film was about dinosaurs. They did not mention whether Wendy was closing her eyes or not. They did not say anything about whether they will see it again. I think the answer is D, whether the film was frightening. But C was also possible. ... Oh no, it's not this option because this was between the father and the daughter. It's not A, either because the father said that, not the mother. Later, the mother mentioned that the film was P or G rated. About option B, the daughter said the film was ..... I forgot. I think the answer is C.*

When she was half way through her verbal report, Thu suddenly remembered that there were three people in the conversation and in her interpretation, they were father, mother and daughter. She immediately recalled in her memory who said what and to whom. This process seemed to add more pressure on Thu's memory since she had to recall not only the content of the conversation but also the exact detail that this person said to one of the other two interlocutors. It got even more complicated when Thu had to match what she could remember with each of the response options. At one point she seemed to get overwhelmed and admitted "I forgot". Although she was able to pick the correct option from the beginning, Thu was bewildered by the fact that three speakers were involved at the same time in that particular part of the conversation. That eventually led her to the incorrect answer.

*(19) Minh: I remember that I heard Wendy and Adrian talk much more than Wendy and her mum. I heard Mrs. Turner said something like “it’s P.G rated”. Oh yes! I think the answer for item 10 was C because when she said “it’s only P.G rated”, Wendy said “you mean hilarious”. It’s like they didn’t understand each other. I think so. The rest of the conversation, Wendy and Adrian talked to each other more. When Wendy mentioned whether or not the film was frightening, Adrian asked “were you scared”. Wendy said no. I remember that Wendy said she would not buy a video to watch it again, but she did not say that to her mom. Her mom did not disagree about that. The other three options were related to the discussion between Wendy and Adrian.*

Minh started her report with a comment on how much the three speakers interacted with each other in the conversation. This indicates that apart from trying to understand the content, she actually went beyond utterance level to gain an overall impression of how many people were conversing and to what extent they were involved in the conversation. Then, she went on to recall what each of the speakers said and how they responded to each other. Minh was convinced that C was the correct answer because she could hear Mrs. Turner say “it’s only P.G rated” and also because both Mrs. Turner and Wendy were involved in this part of the conversation. She then justified why the other three options were not correct, reasoning that “they were all related to the discussion between Wendy and Adrian”. Minh’s response to item 10 revealed a complicated cognitive process in which she almost simultaneously (a) caught the key words from the input, (b) identified who said what in the conversation, and (c) use both the key words and her memory of the speakers to choose an answer and disconfirm the other three options. Although Minh was able to understand the overall context of the conversation, she still opted for the wrong answer. It seems possible that the input text and the design of item 10 distracted Minh from understanding the intended message. As a result, she was only able to process information at the surface level and choose C because she heard

“P.G rated” in the interaction between Wendy and her mother.

Thu and Minh are the two typical examples of the participants who were confused or distracted by the number of speakers in the input conversation related to item 10. Four other subjects (Chi, Vy, Mai, Van) also reported a similar experience. It was clear from their verbal data that the number of speakers in this conversation contributed to the increased cognitive load on these six participants. While the question focused particularly on the disagreement between Wendy and her mother, the alternatives involved details of both the debate between Wendy and Adrian, and the discussion between Wendy and her mother. Moreover, the fact that Wendy and Adrian dominated that part of the conversation made it even more challenging for the listeners to focus on the interaction between Wendy and her mother. It seems possible that when the participants had to listen to all three speakers at the same time and try to distinguish between them, less attention was paid to other important factors such as the actual content of the conversation, the language use and the speakers’ tone.

With only one correct answer, item 10 stood out as the most difficult item of the LCLT. It is worth noting that this result is consistent with the finding of the quantitative analysis in Chapter 4. So far, this item has been found to be the most challenging item for both the large group of 200 test-takers with varying levels of proficiency and this small group of 10 English majors. The analysis of the verbal data suggests that the number of speakers involved in the conversation was a major factor contributing to the difficulty level of this item. Apart from that, the lexical overlap between the text and the incorrect options was also found to affect item difficulty in this case.

#### **Item 11**

This item was designed for the same conversation and thus shares the same context with item 10 which was discussed above.

What is it they suggested that they do now that the movie has ended?

- A. Go straight home
- B. Go to see a friend
- C. Go for dinner
- D. Go for a drink**

Audio script

**Wendy:** ..... Look! It's only a quarter past ten.

**Mrs. Turner:** Oh, yes, ten minutes till last orders. Why don't we go for a quick one at the King's Head.

**Wendy:** Adrian can't. Don't you remember. He's driving.

Number of correct answers: 3/10

With only three correct answers, this item is difficult one for this group of participants. It should be noted that, all the distractors worked since each of them was picked at least once by the unsuccessful participants.

#### **Data for the correct answer**

**(20) Nam:** *Uhm, I forgot where they wanted to go but they said they had ten minutes to do something. I think with ten minutes, they could only go for a drink. They could not go for dinner or do anything else. So, I choose, go for a drink.*

Nam depended entirely on one single detail which was the phrase "ten minutes". It seemed possible that "ten minutes" was the only phrase Nam could separate from the stream of speech she was listening to. From that detail, she inferred that ten minutes was the amount of time that the three people in the conversation had to do something. She then made another inference which might have come from her own common sense that ten minutes was only enough for a drink. Nam came up with the answer using both bottom-up and top-down processing. The two processes interacted in a way that Nam used her own knowledge and

experience to make an inference about the words she could recognize from the conversation. This inference is actually not unreasonable; however, it does not reflect the right reason why the answer should be D for this item.

**(21) An:** *I think the answer is go for a drink. From the beginning when I heard a name, I did not know it was the name of what but I was thinking drink or dinner. Because it felt like they were going to a restaurant but not sure which restaurant. Then, the daughter disagreed because the father had to drive. So, I inferred that they wanted to go for a drink because if they drank, they could not drive.*

Two particular details helped An come up with the correct answer. First, she heard a name which helped her narrow down the answer to either “go for dinner” or “go for a drink”. Interestingly, An was not able to tell what exactly that name was and no explanation was given about why she thought a particular name was mentioned in the conversation. With her feeling of the overall context, An thought it was the name of a restaurant. However, she was not sure whether they would go there for a drink or for dinner. She kept listening until she heard the key information, “the father had to drive”. This was a turning point at which the answer became clear to her. An could immediately refer to the common sense about drinking and driving, and infer that the three people in the conversation were talking about going for a drink. This decision is the outcome of the inference making process in which An used both her understanding of the surface meaning and her common sense knowledge.

**(22) Minh:** *Because I was still thinking about item 10 so I did not pay attention to item 11. Ah, I did not remember where they were going but I heard some key words ‘don’t you remember that Adrian has to drive’. If he drives then what he should not do is drinking. So, I think the answer is go for a drink.*

Minh seemed lost at first, admitting that she was still thinking about item 10 and not yet ready for item 11. However, immediately after that, she suddenly remembered what she

heard and was able to retrieve from her short-term memory one whole utterance “don’t you remember that Adrian has to drive”. From here, Minh referred to the drinking – driving common sense, made an inference and chose ‘go for a drink’. For Minh, the answer came quickly as soon as she was able to retrieve the right information from her memory. Minh and An both drew on the same kind of world knowledge to make an appropriate inference.

### **Data for the incorrect answers**

Option C – go for dinner

*(23) Chi: I missed most of the information so I can just consider the given alternatives, see which one makes more sense. This is just my guess. Uhm, it was ten to seven, maybe..., not sure whether they wanted to eat or drink. It’s fifty fifty. Uhm, I heard something which I was not quite sure about. There was a name. It could have been the name of a person or of a restaurant. I will choose ‘go for dinner’ because restaurants are usually named after a person. A place to drink might have a more interesting name. So, I choose ‘go for dinner’.*

Chi decided to make a guess based on the given alternatives. This gives hint of a common test-taking strategy. She took advantage of the previewing opportunity to decide which option made more sense. That being said, she did not depend entirely on the written text. Chi still made her guess based on two details she got from the auditory input, the time and the name. Unfortunately, she got the time wrong and had a very vague idea of the name. It was the detail about time that led her to narrowing down the options to “eat” or “drink”. The final decision was made based on her consideration of the name. Indeed, it mainly came from her personal opinion rather than common sense knowledge. Chi intended to make a guess from the beginning but she actually was making inferences using the vague clues she got from the input text. Both bottom-up and top-down processing were observed in her response. However, the influence of top-down processing was stronger since Chi could hardly recognize words from this part of the conversation.

Option B – go to see a friend

*(24) Mai: I heard a name so I thought they would go to see a friend. I was still thinking about question 10 so I didn't pay attention. I heard something but it was vague.*

Mai claimed that she heard a name and associated it with a person. She did not seem able to recognize any particular word and merely guessed.

Option A – go straight home

*(25) Thu: I guess they will go straight home. I think I heard the words come back.*

*(26) Nga: I think it's A. They finish the movie, they will go home.*

*(27) Van: Just my guess. Go straight home. I remember but vaguely. There was something about time. I'm not sure what time it was. I guess so.*

The three participants who picked A were simply making guesses. The information from the text was either misheard or vague to them.

Item 11 is a typical example of items which require the ability to identify details when they are not clearly articulated in a conversation. Among 10 participants, only three of them, An, Minh and Nam, came up with the correct answers. The cognitive processes that these three participants had in common were recalling important words or phrases from the input and making inferences from those clues. When the detail that the item targeted was not mentioned in a straightforward manner, it is apparent that the ability to isolate and recognize known words in a stream of speech did not suffice. Therefore, these three participants had to go a step further by incorporating top-down processes and using their common sense knowledge to infer the implied detail.

If inferencing is typical of the participants who answered item 11 correctly, guessing is the strategy shared by those who failed this item. Among these participants, Chi was the only participant who made her guess based on her very limited understanding of the input



text. All the other ones turned to random guessing as a test-taking strategy since they were not able to comprehend that particular part of the conversation.

### Item 16

This item belongs to the second conversation in Part 2 of the LCLT in which test-takers hear an interview with a singer named Nick Parker who plays in a band called Krispy with his sister Mel.

What does Nick say about life in the band today?

- A. He's pleased to have the chance to travel
- B. It brings them a lot of money
- C. There's no opportunity for them to relax together.
- D. The older members look after him and Mel**

Audio script:

- You've been together a few years now and you're one of the top bands. What's that like?
- Hard work. We travel to concerts all around the world and are never in one place for more than a few nights. The others are like an older brother and sister to me and Mel, which is good. They help us relax on our days off and make sure we eat well. They're strict about practicing, too.

Number of correct answers: 10/10

#### Data for the correct answer

*(28) Chi: I will choose D because he had the chance to travel but he said he was very tired of travelling. I think B is also possible but I'm not quite sure. No, they do not have the opportunity to relax together. I choose D. I am still thinking because I'm not sure whether members mean band members or family members. Anyway, I still choose D for item 16.*

The elimination strategy is central to Chi's approach to answering this item. From her verbalization, there was no evidence that she caught the information relevant to the correct

answer (“the older members look after him and Mel”). Instead, she used her understanding of other details to eliminate the distractors. There was enough information from her short-term memory to confirm that A and C were not the correct answer. Chi thought option B “it brings them a lot of money” was possible but she did not opt for it. This might have come from her understanding of the general context in which Nick talked about the band’s schedule and life rather than money. The facilitating effect of the multiple-choice format was observed in this case. Indeed, Chi missed the particular piece of information related to the correct answer. If it hadn’t been for the given options and the previewing opportunity, she might not have been able to pick the right option.

*(29) Nga: I choose D because he mentioned that the other two members took care of Nick and Mel. At first, I was thinking about A because he did say that they had the chance to travel everywhere for their concerts. However, he did not mention whether he was happy about it. Therefore, I eliminated that option and I choose D. There was no information about the other two options, B and C.*

Nga seemed very confident about her decision to choose option D. She reported hearing that the other two members took care of Nick and Mel. However, it was not clear what particular clue(s) from the conversation led her to that interpretation. A possible explanation might be that Nga was actually able to catch the details related to what the other two members did to help Nick and Mel. Nevertheless, by the time she started reporting, she already forgot the exact words or details. Instead, she could only recall her general interpretation that the other two members took care of Nick and Mel. In order to further justify her answer, Nga also used the elimination strategy to confirm that the other three options were not appropriate.

*(30) Minh: I choose D, the older members look after him and Mel. Because he said that the older members were like brothers and sisters to them. I heard the word ‘good’ but I was not*

*sure what it was about. He said that he travels a lot, that they 'hard work' but he didn't say he was pleased so I think it is not A. Bring a lot of money, I did not hear anything about it. I think it is D because he said the other members were like brothers and sisters and they take care or something. I think take care means look after, they are synonyms. The older members make sure that they eat well and blah blah blah. I think all of that means look after.*

Minh confidently chose D because she understood enough to be certain about this option. She was able to retrieve from her memory key phrases such as “like brothers and sisters”, “make sure they eat well”, which were then interpreted as “take care”. It should be pointed out that the phrase “take care” was not mentioned in the conversation. Minh might have thought she heard the phrase but it was, indeed, her interpretation of the input. After that interpretation, she went a step further by confirming that “take care” and “look after” are synonymous, thus justifying her choice. There was also evidence of the elimination strategy since Minh did explain why she refuted option A and B. The strategy, however, only helped to reassure herself rather than playing an important role in her decision.

Item 16 is an easy item for this group of participants, which is evidenced by the fact that all of them got the correct answer. However, the thinking processes behind it varied from one person to another. There was evidence of elimination strategy from the verbal data. While Chi depended on it to figure out the correct answer, Nga and Minh employed the strategy mainly for reassurance. It also became apparent that bottom-up processing had an important role to play since the identification of key words from the input enabled the test-takers to give the appropriate interpretation. This was possibly implied in Nga's report but salient in Minh's verbalization. Even in Chi's case (16) when the elimination strategy played a key role, she still eliminated the distractors based on the linguistic cues she heard from the input conversation.

**Item 28**

Item 28 belongs to Part 3 of the LCLT and it was written for a lecture on Hawaiian volcanoes. It should be noted that the information necessary for this item was mentioned much earlier in the input text, even before the detail needed for item 27. This mismatch between the order of information in the input text and the order of the corresponding test items is highlighted in Table 5.9.

**Table 5.9.** Auditory input for item 26, 27 and 28

<b>Auditory input</b>	<b>Items as printed on the test paper.</b>
<p>In yesterday's class we discussed the volcanos located in the area known as the ring of fire, an area which basically encircles the Pacific and includes the United States' Mount St Helens as well as Japan's Mount Fuji and Argentina's Aconcagua, the highest mountain in the western hemisphere.</p>	<p>Item 26. What was the topic of yesterday's lecture?</p> <ul style="list-style-type: none"> <li>A. Different types of volcanoes</li> <li>B. What caused the Ring of Fire</li> <li>C. The volcanoes of the Ring of Fire</li> <li>D. Hawaiian volcanoes</li> </ul>
<p>Most of the world's approximately five hundred active volcanos are located along the ring of fire and the eruptions that take place there are among the most violent in the world.</p>	<p>Item 27. What is the topic of today's lecture?</p> <ul style="list-style-type: none"> <li>A. Moana Loa</li> <li>B. The Ring of Fire</li> <li>C. The characteristics of volcanoes in the Ring of Fire</li> <li>A. The volcanoes of Hawaii</li> </ul>
<p>Today, we are going to discuss the volcanos of Hawaii which are quite different from the volcanos in the ring of fire. Hawaiian volcanos are not located along the ring of fire and are therefore not caused by the movement of the earth plates against each other. Instead, Hawaii is located in the middle of the ring of fire, above a massive plate rather than where two plates meet.</p>	<p>Item 28. Where are most of the world's active volcanoes located?</p> <ul style="list-style-type: none"> <li>A. Within the Ring of Fire</li> <li>B. Along the Ring of Fire</li> <li>C. In Hawaii</li> <li>D. In the United States</li> </ul>

As seen from Table 5.9 the very first sentence of the lecture was directly relevant to item 26. Right after this, the information about the location of most of the world's active volcanoes which was needed to answer item 28 was introduced. Then, the lecturer went on to reveal the content of "today's lecture" which was necessary for item 27. This mismatch, which was illustrated by the two arrows in this table, was found to be a source of difficulty for the participants. Among the 10 participants, only 4 were able to choose the correct answer.

### **Data for the correct answer**

The four successful participants were the ones who recognized the mismatch. In their verbal reports, Chi, Vy, Mai and Minh all emphasized that the information necessary for item 28 had already been mentioned before. Fortunately, they were still able to successfully recall the relevant linguistic cues and thus figure out the correct answer. For example:

*(31) Chi: Uhm, the information for this item was mentioned earlier. If I don't misremember, it is now information for item 29. The order of information was funny because as I remember from the beginning, for item 26, he mentioned the topic of yesterday's lecture which was volcanoes along the Ring of Fire. He also said that the most active volcanoes of the world are located along the ring of fire. At that time, there was no information for item 27 so I did not pay much attention. But now I think it's relevant to item 28 so as I remember, I choose option B, along the ring of fire.*

*(32) Minh: I choose B, along the ring of fire. I could hear that. But I'm not sure whether the order of information was the same as the order of items. When I take other tests, like IELTS for example, the information in the conversation will match the order of questions. I have never been in a situation where I have to jump back and forth to answer question. Previously, I heard "the most active" and then I heard "along the ring of fire" so I chose that option.*

*Because I read the questions before listening so it helps me recall the information for this item.*

In the quoted verbal reports, it was apparent that Minh and Chi underwent very similar thinking processes. First, they both recognized the mismatch between the order in which the information was presented in the lecture and the order of the corresponding test items. Then, they used their memory to trace back the information relevant to item 28. Afterwards, they were able to identify key phrases such as “the most active volcanoes” and “along the ring of fire”. The last step was to match these phrases with the option that contains perfect lexical overlap. Responses related to the use of memory such as “I remember”, “if I don’t misremember” and “it helps me recall” were spotted quite frequently in Chi and Minh’s verbal report. They indicated the importance of memory in this particular case. Besides, Minh also revealed that reading the questions carefully before listening helped her recall the right information. This is an example to show that in a tricky situation like this, working memory, the ability to identify key information, and the opportunity to preview the test item all came together to help the listener figure out the correct answer.

#### **Data for the incorrect answer**

For item 28, six unsuccessful participants ended up choosing the same answer, option A (within the ring of fire). A closer inspection of their think-aloud protocols revealed that they went through very similar cognitive processes. An important thing to note about these participants is that they were not aware of the mismatch between the input and the order of item 26, 27 and 28. They followed the stream of speech without questioning and waited for the linguistic cues that matched one of the response options. In this case, all the six participants reported recognizing the phrase “in the middle of the ring of fire” and immediately matching it with option A (“within the ring of fire”) given the obvious association between “in the middle” and “within”. However, the text-to-recording strategy

did not work for them since their focus was only on a lexical phrase while item 28 required comprehension beyond that level as well as careful consideration of the question itself. Some examples of incorrect responses are given below:

*(33) Trang: I think he said the volcanoes in Hawaii are in the middle of the ring of fire. It it's in the middle, then within is the most suitable option. I choose A, within the ring of fire.*

*(34) Van: I don't choose "along". I choose "within". Within means in the middle right? Because he said "in the middle".*

*(35) Trang: It's not "along the ring of fire". It's "in the middle of the ring of fire". So, I choose within the ring of fire.*

The comparison between those who answered item 28 correctly and those who did not showed that the ability to recognize the mismatch between the input text and the item order played a decisive role. The participants who were able to recognize the mismatch came up with the correct answer. In this case, the role of memory must be acknowledged. Meanwhile, those who were not able to notice the mismatch, all ended up with the same incorrect answer. The mismatch between the order in which key information was introduced in the input text and the order of test items did have an impact on the participants' performance. Given the fact that the recording was played once only, it came as no surprise that item 28 was a relatively difficult item for the participants.

### **Item 35**

This is the last item of the LCLT which belongs to a talk about a students' concert. In this talk, Lucy, a music teacher revealed her experience with a new development in one of the music colleges.

In the concert, Tim Travis, a professional musician

- A. Uses only one kind of instrument.
- B. Conducts his own piece of music.

**C. Has to play a strange instrument.**

D. Plays drums.

Audio script:

The orchestra's percussionist, Tim Travis, who was more accustomed to playing the drums seemed fairly good-humoured about it all. He admitted that he had played some peculiar things in his time. He's rattled chains, played a sharp carving knife and also played on funny whistles. In this piece, he has to use a fairly heavy hammer and keeps beating one particular place.

Number of correct answers: 3/10

**Data for the correct answer**

*(36) Nga: The answer for question 35 is C because she mentioned names of different instruments and those instruments were used by Tim Travis in a strange and fun way. Therefore, I choose C.*

*(37) Mai: For question 35, I choose C because I heard her mention names of some instruments that he played and they all sounded strange so I thought the answer was C. She mentioned drums, too but I think there was more evidence for C.*

*(38) Trang: I think the answer is C because from the beginning 'plays drums' was mentioned but it didn't have anything to do with him. Then, I heard a list of other instruments and he admitted that he has to play those instruments so I thought the answer is C.*

Nga, Mai and Trang were the only three participants who answered item 35 correctly. Two of them confirmed that 'drums' was mentioned; however, they were not trapped by that word. All the three participants admitted that they heard names of instruments, other than 'drums'. Interestingly enough, they were not able to name any of those instruments, which shows that bottom-up processing did not work for them in this situation. They seemed to develop a sense for what they heard to compensate for their inadequate linguistic knowledge.



A possible explanation is that from their understanding of the general context and the clues from the written MCQ, these participants knew that musical instruments would be mentioned. Therefore, they could all claim that they heard “a list of other instruments” without being able to name them. If that was the case, the use of top-down processes had an important role to play.

After reporting that they heard names of instruments, Nga and Mai explained why the instruments were “strange”. It is important to note that none of them seemed to catch the word “peculiar” from the input text. Rather, they used other clues to help them decide. Nga thought the instruments were played in a “strange and fun” way without giving any reason. This could have been just her general impression. Meanwhile, Mai admitted that all the instruments sounded strange to her so she decided that they were strange instruments. Trang, on the other hand, could catch the phrase “has to” and thus chose option C (“has to play strange instrument”), which best matched what she heard.

#### **Data for incorrect answer**

Option D – play drums

*(39) Thu: From the beginning, I could hear ‘plays drums’, then I heard that he played many other instruments so A is not correct. The option ‘has to play strange instrument, .... it was probably mentioned. He found it kind of interesting and he had fun doing something. But I heard ‘plays drums’ first so I will choose D.*

*(40) Vy: I heard that he played drums but there was so much information after that so there's got to be something more. Uhm, C is possible but still I choose D. He definitely played drums.*

*(41) An: I choose D, “plays drums”. All the information that followed was about other things but they did not seem to have anything to do with the concert. That might be additional information about him. From the beginning, it was mentioned that he played drums.*

These three participants, Thu, Vy and An all chose ‘plays drums’ for item 35. It is clear from their explanation that the only phrase they could hear clearly was ‘play drums’. Interestingly enough, they all acknowledged that there was more information than just ‘play drums’, and that some other options were also possible. However, due to the lack of comprehension of other details given in the text, they chose to stick with option D, which provided complete lexical overlap with the phrase they recognized. In this case, the participants depended solely on bottom-up processing to answer the item, holding on to the only linguistic clue they had. The thinking processes that Thu, Vy and An underwent were typical of all the participants who chose D and ending up failing this item.

In summary, item 35 is a difficult item for which the participants were clearly divided in their responses. Only three of them came up with the correct answer, option C, while all the others opted for D. It became apparent from the analysis of verbal data that all those who depended merely on bottom-up processing, ended up with an incorrect answer. By sticking with the phrase “play drums”, they fell into the trap of lexical overlap. This was probably the reason why this item was difficult since as Buck (2001) once noted, lexical overlap between the text and incorrect options is the best predictor of difficult items. In order to identify the particular detail that this item requires, successful participants had to go beyond word level, using their knowledge of the context, background knowledge and even personal feeling to compensate for their inadequate linguistic knowledge. Top-down processing had a clear advantage in this case.

### **Summary**

Thus far, the in-depth analyses of participants’ verbal reports on item 10, 11, 16, 28 and 35 have cast light on the cognitive processes behind the ability to recognize details of a conversation when they are not clearly articulated. The think-aloud protocols clearly indicated that linguistic cues and linguistic knowledge are essential for the identification of

the required details. Correct responses to item 11, 16 and 28 are good examples to illustrate this point. In particular, excerpts (21), (22), (28), (29), (30), (31) and (32) show how the subjects came up with the correct answers from the words they recognized. In the meantime, failure in linguistic processing induced groundless guessing which later led to incorrect responses (excerpt (23) – (27)). For this group of English majors, bottom-up processing was important in a sense that their success in it greatly contributed to their correct answers.

That being said, top-down processing also had a crucial role to play. The use of relevant non-linguistic knowledge was evident in the responses for item 11 and 35. In item 11, the participants' responses revealed that the interaction between linguistic cues and background knowledge was necessary for successful identification of the required detail (excerpt (21) and (22)). While in item 35, the understanding of the context was used by the successful participants to compensate for their inadequate linguistic knowledge (excerpt (36) – (38)). For this item, successful and unsuccessful subjects differed in the ability to go beyond word level and use top-down processing to understand the targeted detail.

The think-aloud protocols also revealed considerable impacts of the multiple-choice format. First, this format allowed much uninformed guessing which was often a result of inadequate linguistic processing. However, none of the attempts in guessing was successful. Examples are found in excerpt (24) - (27). Second, the amount of lexical overlap between the input and the incorrect options clearly distracted the listeners. This is evident in the incorrect responses for item 10 (excerpt (16) – (19)), item 28 (excerpt (33) – (35)), and item 35 (excerpt (39) – (41)).

Although these five items were seen as targeting the same listening subskill, they varied greatly in terms of difficulty levels. In terms of the participants' scores, item 10 was very difficult with only 1 correct answer. Item 11 and 35 were difficult with only 3 correct answers for each. Item 28 was relatively difficult with 4 correct responses. Item 16 can be

considered very easy with a total number of 10 correct answers. The varying levels of difficulty could be partly attributed to the text and task-related factors identified in the verbal data. For item 10, the fact that three speakers were involved in the input conversation was clearly responsible for the increased cognitive load on the listeners. Meanwhile, item 28 challenged the test-takers by not matching the flow of ideas in the input with the order of test items. For items 11 and 35, it seemed likely that the content of the input was not very familiar to the listeners. That might have been a contributing factor to the participants' poor performance on these items.

#### **5.4.1.3. Making simple inferences about the context**

The ability to make simple inferences about the context was referred to 32 times in the verbal data. As determined in the test specifications, this subskill is targeted by Part 1 of the LCLT. Three items in this part (item 3, 5, 8) were found to successfully activate this ability. The participants' responses to these items are provided below.

##### **Item 3**

What will they do tomorrow afternoon?

- A. Shopping
- B. Swimming in the sea**
- C. Boating
- D. Biking

Audio script:

- This is a beautiful beach. Shall we come again tomorrow?
- Don't we have to go shopping?
- We can do that in the morning and come here in the afternoon.
- Okay, and let's bring Joe and Linda with us.

Number of correct answers: 10/10

**Data for the correct answer**

*(42) Van: Ah, I chose swimming in the sea because the girl said the beach was ... The answer is not straightforward. That makes me a little confused because they mentioned 'shopping', too but from the beginning, she said the beach was beautiful and she wanted to come back tomorrow. But the guy said how about shopping? Then she said they could go shopping in the morning and come back here in the afternoon. And the question is about tomorrow afternoon. So, swimming in the sea.*

*Researcher: Did they mention the word swimming?*

*Van: No, but swimming in the sea is the best match.*

Van stated the answer immediately but delayed her explanation for a few seconds. By admitting that the answer was not straightforward, she seemed to understand that she needed to make an inference. Van's response continued with her recalling most of the details in the conversation. It should be noted that she was not simply quoting them one by one but trying to connect them in a coherent and organized manner. This was observed in the way she used connectives such as "because", "but", "from the beginning" and "then". Van must have been trying to understand the context better so that she was able to make an appropriate inference. Once she came up with a clear storyline, Van repeated the time reference in the question in order to confirm the answer she picked from the beginning (swimming in the sea). Although she admitted not hearing the word "swimming", Van went on to emphasize that her answer was "the best match". It seems likely that she associated the word "beach" that was recognized from the text with the word "sea" in the response option, thus, choosing the option without even hearing the word "swimming".

In her response to item 3, Van was able to identify the required detail by making a simple inference about the context. The inference was made based on her good understanding of the input. It was simple in a way that it did not involve high-level cognitive processing or

much background knowledge. Van only needed to use her basic knowledge of coherence and cohesion to figure out the relationship between the ideas in the input text. The association between the word “beach” from the input and “sea” in the correct answer was also not difficult to recognize.

*(43) An: I choose swimming in the sea because from the beginning, one of them said that she wanted to come back to the beach tomorrow. Then the other one suggested going shopping in the morning and going swimming, going back to that beach in the afternoon. The question is about tomorrow afternoon so they would go swimming in the sea then.*

**Researcher:** *But did they mention the word swimming?*

*An: No, they didn't mention the word swimming. They only mentioned the beach. There are actually two possible answers, boating and swimming in the sea so I'm a little confused here. However, when they go to the beach, I'm more convinced that they will swim. Boating is not that common and if the answer was boating, I believe there would be some detail to emphasize the difference between boating and swimming.*

At first, An's response to item 3 reflected similar cognitive processes to Van's; however, more was revealed later about why she decided to stick with option B, 'swimming in the sea'. When asked whether the word 'swimming' was mentioned in the conversation, An confirmed that it was not. Then, she started to consider 'boating' as well; however, this option was quickly refuted by her common sense knowledge. She went beyond the text itself to assert that if the answer had been boating, there should have been further details to differentiate between boating and swimming.

Van's and An's verbalizations are two typical examples of the participants' responses to item 3. Although the input did not seem difficult to understand, this item was a little tricky because the required detail could not be identified immediately. In addition, the correct response option included “swimming”, an action which was not mentioned in the input text.

Therefore, the participants had to go a little beyond the linguistic cues to make a simple inference about the context.

### Item 5

Where are the glasses?

- A. On the table
- B. On the chair
- C. By the phone
- D. Under the table**

Audio script

- Are my glasses by the telephone?
- No, where did you leave them?
- Well ... I had them when I phoned Alice.
- Here they are. On the floor.

Number of correct answers: 9/10

#### Data for the correct answer

*(44) Mai: Uhm, the woman asked whether her glasses were next to the telephone. The man answered no and asked where she last put her glasses. I could hear 'on the floor', so I choose 'under the table'.*

*(45) Thu: She asked whether her glasses were near the telephone. He said no and asked where she left them. She said, he said "here on the floor". So, I think, it's definitely not "by the phone", "on the chair" or "on the table". If it's on the floor, it is definitely under the table.*

Mai and Thu had no problem understanding this conversation. They picked up the relevant, salient linguistic cues and inferred the correct answer. Both of them were able to repeat the exact location of the glasses as mentioned by one of the interlocutors. It should be

noted that the exact words that they heard (“on the floor”) did not overlap with any of the options. Therefore, they had to use their common sense to make a simple inference of the context. Thu explained this in more detail by pointing out that “on the floor” could only mean “under the table” in this case.

#### **Data for the incorrect answer**

*(46) Nam: I could hear "on the floor" but ... ah, it is by the phone because she just made a phone call, she put the phone down and she put her glasses down as well. So, I think it has to be "by the phone".*

Like Mai and Thu, Nam was also able to recognize the phrase “on the floor” but instead of sticking with it, she got distracted by the word “phone”. Her focus of attention shifted to option C since the word “phone” was also found in this option. To justify why she chose C, Nam inferred that the woman put the phone down and put her glasses next to the phone. It seems clear that the lexical overlap between the input and option C led Nam to ignore the most important linguistic cue and allowed her belief to override any other possibilities. Therefore, her attempt to make an inference about the context was not successful.

#### **Item 8**

How is Patty going to travel?

- A. By train
- B. By car**
- C. By plane
- D. By bus

Audio script:

- I'm going to go to Vienna on Saturday.
- How long does it take to fly there, Patty?



- Oh. I'm going to drive.
- Oh, yes. You get ill on planes, don't you?
- Yes, and trains.

Number of correct answers: 9/10

#### **Data for the correct answer**

*(47) Chi: Uhm, the guy asked how long it would take to fly but the girl said she would drive. Although she did not say exactly how she would travel but among the four options, car is the only thing that she can drive. So, I choose B, by car.*

*(48) Trang: At first, the guy asked how long it would take her to fly there but the girl said she was going to drive. So, I choose B, by car. After that, she also explained that she got sick on plane or bus.*

In order to answer item 8, both Chi and Trang relied on the word 'drive', the most important and relevant clue. This word led them directly to the correct answer which was 'by car'. The underlying inference was that among the given alternatives, "car" was the only word that could serve as an appropriate object to the verb "drive". Both participants made a simple inference about the context based on the same linguistic cue and drew on the same collocational knowledge. This is also the common cognitive process experienced by seven other participants.

#### **Data for the incorrect answer**

*(49) Van: Why are there two possible answers? At first, I heard the word 'plane', then she said 'and train'. I'm confused. I usually trust my first thought so I choose C, by plane.*

As for Van, the cognitive process that she underwent for item 8 had a lot to do with word matching rather than inferencing. It seems that all of her attention was drawn to identifying the exact word that matched one of the given alternatives. Therefore, she got confused when she heard both 'plane' and 'train'. Her decision to choose "plane" was not

elaborated. Rather, Van simply explained that she trusted the first thought that came to her mind.

### **Summary**

The participants' responses to items 3, 5 and 8 have demonstrated how these items engaged the ability to make simple inferences about the context. In terms of design, they were written in a way that no lexical overlap was found between the input texts and the correct answers. That way, the listeners could not depend entirely on simple word matching but had to make inferences to answer the items. The participants who answered these items correctly reported cognitive processes which included identifying the key words or phrases, making inferences from these clues, and matching the inferences with the correct options. Their inferences involved basic knowledge of coherence, cohesion and synonyms as in excerpt (42) and (43), common sense knowledge of the logical arrangement of objects as in excerpt (44) and (45), and basic collocational knowledge as in (47) and (48). The think-aloud protocols for item 3, 5, and 8 also indicated that low-level processing which involved straightforward lexical matching was not effective (excerpt (46) and (49)).

#### **5.4.1.4. Making inferences about the speaker's attitude and opinion**

There was extensive evidence of participants engaging in a process of making inferences about the speakers' attitudes and opinions. Such evidence was elicited from the participants' verbal reports on items 9, 10, 18, 19, 20, 31. In this section, items 9, 19, 20 and 31 will be subjected to in-depth analysis.

#### **Item 9**

This item belongs to the first conversation in part 2 of the LCLT. In this conversation, test-takers hear three people Wendy, Mrs. Turner and Adrian discussing a film they have just seen at the cinema.

What did Wendy's mum think of the film?

- A. She loved the film.
- B. It was not her favourite film.**
- C. She absolutely hated it.
- D. It made her feel sick.

Audio script:

**Wendy:** What a load of old rubbish! You hated it, too, didn't you, mum!

**Mrs. Turner:** Well, I have seen better films. I mean it wasn't as good as Aliens or that Werewolf film, you know the one?

**Adrian:** Oh, come on, some bits were brilliant!

Number of correct answers: 9/10

#### **Data for the correct answer**

*(50) An: I heard that Wendy did not like this film, she really hated it. She asked her mom but her mom said the film was ..... there were better films. She did not have anything bad to say about the film. She didn't hate it or felt sick of it. It was just not her favourite film.*

An could easily recall the general content of the conversation and retrieve from her memory an important utterance "there were better films". Although she was not able to repeat the exact words but An seemed to have the sensitivity to the speakers' implication. From what Mrs. Turner said, An inferred that Mrs. Turner thought the film was not bad. Therefore, she decided to eliminate option C and D, and came to the final decision that "it was just not her favourite film".

*(51) Minh: Wendy thought it was rubbish but her mom said 'it wasn't as good as the Aliens film, blah, blah, blah. Adrian said the film was brilliant or something. I didn't hear exactly what he said but it was not important anyway. Here, I only need to pay attention to what Wendy's mom said. Well, she didn't hate it. Option C, absolutely hate it, it's more like Wendy*

*when she called the film rubbish. Option D, it makes her feel sick, was the same. Love is definitely not true so I think B is the best. Not her favourite means she didn't hate it but she didn't love it, either. It was not as good as the films that she had seen before.*

Minh started her report by recalling what each of the speakers said about the movie. The most relevant and important linguistic cue that she decided to focus her attention on was Mrs. Turner's comment, "it was not as good as the Aliens film". From this utterance, she began to make inferences about Mrs. Turner's opinion and compared her interpretations with each of the response options. That way, she was able to eliminate option C, D, A and picked B (it was not her favourite film) as the correct answer. However, Minh did not stop there but went on to explain that "not her favourite means she didn't hate it but she didn't love it either". Finally, she referred back to what Mrs. Turner said about the film to reassure herself of the inference she made. For this item, Minh came up with the correct answer by making inferences based on a relevant linguistic cue and the elimination strategy. It should be pointed out that Minh's interpretation of Mrs. Turner's opinion was mainly shaped by the consideration of the language use rather than the prosodic features of the utterance.

*(52) Trang: I'm still thinking about item 9. I think the important information has come but I missed it. I guess her mom did not hate this film but because she used a tag question to ask .... But it was not her favourite film. I'm thinking B or D but I can't make up my mind now. (Trang waited until the end of the conversation).*

**Researcher:** *So, what's your answer for item 9?*

**Trang:** *I choose B. I think the film was not so bad that it made her feel sick. It was only about dinosaurs.*

**Researcher:** *So, you choose B based on your general understanding after you have listened to the whole conversation or because you could hear some information related to option B?*

**Trang:** *I wait until the end of the conversation and then got back to item 9 and chose option*

*B because I think it made sense. Also, because I was considering other options but they were not as reasonable. I used the elimination strategy here as well.*

When the conversation was half way through, Trang was still unable to decide on the answer for item 9. I paused the conversation so that she had the opportunity to report what she was thinking before there was too much information to remember. At this point, Trang had only a vague idea of the conversation, admitting that she might have missed the key information. She had a feeling that this was not Mrs. Turner's favourite film; however, Trang could not convince herself of this option. She then narrowed down her choices to only option B and D but it was not clear why she thought so. Since she found it hard to decide, Trang delayed the answer until the end of the conversation. Once the conversation ended, Trang confirmed her final choice which was B (It was not her favourite film) and rejected option D (It made her feel sick). She admitted that she waited until the end of the conversation to be sure that option B was more reasonable than the other options. There was no evidence from her think-aloud protocol that Trang could hear what Mrs. Turner said and make any inference from that. Instead, she relied on her overall impression of the conversation and the elimination strategy. Trang, eventually, came up with the correct answer but her thinking process did not involve the ability to make inferences about Mrs. Turner's opinion based on the consideration of tone or language use.

#### **Data for the incorrect answer**

Option C – She absolutely hated it.

*(53) Nam: The answer is C, she said she had seen better films so I think she did not like the movie.*

Nam chose the wrong answer for item 9 which was quite surprising since she was actually able to pick a relevant clue from the conversation (Mrs. Turner said he had seen better films). From that clue, Nam inferred that Mrs. Turner did not like the movie and

quickly matched her interpretation with C (She absolutely hated it). Since Nam gave a very quick and short response, it was not likely that she gave her answer enough thought. She attempted to make an inference based on Mrs. Turner's comment; however, she seemed to capture only the surface meaning rather than the subtle implication of the utterance. Although Nam was only able to interpret that Mrs. Turner did not like the movie, she could not tell from her language and tone of voice that Mrs. Turner did not hate the movie, either. It seems possible that her limited pragmatic knowledge led her to choose the incorrect option. Another possibility might be that Nam did not take the time to read the response options carefully.

There was much evidence from the think-aloud protocols showing that item 9 engaged the ability to make inferences about the speaker's attitude and opinion. This subskill was present in the cognitive processes reported by most participants. An and Minh are two typical examples of the subjects who were able to identify relevant linguistic cues and make appropriate inferences from the language used by the speakers. It was also found that these participants used the elimination strategy to aid their decision-making process. From the verbal data, there was also evidence of a different approach in which the overall impression of the conversation was taken into account to compensate for missing information. This approach was used by only one participant (Trang) when she found herself unable to make the decision and thus delayed the answer until the end of the conversation (excerpt (52)). This approach, though different from the one taken by all other participants, still led Trang to the correct answer. The only incorrect response from Nam could possibly be attributed to the listener's inadequate pragmatic knowledge.

### **Item 19**

This item belongs to the third conversation in Part 2 of the LCLT. In this conversation, test-takers hear an interview with an engineer called Roger Moffet, whose working life has changed dramatically over the last ten years.

Roger regards his early days in business as

- A. **demanding**
- B. frustrating
- C. irrelevant
- D. boring

Audio script:

**Interviewer:** And what kind of success did you have in the early days?

**Roger:** You could say it was a bit like taking a roller coaster ride and wondering when you're going to come flying off at breakneck speed. Everything was a challenge. Finance, production, marketing.

Number of correct answers: 5/10

**Data for the correct answer**

*(54) Chi: In item 19 he compared his early days to a roller coaster, going up and down. So, it's definitely not boring and irrelevant. It might be frustrating and demanding. Demanding makes more sense because frustrating is ... . If it is a roller coaster, it can't be ... . Business is kind of up and down. There are many demands and he has to change himself to adapt and meet those demands. So, my answer is A, demanding.*

So as to come up with the correct answer for item 19, Chi underwent a thinking process in which more than one inference were made. From the beginning, she caught the key word 'roller coaster' and quickly inferred that Roger's business went up and down in the early days. With this inference, she eliminated option C (irrelevant) and D (boring). Left with two remaining options, Chi had to make another inference to decide on her final answer. She thought "demanding" made more sense; however, uncertainty was evident in the way she paused and left her sentences unfinished. Finally, she reasoned that when Roger's business was up and down, there were many demands that he had to change himself to adapt to. This

inference made practical sense; however, it was not directly linked to any particular linguistic cue from the input. Chi might have heard some relevant key words from the conversation but was not able to retrieve them from her memory. It seems more likely that she used her own common sense, logic and background knowledge to decide on the final answer.

*(55) Mai: I am a little confused. I heard the word roller coaster. It goes up and down so it's kind of frustrating. But after that, I heard the word "challenge" and "finance" and some other things so I think demanding will be a better choice.*

Mai's explanation reveals that her focus of attention seemed to be mainly at the lexical level. When Mai recognized the word "roller coaster", she immediately associated it with the up and down movement and then the feeling of frustration. Nonetheless, she completely changed her mind when she heard the words "challenge" and "finance". From here, Mai decided to choose 'demanding' as her final answer for item 19. It was clear that she made an inference, associating "challenge" and "finance" with "demanding" in this particular context.

*(56) An: I'm not sure whether the information for item 19 has passed or is it item 20 now? But I think this part of the conversation is quite difficult. They were talking non-stop and it was not as clear as the previous conversations. So, it is more difficult to keep track. I think he made a comparison here. When he talked about his early days, he compared them to a roller coaster. He felt like he was on a roller coaster. That means demanding. Like the job has many demands. So, he had to ... I don't know how to explain that but I will choose demanding. He compared it to a roller coaster. It feels like demanding.*

In her response, An admitted that this item was difficult since the conversation was fast and not as clear the previous ones. The only piece of information she could recall was that Roger compared his early days to a roller coaster ride. From here, she started to make an inference about how Roger felt when he found himself on a roller coaster. She immediately



thought it was demanding but struggled to explain why. Bottom-up processing was observed here since An clearly depended on the linguistic cue she got from the input. However, in order to link that information with the correct answer, she must have used her common sense knowledge as well as her understanding of the general context because “roller coaster” and “demanding” are not usually associated with each other. It could be argued that top-down processing also had a crucial role to play in the way An tackled this item.

### **Data for the incorrect answers**

Option B – frustrating

*(57) Vy: Uhm ... Roger compared his early days to a roller coaster ride. Then, he was frustrated. He was spinning. He seemed to have no idea what direction to take. So, I think I will choose something frustrating.*

What Vy was drawing her attention to was the fact that Roger compared his early days to a roller coaster ride. From that only cue, she interpreted that Roger’s business back then was frustrating. Instead of associating the roller coaster ride with the up and down movement, Vy linked it to a spinning movement. She then related that to Roger’s situation, explaining that “he seemed to have no idea what direction to take”. Vy’s inference was made based on a particular linguistic cue and at the same time shaped by her own background knowledge and her personal opinion. This is an interesting example showing that inference making is a personal process in which different listeners can make different inferences about the same clue. Similar to Chi, Mai and An, Vy was able to recognize the word “roller coaster” and made inferences based on this word. However, her background knowledge and personal opinion led her to a different inference which unfortunately matched the incorrect answer.

*(58) Van: For item 19, I eliminate boring. I don’t think the answer is “boring”, “irrelevant” is not possible, either.*

**Researcher:** *Why do you eliminate irrelevant and boring?*

**Van:** *Well, it is just not logical. That's how I feel about it.*

**Researcher:** *So, what's your answer for item 19 if it's not 'boring', 'not 'irrelevant'*

**Van:** *I like B more.*

**Researcher:** *Simply because you like it? Is there any information you can remember from the conversation? Anything that gives you the reason to choose 'frustrating'?*

**Van:** *Well, they talked really fast. All the sounds are linked together. If I miss just one piece of information, I will miss all of it.*

Van's response is another example of test-taking strategy being used to compensate for the subject's lack of comprehension. Having no clue of the conversation, Van became entirely reliant on the way she felt about each alternative. Specifically, she eliminated 'irrelevant' and 'boring' because they did not sound logical to her. Then, she decided to choose 'frustrating' over 'demanding' simply because she liked it more. Since no evidence of her listening ability was found, Van's report is another case in which construct-irrelevant variance was present.

Option C – irrelevant

**(59) Thu:** *In item 19, he described his success in the early days. He talked a lot but I couldn't understand anything. I guess his success in the early days was too little to mention. So, it will not be frustrating, demanding or boring because success does not mean any of those. I think the answer is C. I missed all the information. They talked too fast.*

Thu started by admitting that she missed all the information related to Roger's answer. She focused her attention on the word "success" which was not a relevant and useful clue for this item. From here, she made an ungrounded guess that "his success in the early days was too little to mention" and thus eliminated "frustrating", "demanding" and "boring".

Thu finally chose “irrelevant” as her final answer simply because she thought the other three options were not related to “success”. No reasonable explanation was given for the particular option that she picked. In her response to item 19, Thu depended largely on an irrelevant linguistic cue and the written text. Her lack of comprehension was compensated for by her test-taking strategy which was to consider the response options using her own knowledge and logic. It can be argued that construct-irrelevant variance was present in Thu’s report since there was almost no evidence that item 19 engaged her listening ability.

Evidence from the think-aloud protocols reveal that item 19 engaged the ability to make inferences about the speaker’s attitude and opinion. For those who came up with the correct answer, their verbal reports all show a lack of comprehension of the input. From excerpt (54), (55) and (56), it could be noticed that the subjects were only able to recall separate linguistic cues rather than retelling the content of the conversation. They tended to depend on the same clue which was the fact that the speaker compared his early days in business to a roller coaster. In their attempts to make inferences about the speaker’s attitude, these subjects activated their non-linguistic knowledge through the linguistic cues they got from the input. The word “roller coaster”, in particular, played a central role since all the successful subjects made inferences based on this word. It was also clear that in their inference making process, only language use was taken into consideration. There was no evidence that the subject paid attention to the speaker’s tone.

For those who came up with the incorrect answers, two tendencies presented themselves regarding how they tackled the question. First, a relevant linguistic cue (“roller coaster”) was recognized; however, the subject had different interpretations which subsequently led to the incorrect answer (excerpt (57)). Second, the participants were not able to understand the input and depended entirely on the written text to figure out the answer (excerpt (58) and (59)).

**Item 20**

This item belongs to the same conversation as item 19.

What does Roger feel is the greatest benefit of running his own business?

- A. He arranges his free time as he pleases.
- B. He gets on better with other people.
- C. He is free of an environment he disliked.**
- D. He has more leisure time than before.

Audio script:

**Interviewer:** Do you have any regrets about the way things have gone, about the way your life has taken a different turn?

**Roger:** To be honest, not at all. I feel that I've escaped being a slave to a regular income. From commuting, from having to justify my actions to everyone, from having to attend the office party, from having to book my holidays in advance. Actually, I don't have any holidays at all at the moment, come to think of it. I'm too busy. But best of all, I've nothing to do with office politics.

Number of correct answers: 6/10

**Data for the correct answer**

*(60) Nam: I choose C. He said that he could escape from the environment that he didn't like. Like...like... he has to please other people or something. So, I don't think he gets on better with other people because he doesn't like to please them. There is no way he can get on better with them. Moreover, he is too busy and does not have time for himself. So, the greatest benefit cannot be 'arrange his free time as he pleases'.*

Nam started by justifying her choice and then explained why the other options were not appropriate. She chose C, claiming that Roger could escape from the environment that he

didn't like. Although Nam was unable to provide details of Roger's working environment, she could at least give an overall description to justify her claim ("he has to please other people or something"). This overall description can be seen as an attempt to paraphrase the input. It seems possible that Nam understood the input but could not retrieve the relevant linguistic cues from her memory. Therefore, she chose to paraphrase the given information based on her overall comprehension of the input. This attempt was successful since the paraphrased information captured an important aspect in Roger's previous working environment. From the impression that Roger did not like an environment in which he had to please other people, Nam went on to infer that Roger did not get on better with other people, thus eliminating option B. From the input, Nam also knew that Roger was too busy and had no time for himself. With that in mind, she eliminated option A (he arranges his free time as he pleases). Although option D was not mentioned, it seems likely that Nam did not pick D for the same reason.

*(61) Minh: I choose C "free of the environment he dislikes". The woman asked him whether he had any regret but he said no. He said he could escape from being a slave, being judged. Like he could not really be himself. Then he said he did not have holiday. So, it can't be D because it's about leisure. He does not have free time as well. He didn't say much about people. He said that he was very busy. The important point is he emphasized twice that he had nothing to do with office life any more. So, the answer is C.*

Similar to Nam, Minh could catch the word "escape" and was able to describe the environment that Roger had worked in before. Her description included both a word taken directly from the input ("slave") and paraphrased information ("being judged", "he could not really be himself"). The paraphrased information was evidence of Minh's comprehension since it remained relatively close to what Roger said. After justifying her decision to choose C, Nam started recalling other details in the conversation to eliminate options which were not

suitable. Hearing that Roger did not have a holiday, Minh associated that with ‘leisure’ and ‘free time’, thus eliminating option A and D. She was also able to recognize that Roger did not talk much about people, which made option B impossible. Finally, she reconfirmed her decision to choose option C. At this point, she did not merely recall information from memory but went beyond that level by synthesizing the information she could hear. Minh must have put together what Roger said at the beginning and at the end of his answer to finally assert that “he emphasized twice that he had nothing to do with office life any more”. In general, the cognitive processes that Minh underwent while answering item 20 was well explained. She made inferences based on her understanding of the input and used the elimination strategy to back her decision.

*(62) Trang: I think the answer is C although I did not think about it at all from the beginning. Then he said “best of all” and something related to politics so I infer that it was an environment he didn’t like. So, I chose C.*

*(63) Mai: In his last sentence, I could hear something related to office so I think it has something to do with the environment he didn’t like. “Best of all” means “greatest benefit”. He mentioned free time and leisure time but I think the answer is C.*

It appeared that Trang and Mai were not able to get the main ideas from Roger’s answer to the interview question. All they could hear was some fragments from the input which were “best of all”, “office” and “politics”. However, Trang and Minh did not seem able to connect these linguistic cues. Instead, they made inferences which were most likely based on their personal logic and experience. In particular, Trang heard “best of all” and “politics” and inferred that it was the environment Roger did not like. Mai made the same inference, reporting that she heard something related to “office”. No further description of the working environment was given as in excerpt (60) and (61). It seems clear that Trang and Mai did not come up with the correct answer because they understood the input well. Rather,

their success could be attributed to the ability to make inferences which were not well-grounded but still appropriate.

*(64) Chi: I can't hear anything related to option B. Option A and D are not possible since he said "I'm too busy" and something else. So, I choose C.*

Chi used the elimination strategy to figure out the correct answer. She was able to understand enough to be certain that A and D were not possible. At the same time, she also recognized that the information about whether Roger got on better with other people (option B) was not mentioned in the input. Therefore, Chi decided to choose C without understanding the details relevant to this option. It should be pointed out that the multiple-choice format contributed greatly to Chi's success. Without the given options, she might not have been able to come up with the correct answer.

*(65) Thu: I think the answer is not A or D. So there are only B and C left. I'm not quite sure whether he liked it (the working environment) or not. His answer was long. I think I will choose C. It's my guess.*

Thu's report also showed signs of an elimination strategy. However, unlike Chi, Thu made ungrounded guesses and there was no evidence of her comprehension of the input. The elimination strategy was used as a sole approach to answering the test items. It seems possible that Thu depended largely on the written text rather than the auditory input.

#### **Data for the incorrect answers**

Option B – get on better with other people

*(66) Van: I will eliminate D because he said he was busy. So, he does not have leisure time. That means he does not have free time so option A, "arrange his free time as he pleases" is also not possible. I choose B because I could hear the word "communication".*

Van chose to use the elimination strategy first. Since she could hear that Roger was very busy, she eliminated option A and D. Then she claimed that she heard the word

“communication” and associated the word with option B (he gets on better with other people). However, the word that was actually mentioned in the conversation was “commuting”, not “communication”. Mishearing “commuting” lead Van to the incorrect answer.

*(67) Vy: A, C and D refer to one point while B refers to another. I think he said that the greatest benefit was B. I choose B.*

Vy was not able to show her understanding of the input. Her explanation was merely her own guesses which were most probably based on the written text.

Option A – He arranges his free time as he pleases

*(68) Nga: I remember vaguely that he mentioned something related to arranging his free time. So I choose A.*

There was no evidence of comprehension in Nga’s verbal report. She depended on a clue that she remembered vaguely and matched it with option A without further explanation.

*(69) An: I choose A, ‘he arranges his free time as he pleases’. Because I could hear that at one point he said he didn’t need to ..... It has something to do with leisure time. So, I think it’s either A or D. Then he said that he didn’t have to book holiday in advance, or ask for permission, or join the company party when he didn’t want to. He also said that he didn’t have as much free time as before because he was very busy. So, I will not choose D. I choose A because although he is busy, he can still arrange his free time as he wants because he is the boss.*

Since much of what she heard was related to leisure time, An narrowed down the options to only A (He arranges his time as he pleases) and D (He has more leisure time than before). Then, she made inferences so that they matched one of these options. Because An was certain that “he (Roger) was very busy”, she inferred that “he didn’t have as much free



time as before” and thus eliminated option D. At this point, An went on to make another inference to convince herself that A was the correct answer. The claim that “although he is busy, he can still arrange his free time as he wants because he is the boss” is not unreasonable. However, it was made based on the listener’s knowledge and experience rather than on the information given in the input.

With 6 out of 10 correct answers, item 20 was rather difficult for this group of participants. The approaches to answering the item varied considerably from person to person. Among the six successful participants, only two appeared to be able to identify relevant linguistic cues and make appropriate inferences from them (excerpt (60), (61)). The other four, on the other hand, could only recognize one or two linguistic cues and had to turn to their non-linguistic knowledge, the elimination strategy or guessing to come up with the correct answer (excerpt (62), (63), (64), (65)).

The two main reasons for incorrect answers were mishearing the input (except (66)) and lack of comprehension (excerpt (67), (68)). An interesting case was observed in excerpt (69) where the listener had a relatively good understanding of the text but failed to catch the most relevant information. This led her to making an inference which made sense but did not reflect the speaker’s opinion.

### **Item 31**

This item belongs to the last talk in Part 3 of the LCLT. Test-takers listen to Lucy, a music teacher, talking about new development in one of the music colleges.

Why did Lucy spend the whole week to listen to the students’ concert?

- A. She had free time.
- B. She had to.
- C. She was attracted to it.**
- D. The students wanted her to.

Audio script:

I spent last week sitting in on a rehearsal for a concert which will be given by final year students at World Seven Colleges of music in a couple of weeks' time. Now, I wouldn't normally spend one whole week doing this but after the first day's rehearsal, I was fascinated by what was going on and wanted to hear more.

Number of correct answers: 7/10

**Data for the correct answer**

*(70) Chi: Uhm, the information is quite clear here. Lucy said she spent the whole week listening to the students' concert. Normally, she would not do that but she said that after a day and a half, she was fascinated and wanted to hear more. So, I choose C, she was attracted to it.*

*(71) Vy: She was attracted to the students' concert because I think the concert last the whole week and after her first time going to it, she was appealed and attracted. So, she decided to go on the other days as well.*

*(72) Nga: The answer is C. She said that although she was not free for a whole week, she was fascinated after the first day listening to that concert. And I think fascinated is synonymous to attracted in option C. So, I choose C.*

The verbal reports from Chi, Vy and Nga show the typical cognitive processes that the seven successful participants experienced while answering item 31. They all chose C and provided similar explanation for that choice. It is worth noticing that none of the participants depended too much on individual words. Instead, they confidently recalled the main ideas of the conversation and make inferences based on their understanding of the context. Since their understanding of the input was adequate, compensatory strategies did not seem necessary in this case.

### Data for the incorrect answers

Option B – she had to

*(73) Nam: I choose “she had to”. I have a feeling that at the beginning, she did not want to be there but after one week, she realized that it was good so she continued to watch. I’m not quite sure. But I think the answer is “she had to”.*

Nam seemed to get the general idea that the speaker (Lucy) was not very interested at first but she then became fascinated by the concert. She thought that Lucy only realized the concert was interesting after one week listening to it. However, as Lucy said in the input text, she was fascinated after the first day and then decided to spend one whole week listening to the students’ concert. It was likely that Nam misheard the time reference or could not remember the exact detail. Therefore, she ended up choosing the incorrect option.

*(74) Mai: I think the answer is A. She mentioned the word “loud” or “bucket”, so I think it was not pleasant to listen to. So, I choose A “she had to”.*

Mai quoted the word “loud” and “bucket” which came later in this talk and was not relevant to item 31. Judging from her verbal report, Nam seemed unable to comprehend the information related to this item. Therefore, she kept waiting until she was able to catch the words that she found useful. The words “loud” and “bucket” came after the information needed for item 31. They were associated with the students’ concert and treated as key words. From such association, Mai inferred that the concert was not pleasant to listen to. Therefore, the reason why Lucy spent the whole week listening to the concert was because “she had to”. In Mai’s response, there was evidence of her making inferences about the speaker’s attitude based on linguistic cues from the input text. Nevertheless, her lack of input comprehension led to the misidentification of key words. This, in turn, resulted in Mai’s unsuccessful attempt in answering item 31.

The analysis of the participants' verbal reports on item 9, 19, 20 and 31 has provided evidence that the LCLT did engage the ability to make inferences about the speakers' attitudes and opinions. The wording of these items indicates the test developers' intention in measuring this ability. In particular, the stems of the items explicitly targeted the speakers' attitudes and opinions with questions or statements such as "What does Wendy's mom **think** of the movie?", "What does Roger **feel** is the greatest benefit of running his own business?", and "Roger **regards** his early days in business as \_\_\_\_". The response options for each of these items also included vocabulary related to feelings such as "she absolutely hated it" (item 9), "demanding", "frustrating" (item 19), and expressions indicating attitudes such as "he is free of an environment he disliked" (item 20) or "she was attracted to it" (item 31). None of the response options for item 9, 19, 20 and 31 perfectly matched the words used in the input text. This shows the test designers' effort in creating a condition for inferences to be made about the speakers' attitudes and opinions.

From the verbal data, there was extensive evidence of participants engaging in the process of making inferences about the speakers' attitudes and opinions. This inference making process was aided by the participants' pragmatic knowledge, common sense knowledge and background knowledge. Since such knowledge varied from person to person, it was not surprising that in certain situations, the listeners made different inferences based on the same linguistic cue(s), thus choosing different answers for a particular item. Examples can be found in responses to item 9 (excerpt (50), (51) and (53)), or responses to item 19 (excerpt (54) and (57)).

The think-aloud protocols also revealed that in the inference making process, the participants paid much attention to the speakers' language use but not their tone of voice. Indeed, there was no evidence of inferences made based on the prosodic features of the utterances. In the participants' responses, the rationale supplied for an inference or an answer

was very frequently 'I heard the word' or 'I heard that...'. The tone of the speakers was never referred to as one of the factors that aid the inference-making and decision-making process.

Although the items analysed in this section target the ability to make inferences which can be seen as involving high-level cognitive processes, the importance of lower-level or bottom-up processing was indeed undeniable. As was clear from the analysis, all the participants who came up with the correct answers for item 9, 19, 20 and 31 reported cognitive processes in which they successfully identified relevant linguistic cues and made inferences based on these cues. Evidence can be found in excerpt (50)-(52) for item 9, (54)-(56) for item 19, (60)-(63) for item 20, and (70)-(72) for item 31. At the same time, failure in answering these items was mainly associated with not being able to understand the input (excerpt (58) and (59)), not being able to recognize relevant information (excerpt (68), (69), (74)), and mishearing the input (excerpt (66)). It was clear that for these four items which targeted a supposedly higher-level listening subskill, lower-level processing was still involved to a large extent. Although top-down processes were salient in the responses to these items, bottom-up processes were showed to be of equal (or perhaps greater) importance.

Another issue emerging from the analysis was the varying difficulty levels among the items. While item 9 was easy for this group of test takers with nine correct responses, item 19 was quite difficult with only five correct answers. Item 20 and 31 had moderate difficulty levels with six and seven correct answers respectively. The reason for the varying levels of item difficulty was not yet clear but it might have had something to do with the clarity of the input. For item 9, 20 and 31, no complaints were reported about how fast and unclear the conversations were. Meanwhile, in the responses to item 19, two of the participants who came up with the incorrect answers appeared to totally get lost. The explanation they gave was "they talked too fast" (excerpt (59)) and "They talked really fast. All the sounds are linked together" (excerpt (58)).

The finding that participants made use of test-taking strategies to make up for their lack of input comprehension caused no surprise. The two main strategies they turned to were elimination strategy and guessing which were made possible by the written text and the multiple-choice format (excerpt (58), (59) and (65)).

#### **5.4.1.5. Understanding discourse function and the purpose of the speaker**

As stated in the test specifications, this subskill ranked high in the difficulty hierarchy (level 5) and should be represented by 2 test items in the second part of the LCLT. The verbal data, however, provided evidence of only one item in Part 3 of the test that targeted the ability to recognize discourse function and the purposes of the speaker. Specifically, the responses to item 21 revealed evidence of cognitive processes in which the participants turned to their discourse knowledge to figure out the correct answer. Details of these processes are provided in the following analysis of the participants' responses to this particular item.

#### **Item 21**

Context: You will hear a talk to a group of visitors who are visiting the Smithsonian museums in America.

Who is probably giving this talk?

- A. An Indian
- B. A professor
- C. An artist
- D. A tour guide**

Audio script:

I hope you've enjoyed your visit so far in Washington DC. Today we're going on a tour of the Smithsonian. The Smithsonian is actually several museums, each with a different focus,

situated together on a mall. These museums in total have more than 60 million items on exhibit.

Number of correct answers: 9/10

**Data for the correct answer**

*(75) Trang: For item 21, I think the speaker was a tour guide who was giving the visitors a tour. She said "I hope you enjoy...". The other three options were not possible. Professors give lectures. There was no information about Indian or artist.*

*(76) Thu: I choose D, tour guide because she was introducing. Like, today we will go on a tour. She must be a tour guide. A professor will not speak like that. Indian and artist were not mentioned.*

*(77) Van: I can eliminate option A because she can't be an Indian. She doesn't have an Indian accent. I can infer from the context. That was the way someone was introducing and guiding other people. A tour guide. Something like she was taking the visitors around and introduce things at the same time. Now we will go on a tour, something like that. She did not say she was a tour guide and I'm taking you on a tour but I understood that.*

Trang, Thu and Van all came up with the correct answer and they underwent similar cognitive processes. First, they were able to catch the key phrases which were "we will go on a tour", "I hope you enjoy...". From these cues, they used their knowledge of discourse to make inferences about who the speaker was. All the three participants recognized that the phrases that they heard was typical of a tour guide's language. It was evident from the way the speaker used the word "tour", from the way she checked whether the visitors enjoyed the trip or introduce the upcoming activity. From such language use, Trang, Thu and Van shared the same impression that the speaker was introducing and guiding the visitors. Although they all seemed confident in their decision to choose 'tour guide' as the final answer, these participants still further confirmed their decision by considering the other three alternatives. It

should be noted that their evaluation of the other options was also based on their sensitivity to discourse types and accent. Trang and Thu eliminated option B (professor) because in their opinion, “a professor will not speak like that” and “a professor gives lecture”. These responses indicated, though indirectly, that Trang and Thu understood the type of discourse generally associated with a professor’s lecture. Van added another perspective when she eliminated option A (an Indian) because in her opinion, the speaker did not have an Indian accent.

Some other participants reported similar cognitive processes; however, their responses also showed the influence of the written text and the context provided prior to listening.

*(78) An: Before listening to the talk, I read that this was a “talk to a group of visitors”, so I had the feeling that tour guide was more suitable than all the other options. It can also be an artist but as I actually listened to the talk, the language use was typical of a tour guide. Like “today, we’re going on a tour”. So, I think, tour guide was the best option. Better than Indian, professor or artist.*

*(79) Minh: Right from the beginning I already thought it would be a tour guide because this was a “talk to a group of visitors”. No, she did not have an Indian accent. I don’t think she was a professor. Not an artist, either. She did not talk about the museum only. She said they visited Washington DC as well, so I think she must be a tour guide.*

Both An and Minh admitted that they had the answer in their minds even before listening. The amount of context given for this talk, though limited, was still enough help these participants predict the correct answer. For An and Minh, the introduction that this was “a talk to a group of visitors” sparked the idea that the speaker would be a tour guide. However, this still remained a hypothesis until they actually listened to the text. Although their decision-making process was aided by the context given for the talk, their listening abilities still played a major role in helping them confirm the hypothesis they had from the



beginning. While the use of discourse knowledge was evident in An's report, general understanding of the context and language use led Minh to the final decision on the correct answer.

**Data for the incorrect answer:**

Option B – A professor

**(80) Nam:** *I think this is a professor because I think a tour guide must sound more excited.*

*This is a professor giving a lecture. It might be a field trip.*

This was the only incorrect response to item 21 in which the participant made an inference based solely on the speaker's tone of voice. It might have come from her own experience that a tour guide usually sounded energetic and excited. Therefore, when the speaker's voice did not seem to match her expectation, Nam discarded the possibility that the speaker was a tour guide. She thought the speaker was a professor who was giving a lecture, adding that the talk might have been given in a field trip. However, no reference to the input text was given to back that claim. It was quite interesting that Nam paid attention to the speaker's tone of voice, which was very rare in the verbal data. Unfortunately, that cue was not very useful for item 21.

In general, item 21 required the participants to go beyond processing at word level and make inferences using their knowledge of discourse. As evidenced by the verbal report, most participants were able to activate the relevant discourse knowledge to come up with the correct answer. There was also evidence of participants using background knowledge and the context provided for the talk to aid their decision-making process. Although item 21 was supposed to engage a high-level listening subskill (level 5), it was found to be easy for this group of participants. This result might be explained by the fact that the input text was relatively clear and easy to understand. Another possible reason is that the type of discourse

(tour-guide talk) targeted by this item is quite common and therefore its typical features were not difficult to recognize.

#### **5.4.2. Factors affecting test-takers' performance on LCLT items**

In this section, interest is attached to the evidence from the think-aloud protocols which showed that the participants' responses to certain LCLT items were considerably influenced by factors related to the input texts or the design of these items. A total number of six factors were identified from the participants' verbal reports. They are:

- multiple-choice format;
- number of speakers;
- order of items;
- speech rate;
- lexical overlap;
- quality of the recording.

The following discussion provides evidence for each of these factors and elucidates the extent to which they influenced the participants' cognitive processes.

##### **5.4.2.1. The multiple-choice format**

The think-aloud protocols suggested that the multiple-choice format encouraged bottom-up processes in which the participants focused their attention mainly on lexical matches rather than on wider meaning. There was extensive evidence of the test-takers using the classic key word strategy to locate relevant information in the auditory input. The most common response from all the participants to justify an answer was "I heard the word" and what came after that was a particular word or phrase written in one of the options. Some reported that they were mainly waiting to catch the word(s) that matched the written text rather than listening to get the ideas.

The lexical matching strategy was mainly noticed in the participants' responses to the LCLT items that targeted the ability to identify implicit information and the ability to make inferences. For certain items in Part 2 and Part 3 of the LCLT, most participants' responses indicated that the lack of comprehension led them to either partly or entirely depend on the written text. In a number of cases, the participants treated the answer options as the guide for their listening process. Once they were able to identify a lexical item that matched one of the response options, they decided to choose that option without understanding the overall meaning of the input. Responses to item 10 and item 35 are typical examples (see excerpt (18)-(19); (39)-(41)). It seems paradoxical that in response to items which required global comprehension of text, many participants tended to process the input only at the local, lexical level. The evidence from the verbal data suggested that the response options might have distracted the participants from attempting to go beyond word level and figure out the broader relationship between details. However, that undesirable impact of the multiple-choice format was not noticed in the responses to items which targeted explicit details and literal meaning of the input. For the LCLT items that targeted the ability to identify explicit details in short and simple conversations, the participants showed very little dependence on the answer options. Most of them were able to correctly recall the content of the input text and then chose the answers that matched their comprehension.

Another issue observed in the verbal data was the use of test-wise behaviours which were directly linked to the multiple-choice format. Random guessing was, not surprisingly, one of the strategies that the participants reported using to compensate for their lack of comprehension. It was noticed in some responses to items 11, 17, 18, 19 and 20 of the LCLT. The other test-specific strategy emerging from the think-aloud protocols was deductive reasoning which was based solely on the written information. The evidence of test-specific behaviours was found in 18 responses which are summarized in Table 5.10.

**Table 5.10.** Evidence of test-taking strategies in the verbal data

Item	Strategy used	No. of references
11	Guessing	3
17	Deductive reasoning based solely on the written information	4
	Guessing	4
18	Guessing	2
19	Deductive reasoning based solely on the written information	2
20	Deductive reasoning based solely on the written information	3

What stands out in Table 5.10 is the dominant use of test-taking strategies in response to item 17. In particular, eight out of ten participants reported using either deductive reasoning or guessing to tackle the item. Their responses will be discussed in the analysis of item 17 as follows.

### **Item 17**

Context: You will hear an interview with an engineer called Roger Moffet, whose working life has changed dramatically over the last ten years.

The interview says that Roger is the kind of person who

- A. Is reluctant to try something different.
- B. Enjoys entertaining others.**
- C. Does not want to spend his money.
- D. Is happy to reveal the trick of his trade.

Auditory input:

It seems only fitting that former construction engineer, Roger Muffet should have used his redundancy money to change directions and break into Hollywood, creating special effects for film and television. For by his own flamboyant admission, he's no conventional engineer

but a born performer who loves an audience. Do you remember a certain car commercial in which a car was driven down the side of a skyscraper? The building façade and windows were built by Roger's own company for a daring stunt whose trade secret he will not divulge. He also constructed sections of a bridge for the film *Mary Riley* which starred Julia Roberts and John Malkovich.

Number of correct answers: 2/10

**Data for the correct answer:**

*(81) Mai: I heard something related to special effects so I think it has something to do with "entertaining others". So, I choose B. I also heard the word "engineer" but I didn't hear anything related to the other three options. And it was quite fast.*

Although Mai did not seem to get the general idea of the input, she was able to recognize an important linguistic cue which later led her to the correct answer. She associated the phrase "special effects" with option B (enjoy entertaining others), which was quite reasonable.

*(82) Minh: It's too fast, I can't hear anything. She just talked non-stop. I don't think he's reluctant to try something different. The question is Roger **IS** the kind of person... . Maybe, he **WAS** reluctant before but he is not any more. Because if he is reluctant to try something different, he can't have a successful business. Does not want to spend his money? No, if he does not want to spend his money, he wouldn't do business. Enjoy entertaining others? Maybe, because he works in the film industry. Is happy to reveal the trick of his trade? I don't think so because they didn't mention any trick in business. So, I choose B, enjoy entertaining others.*

Minh made use of several test-taking strategies to answer item 17. First, she decided to delay the answer until the end of the conversation since she admitted getting lost right from the beginning. The excerpt quoted above was the response she gave after she listened to the

entire interview. Minh only got back to item 17 when she was able to understand the general context and use it to answer this challenging item. This strategy worked for her since with her understanding of the overall context, Minh was able to eliminate the inappropriate options and decide on the correct one.

Mai and Minh, the only two participants who answered item 17 correctly, reported very different thinking-process. While Mai's success was attributed to bottom-up processing, Minh depended largely on top-down processing to figure out the answer. However, they both reported having troubles with the fast speed of delivery.

**Data for the incorrect answers:**

Option A - is reluctant to try something different

*(83) Thu: This third conversation is really difficult to listen because it's really fast, all the sounds are linked together. It feels like everything was all mixed up. I can't hear anything. I have to make a guess. I choose A. If his life dramatically changes, he had to try something different.*

*(84) Nam: I missed it anyway. I think he is reluctant to try something new*

*(85) Nga: I can't hear anything. It's too fast. I have to use my intuition. I choose A (reluctant to try something new). It's just my random guess.*

*(86) Van: I give up. She just talked without pausing. I can't understand. I have a feeling, just a feeling. The answer is A.*

When they failed to understand the input, Thu, Nam, Nga and Van all turned to guessing as a compensatory strategy. While Thu made her guess based on the context given prior to listening, Nam and Nga simply made random guesses.

*(87) An: I missed all the information but I'll try. I think I will eliminate option C, does not want to spend his money because in my opinion if he's a successful person, he can't be stingy or narrow-minded. The option "enjoys entertaining others", I'm not quite sure what it*

*means. I'll eliminate it, too because I think it has nothing to do with his job, an engineer. I'm thinking A or D. But in option A (is reluctant to try something different), I don't know the word "reluctant", not sure whether it means something positive or negative. Uhm, D is also possible but I don't think anybody will be happy to reveal his trick. So, I'll choose A.*

An admitted missing all the necessary information for item 17 so she intentionally used her reading skill as a compensatory strategy. An considered all the alternatives and referred to the context provided to eliminate the options that she thought were not suitable. For example, she learnt from the written introduction that Roger was an engineer, so option B (enjoys entertaining others) seemed irrelevant to his job. Once she was able to narrow down the possible answers to A and D, An made use of her personal logic to eliminate option D (is happy to reveal the trick of his trade). Although she was still uncertain about option A (is reluctant to try something different), she ended up choosing it since it was the only alternative left. An's verbal report indicated that several factors came into play at the same time to make up for her lack of comprehension. In particular, her reading skill, the context provided and also the multiple-choice format all had a role to play in her decision-making process.

Option D - is happy to reveal the trick of his trade

**(88) Chi:** *The answer is not B or C. I think so. If he is reluctant to try something different, it doesn't seem right. If his working life has changed dramatically over the last ten years, he can't be reluctant to try something different. This is an interview, so he should be happy to reveal the trick of his trade.*

**(89) Vy:** *It should be something positive. I think D, he's happy to reveal the trick of his trade. That's probably the reason why she wanted to interview him. Just what I feel.*

In order to compensate for their inadequate comprehension, Chi and Vy turned to the written text and the given context to figure out the answer. They also brought their own feeling and opinion to the decision-making process.

From the verbal data, it also became apparent that the fast speech rate coupled with the lack of pauses caused much trouble for the listeners. When the input was so difficult to understand, it came as no surprise that the participants had to turn to test-taking strategies to figure out the answer. Some of them reported depending entirely on the context provided and the written text to tackle this item (excerpt (82), (87), (88), (89)). At the same time, some others turned to random guessing as their last resort (excerpt (83)-(86)).

Judging from the way it was written and with reference to the relevant input, item 17 was expected to engage the ability to identify a certain detail when it was not clearly articulated. However, in reality, most of the responses to this item clearly indicated the use of test-taking strategies which were made possible by the multiple-choice format. In the responses to item 11, 18, 19, and 20, test-taking strategies were also found but with much lower frequency. As for deductive reasoning, evidence was found in excerpt (58) and (59) for item 19, excerpt (64), (65) and (67) for item 20. Guessing was identified in responses to item 11 (excerpt (25)-(27)). This strategy was also found in two verbal reports for item 18 as follows.

### **Item 18**

How did Roger feel initially about what happened ten years ago?

- A. angry
- B. depressed**
- C. resigned
- D. disinterested



Audio script:

Well, about ten years ago, I had a heart by-pass operation and about the same time, I was made redundant. I was feeling pretty low at the time so I decided that the only thing to do was to take my working life into my own hands and set up my own business.

#### **Data for random guessing**

*(90) Vy: I think he was ... just my guess because I missed all the information. I think he felt depressed.*

*(91) Nga: I was still paying attention to item 17 so I missed the information related to item 18. However, I guess it was depressed.*

These two participants were candid about the way they came up with the answer for item 18. They simply guessed because they missed all the relevant information. Their guesses were correct, which was lucky for them but at the same time revealed nothing about their listening ability.

Among the 18 responses in which the participants attempted to use test-taking strategies to answer the test questions, five of them were correct (excerpt (64), (65), (82), (90) and (91)). Apparently, the multiple-choice format introduced construct-irrelevant variance to the LCLT since it allowed some listeners to arrive at the correct answers without understanding the input text.

#### **5.4.2.2. Number of speakers**

The number of speakers in the input conversation for item 10 was one of the factors that made it the most difficult item of the LCLT. The involvement of three speakers with two female voices and 1 male voice was found to increase the cognitive load associated with this question. As the listeners were trying to identify the particular detail required by the item, they also needed to distinguish between the voices and decide which information was exchanged between which speakers. Previous studies have found that the more voices there

are on a sound file and the more overlap there is between them, the more difficult it becomes for the listeners to discern who is saying what (Green, 2017, p.18). Evidence in the verbal data showed that most of the participants were not able to tell exactly who was talking to whom and about what. In their verbal reports, some simply referred to the speakers as “they” rather than their exact names or roles in the conversation (excerpt (16)-(18)). Memory was also an issue when the subjects had to remember not only the content of the conversation but also the speakers who were involved in the discussion of one particular detail (excerpt (18)).

#### **5.4.2.3. Order of items**

This factor was found to affect the participants’ performance on item 28. In the input text, the information necessary for item 28 was mentioned even before the information relevant to item 27. This requires not only the listening ability but also a good memory that allows the listeners to refer back and retrieve the right piece of information. As observed from the verbal data, the responses to item 28 were clearly polarizing. Only the 4 participants who noticed the mismatch and still had the relevant detail in their memory came up with the correct answer ((excerpt (31), (32)). Meanwhile, all the other ones who were not aware of the mismatch ended up with the incorrect answers.

For item 28, even the successful participants found themselves to be confused and somewhat annoyed by the fact that the order of items did not match the flow of the input text. Such a mismatch disrupted their metacognitive processes (Chi, 2011) and added unnecessary cognitive load on their attempts to construct meaning (Kintsch, 1998). My hypothesis was that the test developers might have intentionally put item 27 and 28 in an unexpected order with the view to discriminating between good and not-so-good listeners. However, in this case, I would argue that this arrangement was unfair for both the successful and unsuccessful participants. Those who were able to notice the strange order had to experience confusion and more pressure on their memory. For unsuccessful participants, it was not their fault to not

notice the mismatch between the item order and the information flow. In a listening test that allows the test-takers to listen to the recording once only, anyone should expect the item order to match the flow of information. The participants' responses to item 28 suggests that scrambling the sequence of items placed unnecessary cognitive burden on the test-takers and failed to test their true listening ability. This technique should not be used to manipulate item difficulty. Elliott and Wilson (2013) cautioned that there needs to be sound justification for presenting items out of their chronological order within the text.

#### **5.4.2.4. Speech rate**

Throughout the LCLT, the speech rate seemed reasonable for the participants. Item 17 was the only item that the participants found extremely challenging because of the fast speech rate. Most participants started their report with a comment on the fact that the speakers "talked to fast", "talked non-stop" and "talked without pausing" (excerpt (82)-(86)). For that reason, they got lost and failed to make sense of the input. They eventually gave up and turned to test-taking strategies to answer this item.

There might be two reasons why the input for item 17 was so difficult for the participants. First, the lack of pauses was definitely a contributing factor. In the part of the recording which was relevant to item 17, the speaker did not pause very often and each time she did, the pause was too short for the listeners to process the input. This lack of pauses, as evidenced in most participants' verbal reports, made it difficult for them to identify word boundaries and keep up with the flow of speech. Second, the conversation related to item 17 was delivered at a faster rate than the input for the previous items. It should be noted that Part 2 of the LCLT included 3 separate conversations and four items were written for each conversation as follows:

- Conversation 1 (item 9-12)
- Conversation 2 (item 13-16)

- Conversation 3 (item 17-20)

The sudden increase in speech rate was only noticed by the participants when the third conversation began. Most of the participants started their report by an exclamation of how fast the speech rate was while such comments were not recorded for the previous conversations. This indicated that the participants might have expected the third conversation in Part 2 to be delivered at a comparable speed with the other two and thus, were totally unprepared for the increase in speech rate. This unexpected change, most probably, caused the participants to lose confidence as soon as the conversation started.

Evidence from the think-aloud protocols clearly showed that speech rate had a detrimental effect on comprehension. Nonetheless, since speech rate itself is a fairly crude measure and might not be the sole source of the participants' problems. The consideration of speech rate should be extended to the frequency of pauses (Goldman-Eisler, 1961) and the consistency of speed throughout one certain part of a test.

#### **5.4.2.5. Lexical overlap**

In the LCLT, complete or partial lexical overlap between the input text and the response options were observed in the items that engaged the ability to identify details of a talk or conversation. Specifically, in most of the items that involved the ability to understand clearly articulated details, lexical overlap was found between the input and the correct answers (item 1, 2, 7, 15). The think-aloud protocols revealed that most of the participants answered these items with ease by matching the words they heard with the options containing those words (excerpt (1)-(8)). Apart from the clarity and the simple language use of the passages, the lexical overlap between the auditory input and the correct answers apparently made the items very easy for the participants.

On the other hand, in the items that engaged the ability to identify details which were not clearly articulated, complete and partial lexical overlap was mainly seen between the

input and the distractors (item 10, 28, 35). Consequently, it diverted the participants' attention away from the implied details and tricked the listeners into choosing the incorrect answers (excerpt (16)-(19), (33)-(35), (39)-(41)).

A previous study by (Freedle & Kostin, 1996) indicated that lexical overlap between the correct option and the text is the best predictor of easy items while the overlap between the incorrect options and the text is the best predictor of difficult items. That holds true in this study since the difficulty levels of certain items were clearly influenced by whether lexical overlap was found in the correct answers or the distractors.

#### **5.4.2.6. Quality of the recording**

The inconsistency in the quality of the LCLT recordings represented a source of construct-irrelevant variance which was found to have a noticeable impact on the participants' performance. As reported by all the participants, the quality of the last sound file in the LCLT was noticeably worse than the files used previously in the test. They also acknowledged that the poor sound quality made the last talk difficult to understand. The reports from Trang, Thu, and Minh are typical examples.

*(92) Trang: She talks really fast and the sound quality is not as good as the previous ones. I can't hear much. I can't hear anything for item 31.*

*(93) Thu: I can't hear very clearly. The sounds were a bit distorted.*

*(94) Minh: Oh, what's wrong with the sound. It's not clear any more. I can't understand anything.*

Since the sound quality was poor for the entire input text for item 31 to 35, most of the participants reported having difficulty answering these items. From their think-aloud protocols, it was clear that they chose the answers mainly by matching an option with a word or phrase that they were able to catch in a stream of blurry sounds. The poor sound quality apparently had a negative impact on the participants' performance on item 31 to 35.

## Summary

The evidence from the verbal data revealed that the cognitive load of the LCLT items was governed by not only the listening subskills they engaged but also a number of text and task-related factors. The characteristics of the input text that were found to affect test-takers' performance were the number of speakers in a conversation and speech rate. The task-related factors emerging from the participants' verbal reports were the multiple-choice format, order of items and lexical overlap. Apart from that, poor sound quality was also identified as a considerable source of construct-irrelevant variance. Taken together, all these factors highlight the multifaceted nature of listening task difficulty. In Chapter 7, they will be further discussed in terms of the practical considerations for the assessment of listening in general.

### **5.5. Judgement on the first warrant of the explanation inference**

This study provides evidence for the warrant which holds that the LCLT items actually engaged the listening subskills described for level 3, 4 and 5 of the CEFR-V. The most important findings will be summarised in this section so as to lay the ground for the final judgement on this warrant.

#### **5.5.1. Poor coverage of the targeted listening construct**

Evidence gathered from the participants' responses revealed that the LCLT items provided very limited coverage of the targeted construct which stretched over a wide range of 16 listening subskills. In fact, only 5 subskills were covered by the test and they are:

- Understanding details of discussion around him/her provided speech is clearly articulated (level 3);
- Understanding details of discussion around him/her when they are not clearly articulated (level 4);
- Listening and making simple inferences about the context (level 4);

- Listening and making inferences about the speakers' attitude and opinion through his/her tone and language use (level 4);
- Understanding discourse function and the purpose of the speaker (level 5).

Despite a strong emphasis on measuring the ability to understand the main ideas of the input text, there was no evidence of this subskill both in the way the test items were written and in the participants' responses. While listening for main ideas has been regarded as a key component in not only traditional taxonomies (Richards, 1983; Weir, 1993) but also the CEFR and CEFR-V scale descriptors, it was completely ignored in the LCLT. Therefore, none of the participants had the opportunity to use their ability to process input and synthesize information to achieve global understanding of the listening text. Also, there was no evidence of higher-level listening subskills such as:

- Understanding the purpose of the speaker even when it is not clearly and directly articulated (level 5);
- Understanding details in propositionally and linguistically complex speech (level 5);
- Understanding idiomatic expressions (level 5).

Among the five subskills that were actually covered by the LCLT, the strongest focus, was on the ability to identify details of a discussion both when they are clearly articulated and when they are implied. The ability to make inferences was also frequently reported in the participants' responses. Meanwhile, not much evidence was found about how the subjects made use of their discourse knowledge to understand the purpose of the speaker.

The results of this study indicated a significant gap between the target construct and the listening abilities that the LCLT was able to measure. This gap represents one of the major sources of invalidity, the situation in which the assessment is too narrow and fails to include important dimensions of the construct (Messick, 1995).

### 5.5.2. Presence of construct-irrelevant variance

Overall, the evidence considered suggests that the LCLT's validity was threatened by construct-irrelevant variance. The multiple-choice format allowed the participants to make guesses and depend on the written text when they were not able to understand the auditory input. Apart from that, the poor sound quality which should be avoided at all costs was found to be a problem for the last sound file of the test. Such evidence of construct contamination apparently casts doubts on the interpretations of the LCLT scores.

Haladyna and Downing (2004) suggested a wide range of factors to be considered which include test preparation, test development, test administration, scoring, student-related characteristics and cheating. However, given the nature of this study, a comprehensive investigation of all these factors was not possible. Since all the participants were invited to take the test in a non-test condition, their performance was not influenced by student-related factors such as test anxiety or motivation. Moreover, in a one-on-one, worry-free setting, aspects of test administration as in a real test event were clearly not applicable. All the 10 students participating in this study were LCLT first-timers and therefore test preparation was not an issue. Also, there were no concerns related to scoring and cheating because scoring was objective and the participants had no reason to cheat. With all other factors screened out, this study can only shed light on the construct-irrelevant variance induced by the input text and the design of test items. However, in a real test event, the LCLT's validity might have been threatened by other sources of construct-irrelevant variance.

In sum, evidence from the verbal data suggests that the LCLT suffered from both construct under-representation and construct-irrelevant variance. Therefore, the warrant which holds that the target construct was well-represented by the LCLT items was not supported.



## **Chapter 6. The comparability between the LCLT and IELTS Listening**

This chapter contributes to the evaluation of the explanation inference by providing evidence for the second warrant which holds that test-takers' scores on the LCLT correlated well with scores on an international English proficiency test that measured a similar construct. It reports on the comparability study between the LCLT and an IELTS Listening test. In this chapter, the background of the study is first presented. This overview is followed by a review of comparability studies on language tests. The methodology and findings of this study are then reported before a decision is made about the second warrant of the explanation inference.

### **6.1. Background of the study**

Although the localized test of English proficiency in Vietnam is currently developed and intended for domestic use only, it claims alignment with the CEFR and aims to reach a level of comparability with international tests of English proficiency. This broader vision is of great importance since it is the desire of test-takers and test scores users, nowadays, to look further than one test and examine the equivalence among various tests of English proficiency (Lim & Khalifa, 2013). With this being the case, a study that compares the Vietnamese localized test to an international measure of English proficiency will be useful because it sheds light on how the local test played out in terms of international benchmarks of proficiency. This chapter reports on a study that compared the LCLT with the listening component of the International English Language Testing System, also known as the IELTS test.

There are two main reasons why the IELTS test was chosen for this comparability study. First of all, among all the current tests of English proficiency in Vietnam such as the Test of English as a Foreign Language (TOEFL), the Test of English for International Communication (TOEIC) or the newly introduced Aptis test, IELTS is still the most popular

one. Over the past 20 years, this test has gained public acceptance as a reliable indicator of test-takers' proficiency in English. Therefore, it is not uncommon for any new test of English proficiency to be considered against the standard that the IELTS has set. Secondly, test-takers' results on the LCLT and IELTS can be used in Vietnam for the same purposes which are higher education and employment. However, different universities, organizations and companies may have different policies regarding which test is accepted. In general, the popularity of IELTS and the common purposes that it shares with the LCLT has made this test an obvious choice for an external measure of proficiency that the LCLT can be compared to. Although IELTS is a well-established test of English proficiency which has been accepted internationally, it is not looked at, in this study, as a kind of standard that the LCLT should follow. The two tests were compared in an objective manner with the view to informing test developers, test-takers and test-score users of how comparable they are rather than implying a superior-inferior relationship between them.

## **6.2. Comparability studies on language tests**

This section is intended to give an overview of comparative test analysis and validation studies which are relevant to the context and purposes of this present study. In particular, attention will be paid to a number of studies in which well-established international tests of English proficiency were compared to each other. Another important part is also dedicated to studies that compared locally developed tests in Asia to internationally recognized tests of English proficiency. In so doing, this section aims to provide insights into different approaches to investigating cross-test comparability which has become increasingly important to test-takers and test score users nowadays (Lim & Khalifa, 2013).

A seminal project in terms of developing a systematic approach to the comparison of language tests is the Cambridge-TOEFL comparability study conducted by Bachman, Davidson, Ryan and Choi (1995). This study compared the First Certificate of English

(FCE) administered by the University of Cambridge Local Examinations Syndicate and the paper-based version of the Test of English as a Foreign Language (TOEFL) administered by the Educational Testing Service (ETS), Princeton. Data analysis for this study included a qualitative content analysis conducted by expert judges using Bachman's (1990) model of Communicative Language Ability and Test Method Facets. A range of quantitative analyses were also utilized to investigate the test performance and factor structure of the test batteries. Although the research started with the hypothesis that the different measurement traditions from which the tests were developed would result in “clear and striking differences” in both test content and test performance (Davidson & Bachman, 1990, p. 28), the research results revealed a substantial amount of overlap. The study concluded that there were more similarities than differences in terms of test content and the two tests were generally measuring similar language abilities.

Around the same time with the Cambridge – TOEFL project, Geranpayeh (1994) conducted a comparability study on IELTS and TOEFL, focusing solely on test scores. In an attempt to justify score comparison between the two tests, Geranpayeh (1994) compared the results of 216 subjects who took both tests from 1990 to 1992. Moderate to high correlations between the paper-based TOEFL and IELTS scores were found. The results of factor analysis suggested that IELTS and TOEFL were likely to share similar internal structure and thus might provide similar information of test-takers' language ability.

Along the same line, more recently, a study was carried out by ETS (2010) comparing test-takers' performance on the new internet-based TOEFL and IELTS. This study compared the results of 1,153 test-takers on each section, i.e., Listening, Speaking, Reading, and Writing as well as their overall scores on both tests. The quantitative

analysis of test performance revealed that the correlations between IELTS and TOEFL in terms of Writing and Speaking scores were not very strong. Meanwhile, moderately high correlations were observed in the Listening and Reading scores. The correlation between total scores on both tests was indeed higher than any section score correlations. Even though the results of this study showed that test-takers' overall scores on IELTS and TOEFL were comparable, it was not clear whether IELTS and TOEFL were sharing the same underlying construct.

Influential as they are, major international examinations, such as IELTS and TOEFL, are only likely to work when they are used for the right purpose and on the population they have been designed for (O'Sullivan, 2012). The consideration of the test-taking population and its own context has given rise to localized tests of English proficiency. In Asia, the General English Proficiency Test (GEPT) of Taiwan and the EIKEN test of Japan are the two much-quoted examples of localized tests that have been recognized in other countries. Research has been carried out to seek evidence of the comparability between these localized tests and international tests of English proficiency.

In 2013, Weir, Chan and Nakatsuhara (2013) conducted a study to establish the statistical cross test comparability of the GEPT Advanced and IELTS in their reading and writing subtests. The correlation between GEPT-A and IELTS reading scores was found to be moderate (.520). However, the findings suggested that it was harder to pass the GEPT-A reading test than to score 6.5 or above in the IELTS reading test. Regarding the writing scores, moderate to large correlations were obtained between GEPT-A scores and IELTS writing bands. It was also suggested that passing the GEPT-A writing test was more difficult than scoring 6.5 or above on the IELTS writing test. Significant positive correlations implied that the Reading and Writing tests of GEPT-A and IELTS might have measured the same underlying construct; however, this was not verified by any empirical evidence.

More recently, in 2016, the British Council conducted a comparison study between GEPT and the Aptis test (Wu, Yeh, Dunlea & Spiby, 2016), using the socio-cognitive framework for test development and validation (Weir, 2005; O'Sullivan & Weir, 2011). In this study, the constructs of both tests were analyzed using factor analysis and content analysis while test performances were compared through statistical analysis. The findings revealed that the Aptis subtests correlated significantly with all the GEPT subtests at moderate to high levels. The results of Factor Analysis indicated that GEPT and Aptis measured the same construct. Content analysis which was conducted by means of expert judgement showed that the Aptis Listening, Speaking and Reading was more difficult than the GEPT. However, the GEPT Writing test was considered more cognitively challenging than the Aptis.

Another attempt to compare the GEPT with an well-established international test was the comparability study between the GEPT-A and TOEFL (Kunnan & Carr, 2017) which focused on the reading and writing component. Both qualitative content analysis and quantitative statistical analysis were conducted on the reading component. Meanwhile, for the writing subtests, only quantitative analysis of test-takers' performance was carried out. The reading texts of GEPT-A were found to be more challenging in terms of vocabulary and syntax. The statistical analysis showed that test-takers performed better on the TOEFL iBT reading than on the GEPT-A reading. Similar results were observed for the performance on the writing test. However, it was also found that test-takers' reading and writing scores on the GEPT-A correlated significantly with their TOEFL iBT scores.

Regarding the EIKEN test, the best-known comparability study to date was the one that compared this test with the TOEFL iBT in terms of both construct and score

performance (Brown et al., 2012). This study targeted the three upper-level EIKEN exams (Grade 1, Grade Pre-1 and Grade 2), each of which was divided into reading, listening, writing and speaking subsections. The results of correlational analysis showed that the total scores of EIKEN and TOEFL correlated strongly with each other. This provides evidence that test-takers' total scores on EIKEN can predict their total scores on TOEFL iBT and vice versa. In terms of test construct, a two-component Principle Component Analysis showed a pattern of loadings indicating that the two tests were measuring similar constructs.

The studies reviewed previously reflect the efforts by both local and international researchers in examining the link between various tests of English proficiency available for test-takers these days. In general, the studies on international tests of English proficiency revealed good signs of comparability between them in terms of test scores. However, due attention was not paid to the comparison of test construct when most studies of this kind relied on factor analysis as the only source of evidence for construct comparability. Meanwhile, factor analysis is the method that works out the process by looking at the test results (Field, 2013) rather than investigating the cognitive processes that underlie the design of test items. Therefore, it fails to elucidate what the test construct actually entails. The lack of depth in construct comparison can be seen as a major limitation of most comparability studies mentioned in this review.

### **6.3. Methodology**

This study aims to elicit evidence for the second warrant of the explanation inference which holds that test-takers' scores on the LCLT correlated well with their scores on an IELTS Listening test. It should be emphasized that the comparison of the constructs is a prerequisite for the comparison of test scores. There is no point in comparing test-takers' scores on two tests that measure different constructs. Therefore, this study compares the

LCLT and IELTS in two substantial areas, the test construct, and the test-takers' performance. This comparison requires a mixed method approach that includes both qualitative and quantitative analyses. Data were collected via two mock tests administered to the same group of students from the host university.

### **6.3.1. Participants**

The study participants included 66 Vietnamese students. At the time of the study, they were all second-year English majors at the host university. All of them were female students, aged 19 to 20. These students were relatively familiar with the format of the IELTS Listening test but they never took the LCLT before. Therefore, prior to the administration of the LCLT, the test was properly introduced to ensure that the participants were well-informed and understood what they were expected to do. The information sheet and consent form for this study are provided in Appendix G and H.

### **6.3.2. Test administration**

The instruments needed for data collection are the LCLT and a retired version of an IELTS Listening test. Test materials included test papers and recordings of the listening texts from both tests. While security and confidentiality were required for the LCLT materials, the IELTS Listening test paper and recording were published in the practice test collection by Cambridge University Press.

The group of 66 English majors were invited to take the LCLT first in a mock test event. That said, the test was administered under strict testing conditions and participants were encouraged to try their best. The interval between the LCLT and the IELTS Listening test was one week which was not long enough to risk any significant change in the students' proficiency level, either in terms of gain or loss. Although a counterbalanced design in which a certain number of students take the LCLT first and others take the IELTS first would have been a more robust data collection method, it was not possible due to objective reasons.

After all these steps, the data obtained for analysis included the test papers and recordings of both the LCLT and IELTS. The results of 66 participants for both tests were also recorded for the comparison of test-takers' performance.

### **6.3.3. Expert judges**

Since the qualitative analysis of this study was conducted by means of expert judgement, a team of reviewers was organized for this particular purpose. This team consisted of five reviewers, including myself. The reviewers were Vietnamese teachers of English who had at least 5 years of teaching experience at the time of the study. They all held Master's Degrees of Teaching English to Speakers of Other Languages (TESOL). Each reviewer was from a different university in Vietnam.

This study took a consensus approach to expert judgement, which means the reviewers were first asked to give their expert opinions on certain aspects of the LCLT and the IELTS Listening test. Then, the results were discussed as a group to identify sources of disagreement and resolve any differences. The last step was to arrive at a final evaluation that all the five reviewers agreed on.

### **6.3.4. Data analysis**

The comparison of test construct is an important part of content analysis since it helps to verify whether both the LCLT and the IELTS Listening test were measuring the same construct. This investigation of test construct is usually conducted by means of Factor Analysis and/or expert judgement. The latter was chosen for this study because of two reasons. First, the sample size is not large enough to ensure reliable outcomes for Factor Analysis. While a sample size of 100 is the minimum requirement for Factor Analysis, only 66 participants were recruited for this study. Second, Factor Analysis is the method that works out the process by looking at the test results (Field, 2013) rather than investigating the cognitive processes that underlie the design of test items. Factor analysis therefore does not



provide useful insights into what the test construct might entail. For these reasons, expert judgement was an obvious choice.

At this stage of the study, the team of five item reviewers (expert judges) undertook a qualitative investigation of the cognitive demand for each test item in both tests. An evaluation template was constructed with a list of Listening sub-skills that reviewers could match with each test item. While the underlying construct for the LCLT was based on CEFR-V (which was adapted from CEFR), the construct for the IELTS Listening test was not. Therefore, the comparison of construct between these two tests was done via an unrelated theoretical framework of listening ability. Among a number of descriptive models of communicative competence (Bachman, 1990; Canale, 1983; Canale & Swain, 1980), Bachman and Palmer's (1996) framework is by far the most widely accepted general description of language ability among language testers (Buck, 2001). This framework was adapted by Buck (2001) in an attempt to aid test development by identifying the components of listening. It looks at not only language competence but also strategic competence in listening. However, Buck (2001) did point out that in second-language testing, more emphasis should be put on testing language competence rather than strategic competence. The reason is that, for adult language learners, their cognitive ability is relatively developed and stable while their language competence is only partially developed. For that reason, the differences in performance between individual listeners is generally due to the differences in their language competence rather than their strategic competence. Therefore, the investigation of the listening construct in this study focused only on language competence. The framework of listening ability as suggested by Buck (2001) was used for the comparison of construct between the LCLT and the IELTS (Table 6.1).

**Table 6.1.** The framework for describing listening ability

<b>Language competence</b>	Grammatical knowledge	Understanding short utterances on a literal semantic level. This includes phonology, stress, intonation, spoken vocabulary, spoken syntax
	Discourse knowledge	Understanding longer utterances or interactive discourse between two or more speakers. This includes knowledge of discourse features, such as cohesion, foregrounding, rhetorical schemata, story grammars, and knowledge of the structure of unplanned discourse
	Pragmatic knowledge	Understanding the function or the illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that. This includes understanding whether utterances are intended to convey ideas, manipulate, learn or are for creative expression, as well as understanding indirect speech acts and pragmatic implications.
	Socio-linguistic knowledge	Understanding the language of particular socio-cultural settings, and interpreting utterances in terms of the context of situation. This includes knowledge of appropriate linguistic forms and conventions characteristic of particular socio-linguistic group, and the implications of their use, or non-use, such as slang, idiomatic expressions, dialects, cultural references, figures of speech, level of formality and register.

Based on this framework, the expert judges gave their opinions on the type of knowledge that each item of the LCLT and IELTS required. An evaluation template for this analysis is provided in Appendix J.

Once the reviewing team and the evaluation template were ready, a training session was organized for all the reviewers. This session serves a threefold purpose which was to (a)

familiarize reviewers with the LCLT, (b) provide more insights into the IELTS Listening test, and (c) make sure that the reviewers had correct understanding of the evaluation template. Afterwards, the whole team had three meetings together on different dates. In the first meeting, the team carried out the analysis for the LCLT, following a three-step process. Everyone was initially asked to analyse and evaluate the LCLT's input texts as well as each test item individually. Then, the results were discussed as a group to identify sources of disagreement and resolve any differences. The last step was to arrive at a consensus evaluation of the input text and all the test items. This process followed the recommendation by Alderson et al. (2006) and was repeated in the second meeting which focused on the IELTS Listening test. The outcomes of these two meetings were two consensus evaluation templates which prepared the reviewers for the third and final meeting. At this final stage of the analysis, the whole team was convened one more time to discuss and come to a joint conclusion of the similarities and differences between the LCLT and IELTS in terms of their underlying constructs.

The performance of test-takers on the LCLT and IELTS test was compared by means of descriptive statistics, item analysis, reliability and correlational analysis. First, the descriptive statistics for each test were calculated using SPSS 21. In order to compare both tests in terms of statistical characteristics of test items and reliability, Rasch analysis was run using Winsteps (Version 4.0.1). Next, the correlation between test-takers' scores on the LCLT and their scores on the IELTS Listening test was calculated, also with SPSS 21. A Pearson  $r$  correlation coefficient was used to measure the relationship between the LCLT and IELTS test scores.

#### **6.4. Evidence of comparability between the LCLT and IELTS**

This section reports on the findings obtained from the comparison between the LCLT and the IELTS Listening test. Evidence of the similarities and differences between these two

tests will be presented in terms of both the test construct and the test-takers' performance. These findings provide crucial insights into the comparability between the two tests, making it possible for a conclusion to be made about the second warrant of the explanation inference.

#### **6.4.1. Results of construct analysis**

In the analysis of the listening constructs, the expert judges identified the major processes and subskills required by both tests. We found most of our views on test items complimentary and thus were able to reach a consensus about the final judgement for each test. This section reports on the listening construct for the LCLT and IELTS as agreed by the item reviewers.

##### **6.4.1.1. LCLT listening construct**

The language competence investigated in this study consists of four different types of knowledge which are grammatical knowledge, discourse knowledge, pragmatic knowledge and sociolinguistic knowledge. There was an overwhelming agreement that the first three types of knowledge played a significant role in answering LCLT's test items. Meanwhile, only a small percentage of test items required test-takers to use their sociolinguistic knowledge. It is important to note that in many cases, one single item can cover more than one aspect of the language competence. A summary of construct coverage for the LCLT is presented in Table 6.2.

**Table 6.2.** Expert judgements on the LCLT construct

Language competence	Item	Total number of items
<b>Grammatical knowledge: understanding short utterances on a literal semantic level</b>		
Pre-existing knowledge of grammar, vocabulary, syntax, phonology	All items	35 items – 100%
<b>Discourse knowledge: understanding longer utterances or interactive discourse between two or more speakers</b>		
Cohesion	1, 3, 5, 6, 7, 10, 11, 12, 13, 14, 15, 16, 17, 20, 23, 24, 32, 34, 35	19 items – 54%
Foregrounding	20	
Rhetorical schemata	x	
Story grammars	x	
Structure of unplanned discourse	x	
<b>Pragmatic knowledge: understanding the function or the illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that</b>		
Indirect meaning/hints	1, 3, 7, 9, 10, 11	21 items – 60%
Pragmatic implications	3, 11, 21	
Text-based inferences	1, 3, 5, 8, 10, 11, 14, 16, 17, 18, 19, 20, 21, 22, 29, 31, 32, 34, 35	
<b>Sociolinguistics knowledge: understanding the language of particular socio-cultural settings and interpreting utterances in terms of the context of situation</b>		
Slang	x	2 items – 6%
Idiomatic expressions	19	
Dialects	x	
Cultural references	10	
Figures of speech (comparison, metaphor, euphemism, understatement, sarcasm...)	x	
Levels of formality and registers	x	

As far as grammatical knowledge is concerned, it was agreed that all items of the LCLT tapped into this constituent of language competence with a view to examining whether test-takers were able to understand short utterances on a literal semantic level. This type of knowledge includes phonology, stress, intonation, spoken vocabulary and spoken syntax (Buck, 2001) and we agreed that all items relied on this. However, some items required more difficult processing than others. For example, the item that required test takers to recognize “a variety of” as an equivalence of “many” was considered easier than the one that asked test-takers to match “feeling pretty low” with “feeling depressed”.

In terms of discourse knowledge which underlies test-takers’ ability to understand longer utterances or interactive discourse between two or more speakers, a total number of 19 items were found to measure this type of knowledge. Almost all of them measured test-takers sensitivity to cohesion while only one item was found to target the aspect of spoken discourse known as foregrounding. The other aspects of discourse knowledge, namely rhetorical schemata, story grammars and structure of unplanned discourse were not covered by any test items.

Regarding pragmatic knowledge, the item reviewers were able to identify 21 items which were intended to assess the ability to understand the function of an utterance and interpret its meaning accordingly. As can be seen from Table 6.2, a dominant number of items assessed the ability to make text-based inferences. Indirect meaning/hints and pragmatic implications were also covered by test items but only to a modest extent.

Sociolinguistic knowledge was not a major focus of the LCLT. Only two items tapped into this kind of knowledge, covering idiomatic expressions (item 19) and cultural references (item 10).

#### **6.4.1.2. IELTS listening construct**

Similar to the LCLT, it was broadly agreed among item reviewers that all test items of IELTS measured grammatical knowledge in a way that they required pre-existing knowledge of grammar, vocabulary, syntax, phonology and so on. As seen from Table 6.3, a majority of items were found to measure test-takers' sensitivity to cohesion. This did not come as a surprise given that this test used only long texts and the understanding of cohesion was crucial for comprehension. A total number of 35 items, representing 87.5% of the test, were found to target test-takers knowledge of cohesion. Apart from cohesion, items of the IELTS test did not cover any other constituents of discourse knowledge.

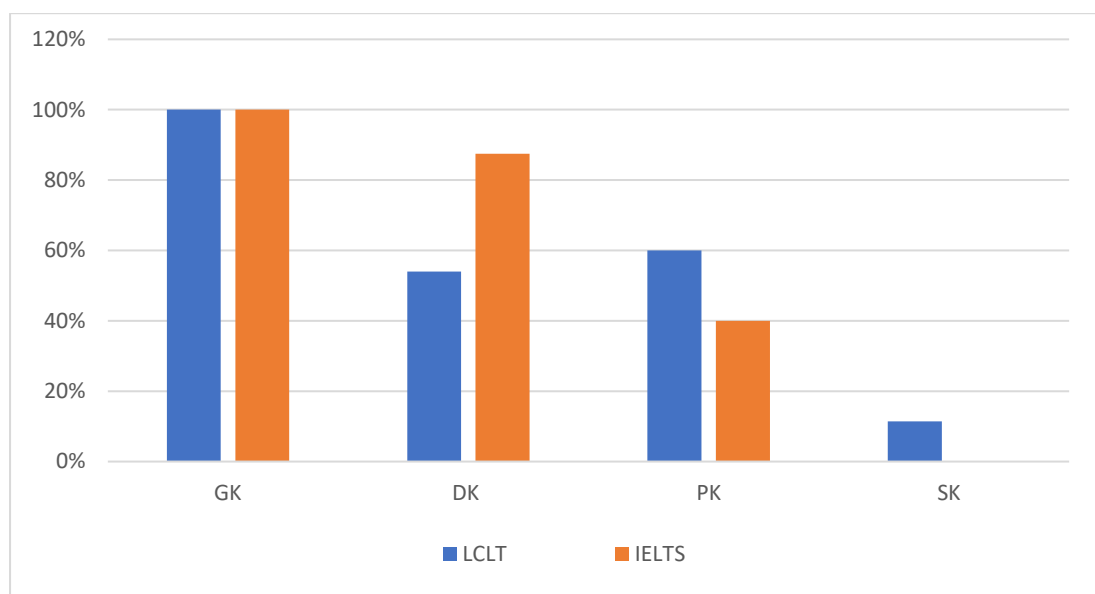
With regard to pragmatic knowledge, there was a strong focus on text-based inferences. While no items were found to target test-takers' sensitivity to indirect meaning and pragmatic implications, the item reviewers agreed on the 16 items that were designed to assess the ability to make text-based inferences. In terms of sociolinguistic knowledge, a consensus was reached between the item reviewers that this component of language competence was not targeted in this test. Other aspects which are usually considered challenging for second language learners like slang, idioms, figures of speech, and levels of formality were also not present in this IELTS test.

**Table 6.3.** Expert judgements on the IELTS construct

Language competence	Item	Total number of items
<b>Grammatical knowledge: understanding short utterances on a literal semantic level</b>		
Pre-existing knowledge of grammar, vocabulary, syntax, phonology	All items	40 items – 100%
<b>Discourse knowledge: understanding longer utterances or interactive discourse between two or more speakers</b>		
Cohesion	3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40	35 items – 87.5%
Foregrounding	x	
Rhetorical schemata	x	
Story grammars	x	
Structure of unplanned discourse	x	
<b>Pragmatic knowledge: understanding the function or the illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that</b>		
Indirect meaning/hints	x	16 items – 40%
Pragmatic implications	x	
Text-based inferences	9, 11, 12, 13, 14, 15, 16, 17, 22, 31, 32, 33, 34, 35, 36, 40	
<b>Sociolinguistics knowledge: understanding the language of particular socio-cultural settings and interpreting utterances in terms of the context of situation</b>		
Slang	x	0 items – 0%
Idiomatic expressions	x	
Dialects	x	
Cultural references	x	
Figures of speech (comparison, metaphor, euphemism, understatement, sarcasm...)	x	
Levels of formality and registers	x	



Despite differences in the extent to which each type of knowledge was covered by the LCLT and IELTS, the overall evaluation revealed that the constructs of both tests matched together relatively well. In general, the LCLT and IELTS were measuring similar listening abilities, focusing on grammatical knowledge (GK), discourse knowledge (DK) and pragmatic knowledge (PK) (Figure 6.1).



**Figure 6.1.** Comparison of the listening construct between the LCLT and IELTS

In terms of grammatical knowledge, both tests were in complete agreement with each other when all of the items required pre-existing knowledge of grammar, vocabulary, lexicon, phonology and so on. As far as discourse knowledge is concerned, both the LCLT and IELTS put a strong emphasis on test-takers' sensitivity to cohesion. Regarding pragmatic knowledge, a dominant focus on text-based inferences was observed in both tests. Despite the minor difference in the involvement of sociolinguistic knowledge (SK), the construct of the LCLT did match that of the IELTS to a reasonable degree. This finding is of great importance since it lays the ground for further comparison to be made between the LCLT and IELTS.

## 6.4.2. Results of the quantitative analysis on test-takers' performance

### 6.4.2.1. Descriptive statistics

Valid test data was collected from 66 volunteers who took both the LCLT and the IELTS Listening test. Table 6.4 presents the descriptive statistics for these test-takers' scores on both tests.

**Table 6.4.** Descriptive statistics for the LCLT and IELTS scores

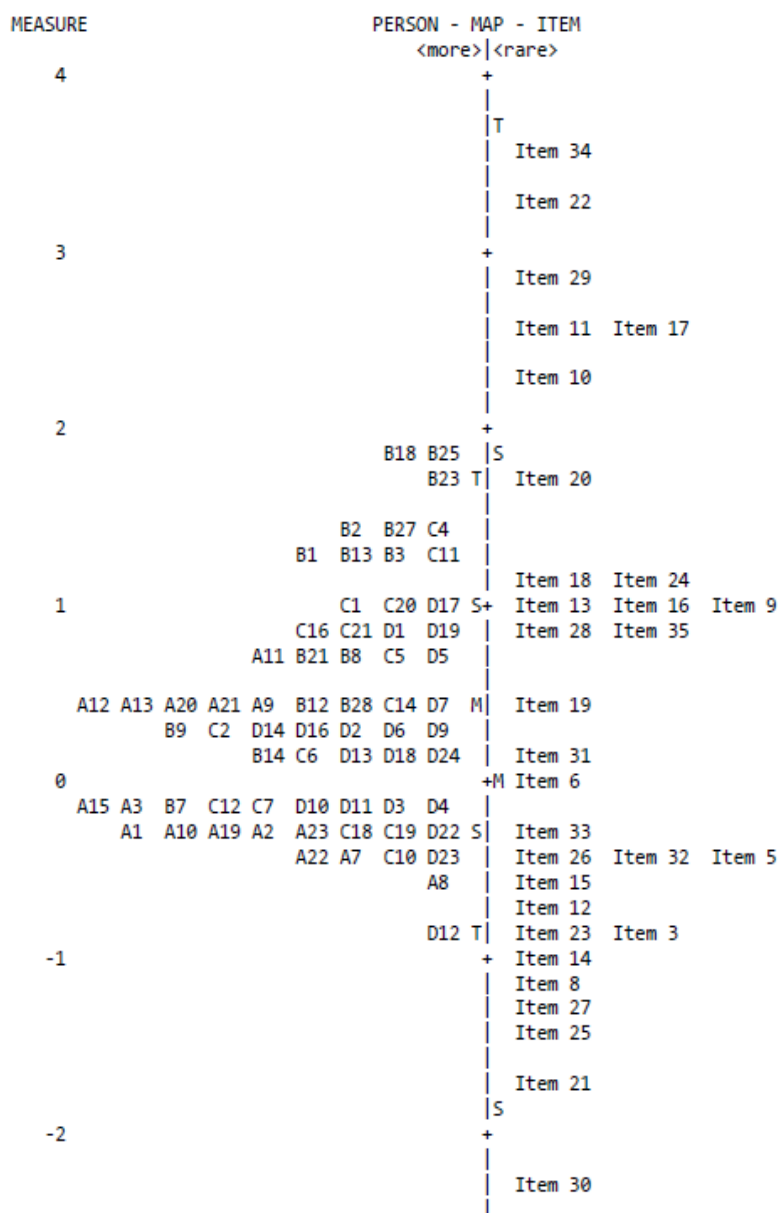
LCLT	IELTS Listening test
Mean = 19.470	Mean = 25.712
Median = 19	Median = 26
Mode = 17	Mode = 26
SD = 3.32	SD = 4.85
Skewness = .410	Skewness = .068
Kurtosis = -.513	Kurtosis = -.247
Min = 13	Min = 15
Max = 27	Max = 37
Range = 14	Range = 22

The differences in test-takers' performance between the LLCT and the IELTS Listening test were clearly seen from the descriptive statistics. While the minimum scores on both tests were quite close to each other (13 for the LCLT and 15 for the IELTS), a remarkable gap was noticed in the maximum scores. Specifically, the maximum score on the LCLT was only 27 out of 35. Given that this group of test-takers were all English majors, this result was quite low. Meanwhile, the maximum score on IELTS was 37 out of 40, which was very high. Mean, median and mode values were noticeably higher for IELTS. The standard deviation was 3.32 for the LCLT and 4.85 for IELTS, suggesting that test scores varied to a greater extent in the IELTS test. Taken together, descriptive statistics showed that this group of English majors performed much better on the IELTS Listening test. This could be inferred that, for this group of test-takers, IELTS was much easier than the LCLT.

#### 6.4.2.2. Item analysis

Item analysis for both the LCLT and IELTS was carried out with Rasch measurement. In this section, the results related to item difficulty, person ability and construct-irrelevant variance will be presented in detail.

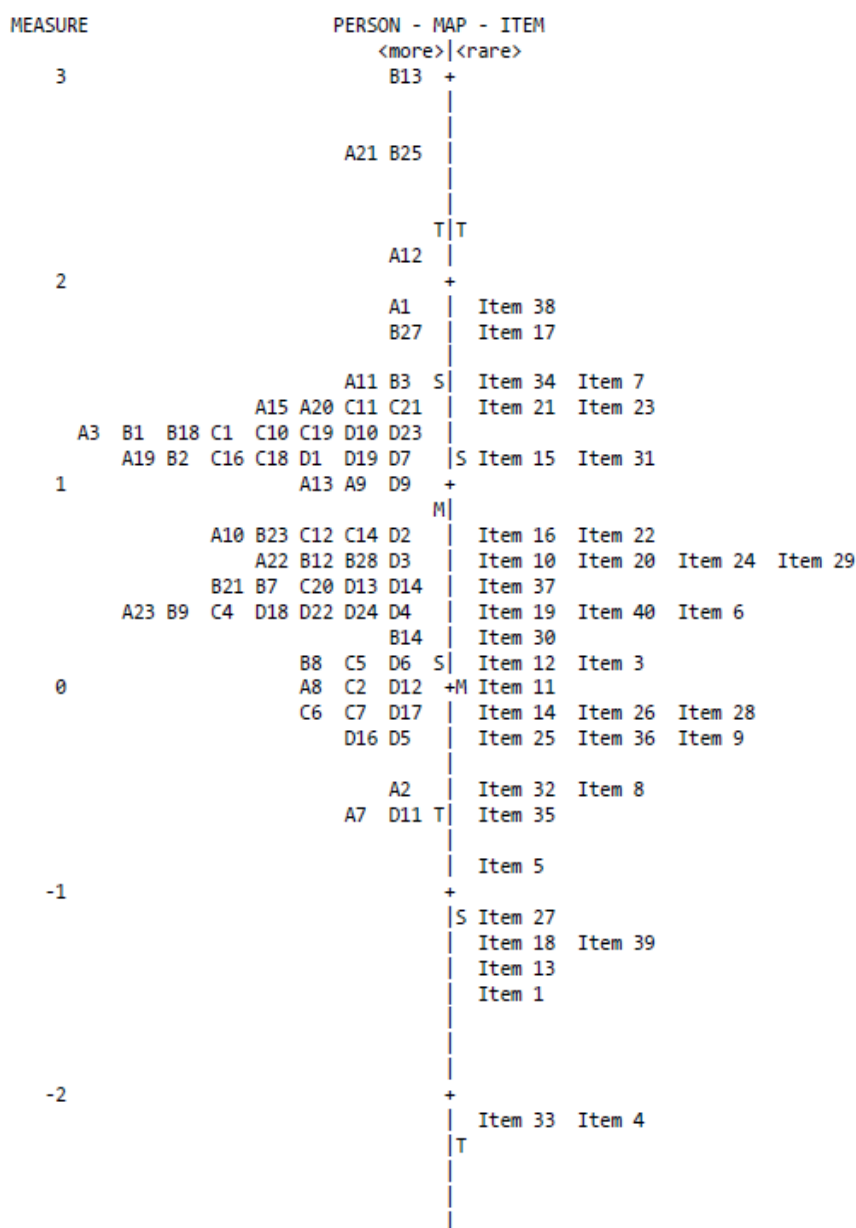
The relationship between item difficulty and test-takers' ability for the LCLT is schematized in Figure 6.2. As can be seen, test-takers clustered in the middle of the map rather than spreading out, which suggested that they belonged to a homogenous group with similar ability levels. Located at the bottom of the map were 10 items which were very easy for this group. Despite this large number of easy items, there were still six very difficult items that prevented test-takers from scoring high. That helps to explain why the highest score recorded was only 27 out of 35. In general, the LCLT was still difficult for this group of 66 English majors. The results shown in Figure 6.2 shed further light on the descriptive statistics presented previously.



**Figure 6.2.** Item difficulty and person ability in the LCLT

Figure 6.3 revealed a very different picture about the relationship between item difficulty and person ability for the IELTS Listening test. The same group of test-takers tended to spread out rather than clustering in the middle. None of the items was too difficult for this group but instead, some test-takers were found to have an ability level even beyond the most difficult item of the test. More specifically, test-takers B13, A21, B25 and A12 were

not matched to any test items since their ability levels were higher. Below these levels, a relatively good match was observed between test-takers' ability and item difficulty. However, at the bottom of the map, in the area from -3.5 to -1 logit, there were 9 items that were not matched to any test-takers because they were too easy. These easy items accounted for 22.5% of the test. In general, the IELTS Listening test was easy for this group of 66 English majors. That also confirms the descriptive statistics reported previously for the IELTS test.



**Figure 6.3.** Item difficulty and person ability in IELTS

Thus far, differences in the test-takers' performance on both the LCLT and the IELTS Listening test have been brought to light. For this group of English majors, the LCLT was shown to be much more difficult than IELTS. It is important to acknowledge, at this point, that test familiarity could have been a contributory factor since most of the participants were familiar with the IELTS Listening test. The LCLT, on the other hand, was part of a newly developed test in Vietnam, and this group of 66 English majors had no prior experience with it before participating in this study.

Another important part of item analysis was the identification of construct-irrelevant variance in both tests. Fit statistics for the LCLT and IELTS are presented in Appendix K. In general, the infit MNSQ values for LCLT items fell into the stringent range between 0.8 to 1.2 and the ZSTD values for these items were also within the acceptable range from -2 to 2. Item 32 was the only item that contained construct irrelevant variance, which was evidenced by its large MNSQ value (1.32) and large ZSTD value (3.0). A similar result was seen for the IELTS Listening test when 39 items were shown to have acceptable MNSQ values. Construct-irrelevant variance was identified in only item 16 which had a large MNSQ value (1.23) and large ZSTD (3.0). Taken together, the quantitative analysis of test scores indicated that construct-irrelevant variance was not a serious threat for both the LCLT and IELTS.

#### 6.4.2.3. Test reliability

Table 6.5 summarizes the reliability indices for both the LCLT and IELTS.

**Table 6.5.** Person and item reliability for the LCLT and IELTS

<b>Reliability index</b>	<b>LCLT</b>	<b>IELTS</b>
Item reliability	0.95	0.91
Person reliability	0.48	0.70

The results of Rasch analysis showed that item reliability was high for both the LCLT and IELTS. These values are 0.95 and 0.91 respectively. With this result, we can be confident about the item difficulty hierarchy for both tests. However, there was a stark contrast between them in terms of person reliability. While the IELTS test had an acceptable value of person reliability (0.7), this value for the LCLT was very low (0.48). The order of test-takers' listening ability based on their LCLT scores was therefore questionable. This low person reliability index also implies a lack of items in the LCLT. Overall, suffice it to say that, for this group of test-takers, the LCLT was much less reliable than IELTS.

#### 6.4.2.4. Correlational analysis

As can be seen from Table 6.6, the correlation coefficient was rather low (.311) and the correlation was highly significant ( $p = .011$ ). This result did not indicate a strong correlation between test-takers scores on the LCLT and their scores on the IELTS Listening test. With the significant difference noticed in test-takers' performance on both tests, this finding was to be expected. As it turns out, high scores on the IELTS Listening test did not predict high scores on the LCLT and vice versa.

**Table 6.6.** Correlation between LCLT and IELTS scores

		IELTS scores	LCLT scores
IELTS scores	Pearson Correlation	1	.311*
	Sig. (2-tailed)		.011
	Covariance	23.562	5.030
	N	66	66
LCLT scores	Pearson Correlation	.311*	1
	Sig (2-tailed)	.011	
	Covariance	5.030	11.084
	N	66	66

In studies that compared a localized test with an international test of English proficiency, it was usually found that test-takers' results for both tests were strongly correlated (Brown, 2012; Wu et al., 2016; Weir et al., 2013). However, that did not appear to be the case in this comparison between the LCLT and IELTS. Although the LCLT and IELTS were found to measure a similar listening construct, test-takers' scores on both tests did not correlate well with each other. A thorough explanation for the low correlation would have to come from in-depth content analyses which are realistically beyond the scope of this thesis. That being said, it is likely that the differences between the two tests in terms of the input texts and the item design resulted in marked differences in the test-takers' performances. In light of the findings from the study on test-takers' cognitive processes (Chapter 5), the number of speakers, speech rate and the sound quality all had adverse impacts on the test-takers' performance on the LCLT. These factors might have been better controlled for the IELTS Listening test, which possibly led to better results for the test-takers. Test format is another aspect that should be taken into consideration. While the LCLT depended entirely on the multiple-choice questions, the IELTS test employed a variety of response formats such as multiple-choice, matching, gap filling or short answers. The use of different formats might have reduced the chances of lucky guesses as well as the use of test-taking strategies associated with a particular type of question.

### **6.5. Judgement on the second warrant of the explanation inference**

The findings of this comparability have revealed that the LCLT and IELTS were measuring similar listening construct. However, test-takers' performance differed greatly with much better results for the IELTS Listening test. The most disappointing finding is the low correlation between scores on the LCLT and scores on the IELTS. This result refuted the warrant which holds that test-takers' performance on the LCLT correlated well with their performance on an international test that measured a similar listening construct.



The significant differences between test-takers' scores on the LCLT and their scores on the IELTS Listening test raised intriguing questions about the causes of such differences. These questions can be answered by an in-depth content analysis for both tests. It is suspected that various aspects of test design might have been responsible for the test-takers' poor performance on the LCLT. Factors related to the input text such as text speed, text length or accent might have had considerable impacts on the test-takers' performance. Other than that, the differences in the nature of the test tasks or in the response formats should also be taken into consideration.

## **Chapter 7. Overall discussion**

This chapter recapitulates the argument-based validation of the LCLT and brings together all the major findings of the research. In light of the evidence gathered from the quantitative study on the LCLT scores (Chapter 4), the qualitative study on test-takers' cognitive processes (Chapter 5), and the comparability study between the LCLT and an IELTS Listening test (Chapter 6), a final judgment will be made about whether the argument-based validity holds up in practice. From there, major problems with the LCLT will be thoroughly discussed. Looking further than this particular listening test, this chapter highlights important issues in localizing high-stakes tests of English proficiency. The rest of the chapter provides a discussion of notable concerns in the assessment of second language listening and a critical reflection on the argument-based approach to validation.

### **7.1. Argument-based validity of the LCLT**

This research project aims to develop a validity argument for the evaluation, generalization and explanation inferences of the LCLT. The evaluation inference relies on three warrants related to the statistical characteristics of the LCLT scores. The generalization inference hinges on one warrant related to the LCLT's reliability. The explanation inference rests on two major warrants about (a) the extent to which the LCLT items engaged the listening subskills described for level 3, 4 and 5 in CEFR-V and (b) the correlation between the test-takers' scores on the LCLT and their scores on an international test of English proficiency. The evidence needed to assess these warrants was gathered from three different studies. Table 7.1 summarizes the inferences, warrants and studies conducted in this research.

**Table 7.1.** Summary of the inferences, warrants and relevant studies

Inference	Warrant	Study
<p><b>Evaluation inference:</b> the test-takers' performances on the LCLT were appropriately observed and scored.</p>	Item difficulty was appropriate for test-takers.	<p>Quantitative study on the LCLT scores (Chapter 4)</p>
	The LCLT items had reasonable discriminating power.	
	Test items were relevant to the targeted listening construct.	
<p><b>Generalization inference:</b> The LCLT scores were reliable</p>	The LCLT has high item and person reliability indices	
<p><b>Explanation inference:</b> test-takers' performances on the LCLT are reflective of the targeted listening construct.</p>	The LCLT items actually engaged the listening subskills described for level 3, 4 and 5 of CEFR-V.	<p>Qualitative study of test-takers' cognitive processes (Chapter 5)</p>
	Test-takers' scores on the LCLT correlates with scores on an international test of English proficiency that measures the same construct.	<p>Comparability study between the LCLT and IELTS (Chapter 6)</p>

While the evaluation inference, to a certain extent, were supported by validity evidence, the other two inferences were not. Before arriving at an overall judgement on the validity of the LCLT, the assessment of each inference can be restated as follows.

### **The evaluation inference**

*The test-takers' performance was appropriately observed and scored.*

This inference was supported for two reasons. First, the discriminating power of the LCLT were quite reasonable. The discrimination index was moderate for the majority of the test items. Eight items were found to discriminate test-takers particularly well and only two items had low discrimination indices. Second, most of the test items were found to be relevant to the targeted construct. As evidenced by the quantitative analysis of test scores, construct-irrelevant variance was present in only three items.

It should be noted that the evaluation inference was supported but only to a certain extent. While the discriminating power and construct-irrelevant variance did not threaten the test's validity, the high level of difficulty was clearly a concern for the LCLT. Three items (item 10, 22 and 29) were found to be too challenging for the test-takers. Most of other items had relatively high difficulty levels which were more suitable for higher-proficiency candidates. There was a shortage of easy items targeting lower-proficiency test-takers. For that reason, these test-takers' ability levels might not have been reliably measured.

### **The generalization inference**

*The LCLT scores were reliable.*

While the item reliability for the LCLT was high (.97), the person reliability (.75) was not high enough for a high-stakes test. This result indicates that we can place confidence on the hierarchy of difficulty for the LCLT items. However, the order of test-takers ability according to their results on the test was not as reliable. The generalization inference was therefore not fully supported.

### **The explanation inference**

*Test-takers' performances on the LCLT are reflective of the targeted listening construct.*

This inference was not supported since both of its warrants were refuted. Specifically, the think-aloud data indicated that only five out of 16 subskills in the test construct were actually covered. The test-takers' performance was affected by poor sound quality, the multiple-choice format, the number of speakers in a conversation, the order of items, the speech rate, and the lexical overlap between the input and the test questions. The comparability study between the LCLT and IELTS revealed that although the LCLT and the IELTS Listening test were found to be similar in terms of construct, the test-takers' scores on both tests did not correlate well with each other. In fact, the correlation was low, implying that the LCLT was far from comparable with an internationally recognized test of English proficiency.

Figure 7.1 recapitulates the validity argument with evidence for each warrant. The validity argument for the LCLT was not successfully supported since the generalization inference and most importantly, the explanation inference were not backed by the research findings.

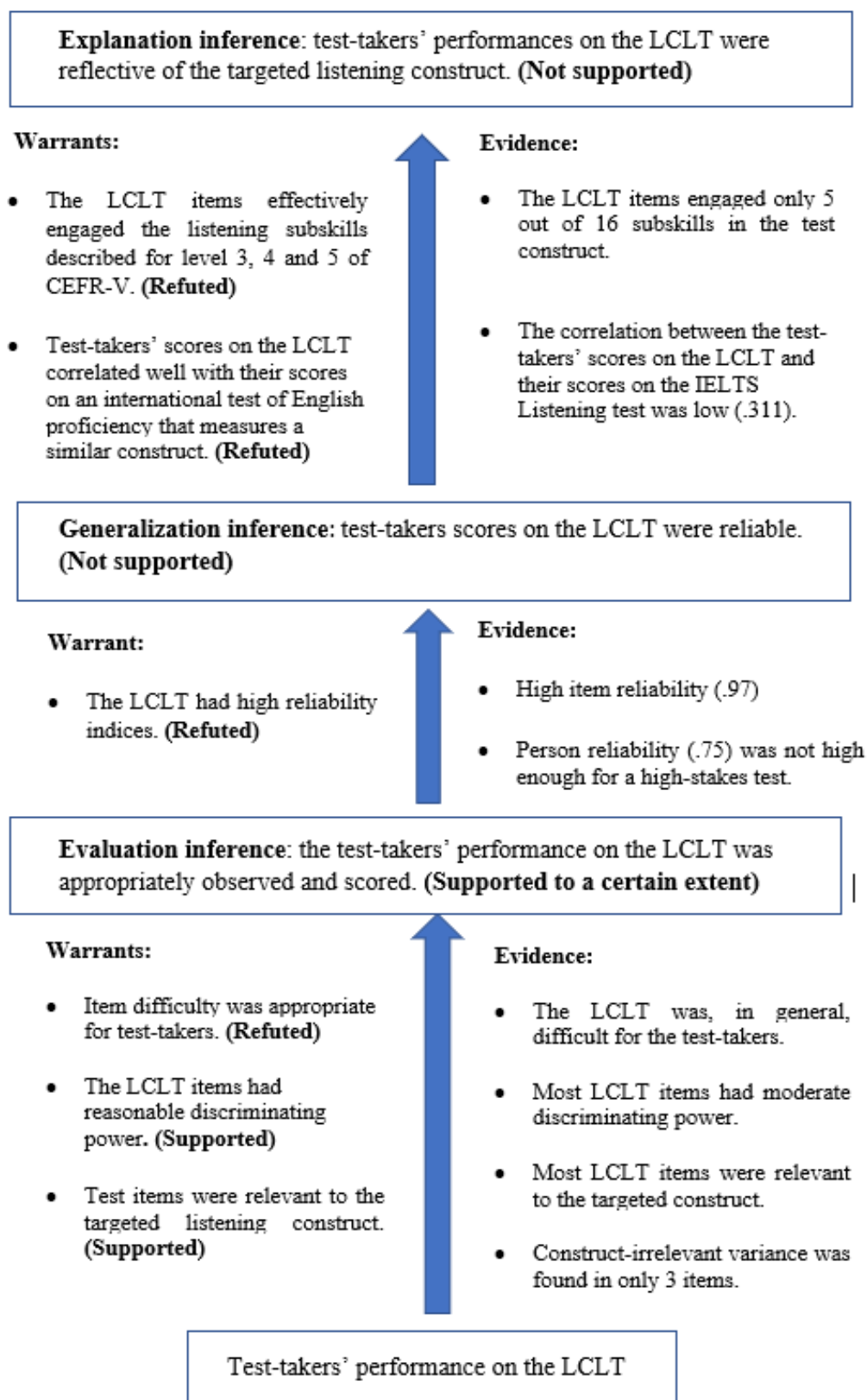


Figure 7.1. Argument-based validity for the LCLT

The results showed in Figure 7.1 are absolutely crucial for the development of future LCLTs. By unveiling the sources of invalidity in this particular test, this research highlights a number of major flaws that should be addressed in the future. The following section discusses in detail both the positive features and noteworthy problems with the LCLT. It also proposes a number of ways in which these problems can be tackled.

## **7.2. Positive features of the LCLT**

Despite the fact that the validity argument for the LCLT was not successfully supported, the test still possesses a number of good qualities that deserve acknowledgment. It should be emphasized that the LCLT was developed by a group of English teachers who were new to the role of item writers, especially for high-stakes listening tests. Due to a lack of language testing expertise at the host university, these teachers were mobilized to take on this new role without substantial training in item writing. However, they managed to create the auditory input and design a relatively decent listening test with 35 multiple-choice items. It was apparent that great effort had gone into both input text selection and item construction.

In terms of content, the LCLT incorporated a variety of topics ranging from daily life, entertainment and places to geography and art. Although some socio-linguistic knowledge might have been necessary for one or two items, the input texts, in general, did not require special topical knowledge. In the study on test-takers' cognitive processes, none of the participants reported any issue with the topics of the listening texts. Apart from the diverse topics, the LCLT also included a good variety of discourse types such as daily conversations, interviews, tour-guide talks or lectures. It should be noted that a large number of individual listening texts was required for the LCLT. Specifically, eight short conversations were needed for Part 1, three long conversations for Part 2 and three long talks for Part 3. It was indeed a challenge to put together in one test a total number of 14 individual texts which had different contents and featured only general background knowledge.

Since effective multiple-choice items are notoriously difficult and time-consuming to write, the developers of the LCLT must have undergone a laborious process constructing 35 MCQs for the test. Although the design of many items was flawed when it came to details, most questions in the test appeared to follow the general rules for writing MCQs. In all the questions, the stems were easy to understand, grammatically correct and had reasonable length. The four alternatives for each question was usually short, clear and grammatically consistent. This way of designing test items was suitable for a listening test since it minimized the amount of reading, thus allowing test-takers to focus on the spoken texts. For certain items (item 10, 11, 13, 19, 20, 32), the distractors did a very good job. The think-aloud data showed that while answering these items, the participants had to consider two or three alternatives and think critically to decide on their final answers. Another positive feature of the LCLT was the absence of ambiguity. For each question, there was always one correct answer and no overlapping content was seen between the alternatives.

In relation to the listening construct, the items that targeted the ability to understand explicit and implicit details were relatively well written. For the latter, in particular, the think-aloud protocols showed that these items did motivate the listeners to use both bottom-up and top-down processes to figure out the answers (Section 5.4.1.2). The ways in which these items were constructed successfully engaged high-level thinking, pushing the test-takers to go beyond word level to comprehend the wider meaning of the texts.

As evidenced by the statistical analysis of test scores, the LCLT had reasonable discriminating power and most of the test items were relevant to the listening construct. This result, along with the positive features discussed previously, is encouraging evidence that local English teachers at the host university are capable of developing good listening tests. Their skills and expertise will be greatly improved as they gain more experience on the job, receive further training and are informed by validation efforts similar to this present research.



### **7.3. Problems with the LCLT and proposed solutions**

The findings of this research bring to light four major problems with the LCLT. They are construct underrepresentation, the inappropriate difficulty levels, construct-irrelevant variance, and unsatisfactory person reliability. This section discusses these issues in light of the research findings and the existing literature. For each of these issues, practical solutions are proposed.

#### **7.3.1. Poor construct coverage**

The most serious problem with the LCLT is construct under-representation. The test represented its target construct very narrowly by focusing primarily on the comprehension of details. The test-takers' ability to make inferences and understand discourse function were also engaged but to a much lesser extent. The subskills that were not observed in the participants' verbal reports included:

- the ability to understand main ideas in various contexts;
- the ability to understand new words in context;
- the ability to understand idiomatic expressions;
- the ability to follow detailed directions or instructions;
- the ability to make inferences about the purpose and the implication of the speaker.

The under-representation of the targeted listening construct, as Messick (1989) emphasized, is a major threat to the LCLT validity. It adversely affects score interpretations and can mislead score users (Kane, 2002).

Although these findings are disappointing, it should be pointed out that the underrepresentation of the targeted listening construct is not unique to the LCLT. In a limited number of studies on the cognitive validity of the IELTS Listening subtest, evidence of construct underrepresentation has been found. In his investigation of the cognitive validity of the lecture-based questions in the IELTS Listening paper, Field (2005) pointed out that the

focus of testing was very much bottom-up and test questions mainly engaged shallow processing at local level. Geranpayeh and Taylor (2008) also described the IELTS Listening test items as focusing on explicit and easily accessible information. More recently, in an attempt to validate the IELTS Listening test, Aryadoust (2013) found that the test underrepresented the listening construct “by tapping only two listening subskills: the ability to understand explicit stated information and to make paraphrases” (p. 218). It is unfortunate that research on how well tests of listening actually engaged their intended construct is scarce. No other published research has been found to focus strongly on the construct representation of a particular listening test. The findings of this research and of previous studies on the IELTS Listening test imply that construct underrepresentation is a major issue not only for locally created listening tests but also for the ones which are internationally recognized and professionally made. The focus on details and local comprehension of the input text appear to be a shared pattern.

Improving the construct coverage should be the top priority in the development of future LCLTs. Better construct coverage for the LCLT can be achieved with the four following practical solutions. First and foremost, it is vitally important to make sure that the test developers fully understand the construct and are consistent when designing test items. For any item writer, in-depth understanding of the targeted listening construct is a prerequisite for the success of the item writing process. However, the construct of a test can never be fully comprehensive and is always open to interpretation (Clapham, 1996; Fulcher & Davidson, 2007). In case of the LCLT, some of the 16 listening subskills that the test aimed to measure can be quite confusing for the item writers. For example, it will be challenging to write an item that assesses the ability to understand the main ideas of propositionally and linguistically complex speech on familiar topics. The interpretation of the term “propositionally and linguistically complex speech” might vary from one person to

another. Unless an agreement is reached among the test developers about what the term means in the particular context of the LCLT, it will be difficult to select suitable input texts and write items to represent the subskill. It might not be a coincidence that the descriptions for all the five listening subskills represented by the LCLT were clear and easy to understand. However, this was not always the case for the unrepresented subskills, for example:

- Listening and identify the main points and structure of propositionally and linguistically complex speech including technical discussions;
- Listen and recognize details in propositionally and linguistically complex speech.

Detailed guidelines and discussion between test developers are necessary to ensure that everyone in the item production team shares the same understanding of the construct and how it should be represented by test items.

Secondly, it is important to ensure that there is variety in the test format. The decision to choose the multiple-choice format for the LCLT was likely based on practical considerations such as objective, low-cost and efficient scoring (Haladyna, 2004). Nonetheless, the test method should first lend itself to the construct targeted by the test tasks (Haladyna & Rodriguez, 2013). The multiple-choice format is unfortunately not ideal to represent all aspects of the LCLT's ambitious construct. The findings from the study on test-takers' cognitive processes (see Chapter 4, section 4.4.2) revealed that while answering LCLT multiple-choice items, the test-takers mainly adopted a procedure of matching the words they could hear from the recording with words written in the questions, thereby largely relying on the lexical overlap. This attention to word level was at the expense of in-depth, global understanding of the input text (Field, 2005), which was apparently an important part of the LCLT's target construct. Since the multiple-choice format alone is not good enough to represent well the LCLT construct, other test methods should be taken into consideration.

Field (2013) suggests a number of other useful formats for listening tasks such as multiple matching, gapped summary or true/false/not mentioned. However, he also cautions that test developers should have sound reasons for choosing a particular format for a particular listening task. The impacts of test methods on test-takers' performance will be further discussed in Section 6.4.

Thirdly, giving the test-takers the second chance to listen to the aural input would be useful. The results of the study on the test-takers' cognitive processes suggested that their performance on many test items was adversely affected by the rapid shift of attention between reading, listening, recalling information, and word matching. While working on the items that required relatively complicated inference making, some candidates decided to delay the answer since they found the items too cognitively demanding to give immediate answers. In this case, test-takers can really benefit from a second chance to listen to the input. Field (2009, 2011) suggests that test-takers carry out different types of processing when given the opportunity to listen twice. On the first listening, they are able to locate relevant evidence in the input text and possibly making initial links with one or more items. On the second listening, test-takers will have the chance to review, confirm or even change the initial answer. In the LCLT construct, many listening subskills involve high-level processing which was not observed in the test-takers' verbal reports. If these test takers could listen to the input the second time, they would be likely to display high-level thinking as they review, confirm or change their initial position of the information they hear. With the double-play format, the test-takers will possibly be engaged in a wider range of listening behaviours targeted by the LCLT. Although this format has often been criticized for being more time-consuming and less authentic (Ruhm et al., 2016), a number of empirical studies have showed that repeated listening had positive effects on test-takers' performance. For example, Otsuka (2004) found that repeated listening reduced item difficulty. Cervantes and Gainer (1992) found that the

double play format had more effect with top-down (gist) tasks than with bottom-up (detail) tasks. In terms of authenticity, single-play tasks can become less relevant these days when, given changes in technology, people can listen to online materials such as recorded lectures, radio programs, as often as they wish (Murray, 2007). Despite much debate surrounding the issue of whether test-takers should be allowed to listen once or twice (Geranpayeh & Talor, 2008), from a cognitive point of view, the double play format is still worth considering for a listening test.

The fourth suggestion concerns review and revision, a very important step in the test development procedure. It is in this step that reviewers have the opportunity to examine the extent to which test items represent the target construct (Green, 2017). Since there is no guarantee that the subskill that the test developer has in mind when designing an item is actually the subskill employed by the test-takers, test trialling and feedback from reviewers are essential. For the LCLT, the feedback on construct representation is best provided by those who are familiar with the construct and the test specifications. It is important that the reviewers are prepared to be completely objective and honest. At the same time, the test developers should also be able to accept the fact that their test items might not adequately represent the construct and that they need to revise, drop or write new items (Green, 2017). Since construct underrepresentation was found to be a serious problem with the LCLT, the work of the reviewers should go beyond superficial aspects such as spelling, test layout, instructions and focus strongly on the evaluation of test items against the target construct. Honest feedback and thorough revision will, most probably, lead to better construct coverage for the LCLT.

### **7.3.2. Inappropriate difficulty levels**

The statistical analysis of the LCLT (Chapter 3) showed that the test items were difficult for the test-takers. The comparability study between the LCLT and IELTS (Chapter

5) found that the same group of English majors from the host university scored significantly higher in the IELTS Listening test than in the LCLT. Taken together, these findings confirm that the LCLT items had very high levels of difficulty.

The study on the cognitive processes of LCLT test-takers (Chapter 5) revealed that certain text and task-related factors contributed to increasing the cognitive demands of many LCLT items. These factors include the number of speakers, order of items, speech rate, and lexical overlap. It is important to note that these factors were found to make some LCLT items more cognitively demanding in ways that did not really represent the construct. Specifically, including three speakers in a conversation and sequencing items out of the chronological order put unnecessary pressure on the listeners' working memory. The sudden increase in speech rate and the lack of pauses in the auditory input caused anxiety. The lexical overlap between the listening text and the response options increased the listeners' dependence on the written input.

It is crucial that the test developers have better control over item difficulty for future LCLTs. This means the difficulty levels of test items should reflect the targeted construct instead of being unreasonably determined by text and task-related factors. Since the LCLT construct does not target the listening ability in settings which feature a large number of speakers, having two speakers in the recordings for part 1 and part 2 of the LCLT would be reasonable. That way, listeners will have more time to normalise to the different speakers' voices (Elliott & Wilson, 2013) and really focus on processing the input. Scrambling the order of multiple-choice items to make them more challenging for the test-takers was also not justified by the LCLT construct. Therefore, it should be ensured that, in future LCLTs, the order of items follows the order of information in the input text. This match between the test items and the input text will help to avoid confusion for the test-takers and reduce the impact of memory. In terms of speech rate, it is important to ensure comparable speed for the

conversations or talks that belong to the same section of the test. The number of pauses should also be taken into consideration.

Although the suggestions above can be useful in mitigating the unnecessary impacts on the LCLT item difficulty of text and task-variables, they are certainly not enough. Geranpayeh (2013) once noted that item writers usually think of an intended difficulty level when they construct the test items but they can never be sure how difficult their items are until they are taken by real candidates. Therefore, trialling the test on a representative sample of test-takers is an essential step in the test production procedure. It is also suggested that statistical analysis should be carried out on the candidates' scores and be followed by content analysis from experts (Geranpayeh, 2013). At this stage, items with acceptable statistical values will be retained while items whose statistical values fall outside the acceptable range will need to be revised or rewritten for a new trial. The whole process of writing, trialling, analyzing, and revising test items takes much time and effort; however, without such a process, it is not possible to control and balance item difficulty in a listening test. The evidence of the LCLT's inappropriate difficulty levels implied that the trialling, analyzing, and revising steps might have been skipped or done half-heartedly. It is therefore important to ensure that the LCLT developers carry out these steps with great care so as to better control the difficulty of future tests.

### **7.3.3. Substantial construct-irrelevant variance**

The study on the statistical characteristics of the LCLT scores indicated that construct-irrelevant variance (CIV) was not a major issue for the LCLT (see Section 3.2.6). However, the qualitative analysis of the test-takers' cognitive processes revealed two sources of CIV that had considerable impacts on test-takers' performance. The poor sound quality of the last recording used in the test resulted in construct-irrelevant difficulty (Messick, 1995) for the last five items. The multiple-choice format allowed the test-takers to use test-taking

strategies such as random guessing or deductive reasoning based solely on the written information. These test-specific behaviours indicate possible flaws in the test design that permit the test-takers to avoid normal processes (Field, 2013). This means that even when they come up with the correct answers, it is difficult to be confident that their answers are reliable indicators of their listening ability. Rather, it only means that the test-takers are lucky to be rewarded by correct guesses (Wu, 1998) or that they successfully apply the approach they were trained in to tackle multiple-choice tasks (Gude & Duckworth, 2008). With evidence that the LCLT test-takers made their decisions on the basis of the written information only, it was their reading ability that affected their performance on a listening test (Field, 2009).

So as to reduce the presence of construct-irrelevant variance, it is generally desirable to employ a variety of response methods (Elliott & Wilson, 2013). This suggestion was given before as a good way in which the construct coverage for the LCLT could be improved (see Section 6.2.1). However, it is discussed here as a solution that helps the LCLT avoid construct-irrelevant variance caused by the over-reliance on the multiple-choice format. Shohamy (1997) notes that all response methods have their advantages as well as drawbacks. By using a blend of response methods, the validity of the entire test will not be overly threatened by the drawbacks of any individual method (Elliott & Wilson, 2013). The dependence on blind guessing and on the written text will be lessened if only a certain percentage of the LCLT items takes the multiple-choice format.

#### **7.3.4. Unsatisfactory person reliability**

Rasch analysis on the LCLT scores showed that the person reliability for this test was .75 which was acceptable (see Section 3.2.6). With this result, one can be relatively certain of the hierarchy of the test-takers' ability in that the high-scorers will be likely to score high on another LCLT and low-scorers will be likely to still score low. Having an acceptable



person reliability index was good news for the LCLT. That said, Linacre (2012) emphasizes that person reliability should be 0.8 for a high-stakes test. He also noted that if this value for a test is less than 0.8, the test needs more items. Therefore, an obvious solution to increase reliability for future LCLT is to increase the number of items in the test, which should reduce error of measurement and provide better construct coverage. It is also important to consider practical constraints such as the time limit of the test, the time that test developers have to write test items, or examinee fatigue.

The decision of how many items to be added to the future LCLTs cannot be intuitively made and the quality of added items is also difficult to determine. These issues can only be addressed by trialling the test on a representative sample size. The test developers may try adding a certain number of items to the LCLT and give the longer test to a representative group of candidates. The statistical analysis of test scores will reveal whether the addition of new items can help to achieve the reliability needed for the LCLT. In order to examine item quality, statistical and content analysis will both be useful.

#### **7.4. Issues in the localization of English proficiency tests**

In recent years, the concept of test localization has attracted more attention since the current use of major international tests of language proficiency, with little or no evidence of local appropriateness of these tests, is no longer supportable (Su et al., 2020). In Asia, the use of locally produced English proficiency tests has become a growing trend in English assessment. During the first decade of the twenty first century, there was an increasing number of English proficiency tests being developed by different institutions in Asia such as the General English Proficiency Test (GEPT) developed by the Language Training & Testing Centre in Taiwan; the Test of English for Academic Purposes (TEAP) developed by the Eiken foundation in Japan or the College English Test (CET) in China (Nguyen, 2020). These tests became the motivation for Vietnamese testing experts to create a new assessment tool

that is valid, reliable and at the same time more affordable and suitable for local context. However, in stark contrast to that rosy picture, the findings of this research suggest that test localization is easier said than done. Although the research focuses only on a particular component of a localized test, it opens up a part of a bigger picture in which a number of drawbacks are noticed in the localization of English proficiency tests in Vietnam. The following discussion of these drawbacks might be useful for teachers, researchers and policy-makers with an interest in language test localization.

#### **7.4.1. Poor construct conceptualization**

As mentioned in the Introduction chapter, the Vietnamese localized tests of English proficiency are developed on the basis of the six-level framework of foreign language proficiency for Vietnamese. Adapted from the CEFR, this framework retains most of the content of the CEFR (Nguyen & Hamid, 2015), without closely examining the deep-rooted nature of English teaching and learning in Vietnam (Hoang, 2010; Nguyen et al., 2018). Despite such criticisms, the Vietnam Foreign Language Framework has still been treated as the single operationalized construct for the locally created English proficiency tests which serve multiple high-stakes purposes. However, that one-size-fits-all approach is highly problematic since it does not take into account the language abilities required for each purpose. In other words, the Vietnamese localized English tests are developed without a clearly defined target language use domain which is “a specific setting outside the test itself that requires the test taker to perform language use tasks” (Bachman & Palmer, 2010, p.60). The investigation of the LCLT’s construct in this research reveals how poorly the listening construct is conceptualized.

The primary purpose for which the Vietnamese localized English proficiency tests are developed is to assess the English proficiency of English teachers in Vietnam (Le, 2017). While this purpose clearly indicates that the tests should target the English proficiency

needed for a particular professional field which is teaching, the listening component in these tests is developed based on a framework for general English proficiency. The sixteen listening subskills listed in the test specifications bear very little relevance to the working environment of English teachers in Vietnam. The listening abilities that English teachers need to demonstrate in order to succeed in their job are not reflected in a general test construct which contains subskills such as:

- understanding the main points of short, clear announcements;
- understanding discourse function and the purpose of the speaker;
- understanding details of discussion around him/her when they are not clearly articulated;
- understanding new words in context.

Without the identification of the target language use domain, the assessment of listening skill for English teachers in Vietnam clearly lacks authenticity and specificity, the two important features of a language test for specific purposes (Douglas, 2000). Therefore, the results of such assessment can scarcely be a predictor of the teachers' language performance at work. This drawback can only be addressed when the listening subtest is developed with a full consideration of the specific features of the teaching field. It is essential that the test developers investigate the extent to which Vietnamese teachers of English make use of their listening skill, focusing on important issues such as:

- Which particular aspects of teaching require the teachers' listening skill?
- Which particular aspects of listening skill are the most important and relevant to the teachers' job?
- Are there any differences in the use of listening skill among English teachers from primary to tertiary level?

These questions can be answered by means of classroom observation, questionnaires or interviews. It is only when the teaching practices in the classroom and the teachers' voices are taken into consideration that the listening construct can be identified. Unfortunately, instead of making the effort to provide specific descriptors of the listening abilities necessary for Vietnamese teachers of English, policy-makers opted for a general framework to assess the teachers' listening skill. This is also the case for other components of the Vietnamese localized test of English proficiency. Le (2017) found that the content of the speaking and writing subtests also fails to cover the abilities needed for English teaching. Nguyen (2017) pointed out that the language content of the locally developed tests is different from the actual teaching practices.

Apart from its use in the evaluation of English teachers nation-wide, the Vietnamese locally created English tests have also been implemented as a graduation benchmark policy for university students (Le, 2017). Under this policy, the graduation threshold is level 5 of the CEFR-V (equivalent to C1-CEFR) for English majors and level 3 (equivalent to B2-CEFR) for non-English major students. However, the CEFR-V does not take into account the teaching curricular used at higher education institutions across the country. In fact, each individual institution has its own right to decide what to teach based on the general guidelines from the Ministry of Education and Training (Pham & Bui, 2019). Bui and Nguyen (2016) pointed out that the most popular English language programs offered by universities and colleges in Vietnam for non-English majors consist of basic English communication grammar, reading, and basic English communication. Therefore, the use of a test that measures level 3 to level 5 of the Vietnam Foreign Language Framework as graduation requirements is argued to be over-ambitious and unachievable for a majority of university students (Nguyen et al., 2018; Van Huy & Hamid, 2015). With regard to the assessment of

listening, the students who are exposed to only basic grammar, reading, and English communication will be disadvantaged by a test that measures high-level subskills such as:

- understanding enough to follow detailed directions or instructions;
- understanding the main ideas of a conversation even when content is not straightforward and clearly structured;
- listening and making inferences about the speaker's attitude and opinion through his/her tone and language use.

To better serve its purpose as a graduation requirement, the development of the Vietnamese localized English proficiency test should be informed by research on the curricular and teaching practices at higher education institutions in the country.

Thus far, it is apparent that the construct targeted by the listening subtest of the Vietnamese localized English proficiency test is appropriate for neither teacher evaluation nor a graduation benchmark. The conceptualization of that construct is simply based on a pre-determined framework of language proficiency instead of the purposes for which the subtest is developed. This is also the case for the reading, speaking and writing components of the test (Le, 2017). As a result, many Vietnamese teachers of English and university students have been assessed against irrelevant standards, which has led to social consequences and negative wash-back effects (Le, 2020; Pham & Bui, 2019). The use of only one test which is based on a borrowed language proficiency framework in the Vietnamese context for different high-stakes purposes does not seem to work. The problem with this one-size-fits-all approach is that the characteristics of English teachers and university students in Vietnam have been largely ignored. Questions of trustworthiness arise when a localized test of English proficiency is not found to reflect the working, learning environment and the unique characteristics of the local test-takers.

#### 7.4.2. Substandard technical quality

Solorzano (2008) cautioned that there are several risks to the use of a standardized high-stakes language test in public education. One of them is the risk that the rush to develop and administer these tests to a large number of test-takers loses sight of important principles of test construction. The disappointing findings about the technical quality of the LCLT clearly illustrate this point. At the surface level, the test looked like a decent multiple-choice test which had clear instructions and was objectively scored. However, the in-depth investigation revealed major flaws underneath that surface. The ambitious listening construct was poorly represented by the test items and the quality of the recordings was not consistent throughout the test. In addition, the difficulty of certain items was unreasonably affected by a number of text and task-related factors. The unsatisfactory person reliability index suggested that the LCLT needed more items. Apart from that, there are still other important aspects that this research was not able to cover. They include but are not limited to the lexical profiles of the listening texts or the design of multiple-choice items (Green, 2017). The investigation of the LCLT's validity, though still insufficient, suggests that not much effort was put into the sophisticated technical aspects of developing a listening test. This sounds the alarm about the quality of the reading, speaking, and writing components of the English proficiency tests developed by the host university. It also casts doubt on the technical quality of the tests developed by the other universities that were chosen to develop the localized English proficiency test.

When language tests are used for high-stakes purposes, their technical quality is a fundamental concern (Solorzano, 2008) since it affects the interpretation of test scores. Kopriva (2000) emphasizes that when high-stakes decisions are made using test results that may be flawed, the test-takers' futures hang in the balance. In Vietnam, the futures of thousands of English teachers and university students depend on their scores on the locally

created tests of English proficiency. If efforts are not made to ensure that these tests are of good quality, unfair decisions will be made about the test-takers. Taking the test will be a waste of their time and money. This research uncovers only a small part of the reality regarding the actual quality of the Vietnamese locally developed English proficiency tests. What has been found is enough to raise concerns about the validity of these tests and the impacts they have on the test-takers' career and future.

#### **7.4.3. Lack of quality control**

The fact that the Vietnamese localized English proficiency tests are developed by eight universities across the country has made quality assurance extremely taxing. Although they are leading universities in Vietnam in terms of foreign language education (Toan, 2013), there is no guarantee that the tests they develop are comparable in terms of quality. In this research, the LCLT was found to have poor validity even when it was developed by one of the most prestigious universities in Vietnam in the area of English education. That result implies that there is little ground for optimism about the quality of the tests developed by the other seven chosen universities. Le (2017) is concerned that the inconsistencies among the different institutions responsible for developing the tests might result in discrepancies in the actual proficiency levels of two candidates achieving the same results. Therefore, a comprehensive procedure for quality management is needed to make sure that the different versions of the Vietnamese localized test of English proficiency created by the eight chosen universities are of comparable quality. The least that will need to be done next will have to be test equating, so that the eight versions of the same test that are developed by eight independent universities can be credibly argued to be capturing similar levels of language competence. Unfortunately, this issue has not been emphasized in both legal documents and academic research.

Concerning the quality of a language test, validity is by far a must-have criterion (Bachman & Palmer, 1996). However, validation research on different versions of the Vietnamese localized test of English proficiency is still very limited. Without more validation studies like this one, equating different versions of the test will be a meaningless endeavour. The lack of published validation works in both domestic and international platforms can be explained by the three following reasons. First, there might have been little interest among researchers in validating the Vietnamese localized tests of English proficiency. Second, validation research might not have been supported by the universities that are responsible for developing the tests. Since this line of research requires access to confidential data such as the actual tests and test-takers' results, it is obviously impossible without the permission from these universities. Third, validation research might have already been done at each individual institution but the research results have not gone public. However, since the Vietnamese localized test of English proficiency is a high-stakes test with considerable social impact, the public should be informed of its validity. Therefore, validation research should be encouraged, supported and published at least in accessible domestic journals and reports. Other localized tests of English proficiency in Asia such as Taiwan's General Test of English Proficiency (GEPT) and the Japan's EIKEN test are good examples to consider. These tests provide grants for validation studies and welcome international testing experts to participate or lead the projects. These studies are updated on their websites and in searchable research reports. Similar support for validation research should be an important value of the Vietnamese localized test of English proficiency. The public, especially the test-takers whose futures depend on the test results deserve information on the validity and other aspects of the test's quality. More importantly, investigating the quality of the tests that have been administered is vital for the improvement of the tests that will be conducted in the future.



#### 7.4.4. Negative social consequences

Although this research does not attempt to investigate the social consequences of the localized English proficiency tests in Vietnam, there is an indirect link between the research results and the tests' negative impacts as revealed by a number of previous studies. The poor construct conceptualization and questionable quality of the tests partly explains why they have not been well received by local teachers of English and university students.

Since the very start of the National Foreign Languages Project 2020, English teachers in Vietnam have been put under the spotlight. When their test results were revealed, the majority of teachers failed to achieve the expected proficiency levels, which attracted a lot of criticisms from the public (Le, 2017). According to these results, a great number of English teachers had poor listening and speaking skills (Vietnamnet, 2016). In an attempt to raise teachers' proficiency in English, in-service language improvement courses were offered to English teachers of all educational levels (Le, 2020). After these courses, the teachers who had failed the test had to take the English proficiency test again. Nonetheless, despite intensive training, the number of teachers who met the proficiency requirement was still far from the expected target (Le, 2020). A great number of teachers felt frustrated as they had to repeatedly take the test without achieving significant progress in test scores (Le, 2020). Moreover, they had to experience the feeling of humility when being classified as unqualified teachers. This feeling negatively affected their professional identities, leading to their demotivation in trying to reach the proficiency levels which are believed to be too challenging for them (Le, 2017). Even when they are able to reach these proficiency levels, there will be no guarantee that their improved English proficiency leads to their improved classroom practices (Nguyen & Mai, 2015).

Although the English teachers' scores on the locally developed English tests have captured public attention and put them under a lot of pressure, not much has been discussed

about the quality of the tests. Given the research findings on the poor validity of the LCLT, I would argue that rather than blaming the teachers for their disappointing test performance, improvements should be made to the tests themselves. It is clear that when taking the locally created tests, the teachers' English proficiency is assessed against ambitious standards which bear very little relevance to their context and teaching practices. This partly explains why they lose the motivation to improve their proficiency in English. It is time that the teachers' context and voices are taken into consideration and reflected in the localized English tests.

For university students in Vietnam, the use of localized English proficiency tests as a graduation requirement has not been much favoured. Under Project 2020, Vietnamese students are able to choose an array of English proficiency tests developed by both domestic universities or international institutions such as the International English Language Testing System (IELTS), the Test of English for International Communication (TOEIC), the Test of English as a Foreign Language (TOEFL), Cambridge Preliminary Test (PET) and First Certificate of English (FCE) (Nguyen, 2013). Le (2017) reports that students still opt for IELTS, TOEFL, TOEIC, PET or FCE because of their recognised international credibility (Le, 2017). Ironically, although a great amount of money has been budgeted for the development of the localized English proficiency tests, they are not trusted by Vietnamese students. Unfortunately, the reality unveiled by this research, to a certain extent, suggests that many students were right to doubt the quality of the locally developed tests of English proficiency.

The use of the locally created English tests was believed to generate extrinsic motivation among teachers and students (Pham & Bui, 2019). However, in reality, the tests have yet to create the social impacts they aimed for. They were found to demotivate English teachers nation-wide and were not favoured by university students. Such consequences clearly show that a locally created test will not be well received by local test-takers if it does

not take into account the local context and test-takers' characteristics. Moreover, just because an English test is locally produced does not mean that the expectation for its quality is lowered. When that test is used for high-stakes purposes and costs the test-takers a certain amount of money, it should be of acceptable quality. The undesirable social impacts will continue unless the Vietnamese localized tests of English proficiency are thoroughly improved.

### **7.5. Concerns about the assessment of second language listening**

Beyond its meaning for a locally created listening test, this research provided a number of useful insights into the assessment of second language listening in general. Most of these insights came from the analysis of the think-aloud protocols provided by 10 English majors from the host university. The protocols shed light on not only the cognitive processes of the participants but also the factors affecting their performance. Taken together, such evidence highlighted the need to rethink the use of the multiple-choice format in listening tasks, in general. Besides, it has practical implications for developing tests of listening.

#### **7.5.1. Undesirable impacts of the multiple-choice format**

The think-aloud protocols in this research suggested that the way in which the participants processed the input was very much bottom-up since they focused their attention mainly on lexical matches rather than on wider meaning. Such behaviour was clearly promoted by the multiple-choice format. There was extensive evidence of the test-takers using the classic key word strategy to locate relevant information in the auditory input. In one of the very few studies on the effects of the multiple-choice format on the listeners' cognitive processes, Field (2005) also found that the listeners' thinking operated in the direction: written lexical input – spoken lexical input. This way, the multiple-choice format somehow changed the purpose of listening. It was no longer listening for comprehension but indeed, listening to check whether the input matched one of the ideas given to them before the

recording was played. The listeners, when given multiple-choice questions, did not listen with an open mind but with a number of expectations shaped by the questions and the response options. As evidenced by the think-aloud protocols, the correct answers, sometimes were merely the result of successful lexical match rather than true comprehension. This result is in agreement with Weir's (1993) criticism that previewing questions before listening changes the nature of the listening tasks. It confirms Yanagawa and Green's (2008) concern that previewing answer options may encourage test-takers to adopt a lexical matching strategy which must have limited validity for the testing of listening skill.

The multiple-choice format is often chosen for listening tests because it is convenient and cost-effective (Yanagawa & Green, 2008). That said, the assumption that a format should be used because of its convenience and efficiency is flawed. The issue of construct validity should be of primary importance in the decision regarding the response formats. Barta (2009) emphasizes that the unique and often stressful nature of listening calls for the application of valid response format even more. The findings of this research and that of Wu (1998) and Field (2009) suggest that the multiple-choice format might be suitable for items that target the understanding of details and bottom-up processing. However, much more research is still needed to investigate which listening subskills can be best embraced by multiple-choice items.

Since a given response method can only test part of the listening construct, a blend of response methods is believed to permit broad construct coverage (Elliott & Wilson, 2013). Testing textbooks often provide lists of testing methods that can be used to assess listening but mostly without discussing the usefulness of these methods in measuring the listening construct (Barta, 2009). In fact, the advantages and disadvantages of response formats such as true/false, multiple matching, short answer, gap filling or note-taking are usually discussed at the surface level. Very little has been known about what particular listening subskills these

formats are actually able to engage. While it is generally desirable to employ a variety of response methods to ensure fairness for test-takers and increase construct coverage for a listening test (Alderson et al., 1995), the decision of which format to use to measure a particular aspect of the intended construct has not been informed by relevant research. As a consequence, test developers might end up using a number of response formats in a listening test only to ensure variety rather than construct validity. Therefore, research that investigates the impacts of different response methods on L2 listeners' cognitive processes is sorely needed. Only with such research will the choice of response formats be well informed and truly useful for measuring a targeted listening construct.

### **7.5.2. Practical considerations for developing listening tests**

In the development of listening tests, it is important to consider the interplay between the listening subskills, test tasks and input text (Ayradoust, 2018). Except for the effects of the multiple-choice format which has been discussed previously, this research also provides insights into the effects of some text and task-related factors. Such insights lead to some practical considerations for the assessment of L2 listening.

First, the consideration of delivery speed for a listening text should go beyond the crude figure of mean speech rate. This research has supported a classic stance that a text with faster speech rate adversely affected test-takers' performance (Elliott & Wilson, 2013). It also became apparent that the lack of pauses had a detrimental effect on comprehension. Evidence from the think-aloud protocols indicated that the lack of pauses in the input text made it difficult for test-takers to identify word boundaries and keep up with the flow of speech. While most of the studies on speech rate revolve around a threshold rate at which comprehension increases or decreases (Foukle, 1968; Griffiths, 1992) or typical speech rates for different discourse modes (Tauroza & Allison, 1990), very little attention has been paid to the importance of pauses. Since pauses assist listeners in word segmentation and allow them

more time to process the input (Goldman-Eisler, 1961), it is necessary to properly incorporate them in a listening text. Elliott and Wilson (2013) suggest that both pause length and the positioning of pauses need to be taken into consideration. Since these aspects of the input text have not been researched enough, they would be interesting areas to investigate further.

Second, the appropriate number of speakers for a text is an issue that needs to be thought through. This factor is believed to affect the test-takers' performance because it brings up the issue of distinguishing between speakers and switching attention from one to another. The belief was confirmed by the evidence from the think-aloud data which showed that having three speakers in an input conversation considerably increased the cognitive load of the related test item. It was observed that the test-takers experienced trouble when they had to quickly shift attention among the three speakers. The most serious problem with the number of speakers seemed to be the extra pressure on the listeners' memory. I would venture to go a step further by arguing that this problem was worsened by the multiple-choice format. Since the listeners had to listen to a short conversation between three speakers then picked a correct answer and at the same time disconfirmed the other three options, most of them reported having difficulty matching what they heard with a particular option since they could not remember who said what. It seems apparent that the multiple-choice format doubled the cognitive demand of the item. The listeners had to switch attention not only among speakers but also among response options and they had to do that almost at the same time. In the selection of input text for a listening test, I would suggest that a maximum of two speakers is enough if the targeted construct does not involve the ability to listen to a conversation with three or more speakers. It is also advisable to have one male and female speaker so that listeners can easily distinguish the speakers' voices (Weir, 2005). In case, more than two speakers have to be involved, other aspects such as the response format, the topic or the lexical demand of the input text should be carefully considered.

Third, lexical overlap between the input text and the response options should not be overused as a strategy to manipulate item difficulty. The findings of previous research have revealed a relationship between task difficulty and the lexical overlap between words in the text and the response options of multiple-choice items (Buck & Tatsuoka, 1998; Jensen, 1997). Buck (2001) emphasizes that the overlap between the correct options and the necessary information in the input text is the best predictor of an easy items. Similarly, the overlap between the input text and the incorrect options is the best predictor of difficult items. The relationship between the lexical overlap and listening task difficulty was supported by this present research. It was also found that using lexical overlap was an ineffective strategy because it manipulated item difficulty at the expense of construct validity. This strategy encourages test-takers to rely heavily on linguistic cues and draw their attention away from a natural interpretation of the input text. Chang (2008) and Field (2005) both shared a similar concern. From a personal point of view, a certain amount of lexical overlap between the input text and the test questions is essential since it helps listeners keep track of the flow of speech. For multiple-choice questions, in particular, the strategic use of lexical overlap to adjust item difficulty should be done with care and with reference to the construct. The amount of overlap in a multiple-choice item and the degree of overlap (complete or partial) are issues that deserve more attention and further research.

The insights gained from the test-takers verbal report indicate that manipulating the speech rate, number of speakers, item order, and lexical overlap invalidly increase item difficulty. These strategies can make listening test items more cognitively demanding but because of them, test-takers might be deprived of the opportunity to show their true listening ability. At the end of the day, adjusting the difficulty of a listening task requires more effort than simply making changes to the input text or to the item design. Evidence from the present study suggests that every decision on which factor to be adjusted and how it can be adjusted

should be made in light of the targeted construct. Moreover, when one factor is changed, other related factors should also be taken into consideration. This helps to balance out the cognitive load of an item and ensure that test-takers do not have to suffer from multiple sources of difficulty at the same time.

#### **7.6. Critical reflection on the argument-based approach to validation**

Applying the argument-based approach in validating the LCLT is a valuable learning opportunity. Throughout this research project, I have gained a greater appreciation for this approach because of three particular benefits it offers researchers in test validation. The first and also the most important advantage, in my opinion, is the opportunity to state the intended interpretations and uses of test scores and let them guide the validation work. Therefore, validating the LCLT became a more meaningful process since the test was assessed against its own claims rather than a generic list of qualities that could be applied to any test. This advantage has also been acknowledged in a number of previous studies (Aryadoust, 2013; Chapelle et al., 2008; 2010; Lim, 2009; Liu 2014). It is directly connected to the second benefit, the logical determination of the types of evidence that would be required. In case of the LCLT, once the evaluation, generalization, and explanation inferences were proposed, the evidence needed to support these inferences became apparent. I was not inclined to list and seek all possible types of validity evidence. Instead, it was more productive to focus on the evidence that was most relevant to the inferences.

The third advantage of the argument-based approach is the application of the Toulmin argument model. I find this model particularly useful since it helped to effectively and logically organize the evidence needed to support each of the inferences in the validity argument for the LCLT. Apart from that, taking rebuttal into consideration is a crucial step that ensures the comprehensiveness of a claim. Identifying and seeking evidence for a rebuttal is absolutely necessary because sometimes, it takes only one rebuttal to cancel out all the validity evidence.



For example, in the evaluation inference for the LCLT, evidence of construct under-representation alone was enough to refute the inference despite validity evidence such as minimal construct-irrelevant variance or reasonable discriminating power (see Chapter 3).

Although the argument-based approach provides a great conceptual tool for outlining and allocating research effort, it lacks methodological guidance for the actual validation work. After stating the three inferences for the LCLT, I was left to figure out which qualitative and quantitative methods to use to elicit evidence necessary for each inference. A review of previous studies that applied the argument-based approach revealed inconsistency in terms of research methodology. For example, to elicit evidence for the evaluation inference of the TOEFL test, Chapelle et al. (2008) investigated statistical characteristics of test items. Supporting the same inference for an integrated listening-speaking task, Frost et al. (2012) used discourse analysis to examine the extent to which test-takers' performance reflected the rating scale descriptors. Although the decision of which research methods to use is largely dependent on the test in question, a common methodological guidance is still much needed. A guidance mapping out the main research methods that could be used to investigate each inference in the interpretive argument will be beneficial for researchers, especially those who attempt to apply the argument-based approach for the first time.

## **Chapter 8. Conclusion**

This final chapter offers reflections on various aspects of this research project. First, it highlights the contribution and practical implications of the research. What then follows is the acknowledgement of the limitations. The chapter ends with suggestions for future research and the researcher's final thoughts on the entire journey.

### **8.1. Contributions of the research**

This entire research provides the first in-depth assessment of a locally created listening test in Vietnam. The results shed light on the poor validity of the test and highlight the major issues that need to be addressed in the future. As was clear from the research findings, the LCLT suffered from both construct under-representation and construct-irrelevant variance, the two major threats to validity (Messick, 1989). These findings suggest that the test scores did not effectively reflect the test-takers' listening ability. The use of such a test for high-stakes purposes such as screening English teachers or setting English proficiency standards for university graduates is, therefore, highly questionable. The LCLT was also found to have high levels of difficulty and for at least eight items, their cognitive demands were increased in ways that did not engage the targeted listening subskills. Although reliability was not a major issue for the test, the research findings still indicated that the LCLT did not have enough items to cover its ambitious construct.

Taken together, the research findings raised an alarm about the technical quality of the LCLT. Since listening is, arguably, the most difficult skill to assess (Field, 2013), the flaws in listening tests, in general, and in the LCLT, in particular, are to be expected. Rather than criticizing the test developers, this research has provided them with an opportunity to reflect on their work and see where the problems were. It also proposed practical solutions for the betterment of future LCLTs.

Although this research focused only on a particular locally created listening test, its findings suggested that the localization of language tests, in general, is easier said than done. In theory, the goal of test localization is to develop tests which are suitable for local test-takers while being aligned to international standards. In reality, the LCLT was found to be neither appropriate for local test-takers nor compatible with international standards. This failure was due to poor construct conceptualization, substandard technical quality, and lack of effective quality control measures. There is still a great deal to be done before Vietnam can establish an effective mechanism for the localization of an English proficiency test. It is important that the main value of test localization is embraced by both test-developers and policy-makers. The gap between theory and practice can only be narrowed when clear guidance is in place and effective actions are taken to ensure the quality of the localized English tests.

This study represented a very limited number of efforts that investigated test-takers' cognitive behaviour in response to the multiple-choice format in a listening test. It further confirmed Field's (2009) finding that the multiple-choice format encouraged the lexical matching strategy and distracted test-takers from the broader meaning of the text. It also supported Weir's (2005) concern that previewing the multiple-choice questions before listening changed the nature of the listening task. The use of the multiple-choice format was found to be counter-productive in items that targeted the ability to identify implicit details and to make inferences. This finding, to a modest extent, addressed the lack of research on the interaction between listening subskills and response formats in general.

Four other factors were found to affect the cognitive demand of items in a listening test. They are the number of speakers, order of items, speech rate, and lexical overlap. These factors are known to have certain impacts on item difficulty (Freedle & Kostin, 1996, 1999; Buck & Tatsuoka, 1998); however, very little has been revealed about how they actually

affect test-takers' thinking process. The investigation of the verbal data indicated that the number of speakers, the order of items and the lexical overlap between the input text and the multiple-choice questions were potential threats to the validity of a listening test. These factors were found to add an extra burden on the test-takers' memory and distract them from a normal listening process. In terms of speech rate, while previous studies mainly focused on the crude figure of articulation speed such as the number of words per minute (Griffiths, 1992; Tazora & Allison, 1990), this study found that the frequency of pauses was another important dimension of speech rate that affected the cognitive demand of test items.

## **8.2. Practical implications of the research**

In light of its findings, this research has practical implications for the development of listening tests. These implications focus on two important aspects, the provision of suitable texts and the design of test items.

With regard to the listening text, sound quality must remain the top priority. This might sound too obvious; however, since that turned out to be an issue for the LCLT, it should still be emphasized. For a listening test that puts together a number of separate sound files, the editing stage is important to make sure that the sound quality is consistently good throughout the test. Buck (2001) provided detailed instructions on ensuring good sound quality. The number of speakers in a conversation should be carefully considered. A maximum of two speakers (one male, one female) is ideal since it does not put too much pressure on the listeners' working memory. The determination of speech rate should go beyond a crude figure of words per minute and take pauses into consideration. The number, the length, and the position of pauses in an input text are important factors that can affect test-takers' performance.

Five practical suggestions are given for the design of listening tests. First, a variety of response formats is needed. Although the multiple-choice format offers undeniable

advantages, it is not ideal in measuring a construct that include both low and high-level listening subskills. Moreover, the opportunity to preview the questions and all the response options may change the nature of listening. Second, multiple-choice items are suitable for items that target bottom-up processing and the ability to understand details. The research findings suggest that this format might not be very useful in assessing the ability to understand the speakers' attitudes or opinions. Third, for listening tests which give test-takers only one chance to listen, test items should be organized in chronological order. Scrambling item order confuses listeners and unreasonably increases item difficulty. Fourth, the double-play format might be useful if a test aims to measure many high-level listening subskills. Finally, in order to control the difficulty of a test item, it is important to consider not only the listening subskill that the item targets but also text- and task-related factors. Although adjusting factors such as vocabulary, speech rate or number of speakers can increase or decrease item difficulty, this strategy should be treated with caution. Since text and task-related factors do interact with each other, it is necessary to adjust one factor in relation with the other ones. Moreover, this manipulative strategy will not be useful if changes in the textual features of the input or in the task design do not reflect the intended construct.

### **8.3. Limitations of the research**

A major limitation of this research project is the absence of the extrapolation and utilization inference in the argument-based validation framework for the LCLT. The localized test of English proficiency that the LCLT is a part of has been used in Vietnam as requirement for graduation, post-graduate studies and job application. Therefore, the investigation of the uses and the consequences of test scores is of great importance. Not including them in this research has reduced the comprehensiveness of the argument-based validation framework.

A small sample participant size is another drawback of this research. For a high-stakes test like the LCLT, a large sample size would be ideal for validation. However, all the three studies in this research were conducted with relatively modest populations. It should also be acknowledged that the participants recruited for (a) the qualitative study on the test-takers' cognitive processes and (b) the comparability study between the LCLT and IELTS were not representative of the test-taking population. They were all English majors with relatively high levels of English proficiency, whereas actual candidates of the locally created English tests vary greatly in terms of their backgrounds and proficiency levels. This apparently limits the generalizability of the findings of these two studies.

The main source of weakness in the quantitative study on LCLT scores was the fact that the two groups of test-takers did not experience the same testing condition. In order to create a sample size large enough for Rasch analysis, a group of English majors from the host university was added to the original group of test-takers. This group benefited from an anxiety-free testing condition since they were aware that the results of the test would not have any impact on their study. Thus, procedural fairness was not ensured since test-takers were not treated equally and consistently (Kane, 2010).

In the qualitative study on the test-takers' cognitive processes, the participants were invited to take the LCLT in a non-test condition. Therefore, the cognitive processes they reported did not fully reflect the processes undergone by test-takers in a real test event. The performances of 10 participants in this research were clearly not affected by factors like time pressure and test anxiety. They took the test in a nice, quiet room where they could sit close to the speaker in order to get the best possible sound quality. The non-test condition was a major trade-off in the design of this study. The insights into the listeners' cognitive processes were prioritized at the expense of authenticity.

Finally, in the comparability study between the LCLT and an IELTS Listening test, the test-takers' familiarity with IELTS might have helped them perform better on this test. While the LCLT was new to the group of 66 test-takers from the host university, the listening component of IELTS was, most probably, more familiar to many of them. Since test familiarity can give test-takers an unfair advantage, it might have contributed to the significant difference in the test-takers' performances on the two tests.

#### **8.4. Suggestions for future research**

A natural progression of this project is to validate the extrapolation and utilization inference of the LCLT. From the perspective of both test-takers and test-score users, a study on the usefulness of this test is of great importance. Apart from the listening sub-test, further studies need to be carried out to validate other components of the English proficiency tests developed by the host university. It is recommended that these future validation studies be conducted on a larger sample size and a more representative test-taking population. This study also calls for validation efforts from other universities that have been responsible to the development and administration of localized English proficiency tests in Vietnam.

The assessment of listening will also be a fruitful area for future work. Qualitative studies on the interaction between listening subskills and response formats are sorely needed. While testing textbooks often provide a list of response formats for a listening test, impacts of test methods on construct validity are rarely discussed. The decision of which response format(s) to use to best engage a particular listening subskill should not be made on the basis of the popularity, convenience or efficiency of a certain format. Instead, it should be informed by relevant research which has not been widely available. Therefore, more research in this area will definitely benefit test developers.

More research is needed to fully understand the extent to which factors that affect listening task difficulty might increase or decrease the cognitive demand of test items. More

importantly, it is necessary to examine whether these factors help to better measure the listening construct or add unwanted dimensions to the listening tasks. Insights into these issues will help test developers design good items and effectively control test difficulty.

### **8.5. Concluding comments**

I embarked on this research project with a focus on a particular locally created test. However, I have completed it with findings which will be meaningful for not only test developers in Vietnam but also, item writers, researchers and teachers who share a common interest in assessing L2 listening. Venturing into this under-researched area was challenging but exciting at the same time. I often got lost but gained new experience in every step of the way.

The poor validity of the LCLT, as evidenced by this research, will be disappointing for the test-developers at the host university. It should be noted that the ultimate purpose of this research was not to criticize the test-developers but to help them create better tests in the future. As a former test item writer, I understand how important it is to be able to identify flaws in a test and have a plan to address them. Therefore, once the problems with the LCLT were identified, I proposed solutions that can help item writers who might be struggling with the development of new LCLTs.

The most exciting part of this research, for me, was the journey into the test-takers' minds. The think-aloud protocols unveiled cognitive behaviours that I had never envisaged. Each participant brought a unique perspective to the study. Analysing their thoughts piqued my curiosity in the nature of listening and the challenges of assessing it. I can see a future in which I, as a researcher and item writer, will keep exploring this elusive, unobservable language skill.



## References

- About AEC*. (n.d). Retrieved from <http://investasean.asean.org/index.php/page/view/asean-economic-community/view/670/newsid/755/about-aec.html>
- Alderson, J. C., Clamham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.
- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, 3(1), 3-30.
- Allami, H., & Aghajari, J. (2014). Pragmatic knowledge assessment in listening sections of IELTS tests. *Theory and Practice in Language Studies*, 4(2), 322-340.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15. <https://doi.org/10.1146/annurev.ps.37.020186.000245>
- Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). W.H. Freeman.
- Aryadoust, V. (2009). Mapping the Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions*, 23(1), 1192–1193.
- Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Cambridge Scholars Publishing.
- Aryadoust, V. (2018). Taxonomies of listening skills. In J. I. Liantas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–8). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118784235.eelt0577>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests (Vol. 1)*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice* (1st ed.). Oxford University Press.
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge TOEFL comparability study*. Cambridge University Press.
- Bacon, S. M. (1992). Phases of listening to authentic input in Spanish: A descriptive study. *Foreign Language Annals*, 25(4), 317–333.
- Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, 22(1), 1145–1146.
- Baghaei, P., & Amrahi, N. (2011). Validation of a multiple-choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research*, 2(5), 1052–1060.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
- Baker, F.B., & Kim, S.H. (2017). *The basics of item response theory using R*. Springer.
- Bardovi-Harlig, K. (1999). Exploring the interlanguage of interlanguage pragmatics: A research agenda for acquisitional pragmatics. *Language Learning*, 49(4), 677–713.

- Barta, E. (2009). Analysis of listening comprehension assessment task. In G. Szabo, J. Horvath, & M. Nikolov (Eds.), *UPRT 2009: Empirical Studies in English Applied Linguistics* (pp. 65-80). Lingua Franca Csoport.
- Berne, J. E. (1992, August 9-13). *The role of text type, assessment task and target language experience in L2 listening comprehension assessment*. [Conference presentation]. Annual meetings of the American Association of Applied Linguistics and the American Association of Teachers of Spanish and Portuguese, Cancun, Mexico.
- Bloomfield, A., Wayland, S. C., Rhoades, E. K., Blodgett, A., Linck, J. A., & Ross, S. J. (2010). *What makes listening difficult? Factors affecting second language listening comprehension*. University of Maryland.
- Bond, T., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences* (1st ed.). Lawrence Erlbaum Associates.
- Bond, T., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Routledge.
- Bowles, M. A. (2010). Concurrent verbal reports in second language acquisition research. *Annual Review of Applied Linguistics*, 30, 111–127.
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171–191.
- Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, 19(4), 369–394.
- Brindley, G., & Wigglesworth, G. (1997). *Access: Issues in language test design and delivery*. National Centre for English Language Teaching and Research, Macquarie University.

- Brown, H. D. (2006). *Principles of language learning and teaching* (5th ed.). Pearson Education.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653–675.
- Brown, J. D., Davis, J. M., Takahashi, C., & Makamura, K. (2012). *Upper-level Eiken examinations: linking, validating, and predicting TOEFL iBT scores at advanced proficiency Eiken levels*. Society for Testing English Proficiency.  
<https://www.eiken.or.jp/eiken/group/result/pdf/eiken-toeflibt-report.pdf>
- Brunfaut, T. (2016). Assessing listening. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 97-112). De Gruyter Mouton.
- Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141-168.
- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67–91.
- Buck, G. (2001). *Assessing Listening*. Cambridge University Press.
- Buck, G., Tatsuoka, K., Kostin, I., & Phelps, M. (1997). The subskills of listening: rule-space analysis of a multiple-choice test of second language listening comprehension. In A. Huhta, V. Kohonen, L. Kurki-Suono, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 589-624). University of Jyväskylä.
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15(2), 119-157.
- Bui, T. T. N., & Nguyen, H. T. M. (2016). Standardizing English for educational and socio-economic betterment? A critical analysis of English language policy reforms in

- Vietnam. In R. Kirkpatrick (Ed.), *English language education policy in Asia* (pp. 363–388). Springer.
- Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller, Jr (Ed.), *Issues in language testing research* (pp. 333-342). Newbury House.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.
- Carroll, J. B. (1972). Defining language comprehension: Some speculations. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge* (pp. 1-29). Winston & Sons.
- Celce-Murcia, M., Brinton, D. M., Snow, M. A., & Bohlke, D. (2013). *Teaching English as a Second or Foreign Language* (4th ed). Cengage Learning.
- Cervantes, R., & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, *26*(4), 767-770.
- Chang, A. C. (2008). Listening strategies of L1 learners with varied test tasks. *TESL Canada Journal*, *25*(2), 1-26.
- Chang, A. C. S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System*, *41*(3), 575–586.
- Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-17). John Wiley & Sons.
- Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education Journal*, 12(2), 68-82.
- Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, 37(4), 544-555.
- Cheung, H. (2012). Academic perspective from Taiwan. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalization of language education policy* (pp. 224-230). Multilingual Matters.
- Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, 26(2), 345–374.
- Circular No. 23/2017/TT-BGDĐT. (2017, September 9). Retrieved from <https://thuvienphapluat.vn/van-ban/giao-duc/Thong-tu-23-2017-TT-BGDĐT-thi-danh-gia-nang-luc-ngoai-ngu-theo-Khung-nang-luc-ngoai-ngu-6-bac-363395.aspx>
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge University Press.
- Cook, M., & Liddicoat, A. J. (2002). The development of comprehension in interlanguage pragmatics: The case of request strategies in English. *Australian Review of Applied Linguistics*, 25(1), 19-39.

- Council of Europe. (2001). *Common European framework for languages: Learning, teaching, assessment*. Cambridge University Press.
- Cronbach, L. J. (1971). Validity. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-597). American Council on Education.
- Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement - Measuring achievement over a decade* (pp. 99-108). Jossey-Bass.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.
- Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Lawrence Erlbaum Associates.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
- Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education*, 3, 265-285.
- Cutler, A., & Clifton, C. (1999). Comprehending spoken language: a blueprint of the listener. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 123-166). Oxford University Press.
- Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795-845). Lawrence Erlbaum Associates.
- Davidson, F., & Bachman, L. F. (1990). The Cambridge-TOEFL comparability study: An example of the cross-national comparison of language tests. *AILA Review*, 7, 24-45.

*Decision No. 1400/QĐ-TTg.* (2008, September 30). Retrieved from

[http://vanban.chinhphu.vn/portal/page/portal/chinhphu/hethongvanban?class\\_id=1&\\_page=2&mode=detail&document\\_id=78437](http://vanban.chinhphu.vn/portal/page/portal/chinhphu/hethongvanban?class_id=1&_page=2&mode=detail&document_id=78437)

*Decision No. 729/QĐ-BGDĐT.* (2015, March 11). Retrieved from

<https://thuvienphapluat.vn/van-ban/Giao-duc/Quyét-dinh-729-QĐ-BGDĐT-2015-de-thi-danh-gia-nang-luc-su-dung-tieng-Anh-tu-bac-3-den-bac-5-267956.aspx>

*Decree No. 01/2014/TT-BGDĐT.* (2014, January 24). Retrieved from

<https://thuvienphapluat.vn/van-ban/giao-duc/Thong-tu-01-2014-TT-BGDĐT-Khung-nang-luc-ngoai-ngu-6-bac-Viet-Nam-220349.aspx>

*Decree No. 15/2014/TT-BGDĐT.* (2014, May 15). Retrieved from

<https://thuvienphapluat.vn/van-ban/giao-duc/Thong-tu-15-2014-TT-BGDĐT-Quy-che-Dao-tao-trinh-do-thac-si-229824.aspx>

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

Dudzic, D. L., & Nguyen, Q. T. N. (2015). Vietnam: Building English competency in preparation for ASEAN 2015. In R. Stroupe & K. Kimura (Eds.), *ASEAN Integration and the role of English language teaching* (pp. 41-70). IELTS.

Dunlea, J., Spiby, R., Nguyen, T. N. Q., Nguyen, T. Q. Y., Nguyen, T. M. H., Nguyen, T. P.

T., Thai, H. L. T., & Bui, T. S. (2018). *Aptis-VSTEP comparability study:*

*Investigating the usage of two EFL tests in the context of higher education in Vietnam.* (Report No. VS/2018/001). British Council.

[https://www.britishcouncil.org/sites/default/files/aptis-vstep\\_study.pdf](https://www.britishcouncil.org/sites/default/files/aptis-vstep_study.pdf)



- Ebadi, S., Zamani, G., & Sharifi, M. M. (2014). Interpretive argument of the academic reading module of IELTS: A fusion of current research. *International Journal of Language Learning and Applied Linguistics World*, 6(4), 317–328.
- Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 152-241). Cambridge University Press.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. The MIT Press.
- ETS. (2010). *Linking TOEFL iBT scores to IELTS scores*.  
[https://www.ets.org/s/toefl/pdf/linking\\_toefl\\_ibt\\_scores\\_to\\_ielts\\_scores.pdf](https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf)
- Eykyn, L.B. (1992). *The effects of listening guides on the comprehension of authentic texts by novice learners of language* [Unpublished doctoral dissertation]. University of South Carolina.
- Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27(3), 209–226.
- Feyton, C. (1991). The power of listening ability: An overlooked dimension in language acquisition. *The Modern Language Journal*, 75(2), 173-180.
- Fidler, E. J. (1983). The reliability and validity of concurrent, retrospective, and interpretive verbal reports: An experimental study. In P. Humphreys, O. Svenson, & A. Vári (Eds.), *Advances in psychology - Vol 14* (pp. 429–440). North-Holland.
- Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal*, 57(4), 325–334.

- Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *System*, 32(3), 363–377.
- Field, J. (2005). The cognitive validity of the lecture-based question in the IELTS listening paper. *IELTS Research Reports*, 9, 17–65.
- Field, J. (2008). Revising segmentation hypotheses in first and second language listening. *System*, 36(1), 35–51.
- Field, J. (2009). *Listening in the language classroom*. Cambridge University Press.
- Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes*, 10(2), 102–112.
- Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77-151). Cambridge University Press.
- Field, J. (2015). *The effects of single and double play upon listening test outcome and cognitive processing* (Report No. ARAG/2015/003). British Council.  
[https://www.britishcouncil.org/sites/default/files/field\\_layout.pdf](https://www.britishcouncil.org/sites/default/files/field_layout.pdf)
- Flowerdew, J., & Miller, L. (2014). Dimensions of academic listening. In M. Celce-Murcia, D. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (pp. 90-103). National Geographic Learning.
- Foulke, E. (1968). Listening comprehension as a function of word rate. *Journal of Communication*, 18(3), 198-206.
- Foster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 257-287). North Holland.
- Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: Implications for construct validity* (Report No.

TOEFL-RR-56). English Testing Service.

[https://www.ets.org/research/policy\\_research\\_reports/publications/report/1996/hybm](https://www.ets.org/research/policy_research_reports/publications/report/1996/hybm)

- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's mini-talks. *Language Testing*, 16(1), 2–32.
- Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, 29(3), 345–369.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.
- Garcia, P. (2004). Pragmatic comprehension of high and low level language learners. *TESL-EJ*, 8(2). <https://eric.ed.gov/?id=EJ1068107>
- Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Lawrence Erlbaum Associates.
- Geranpayeh, A. (1994). Are score comparisons across language proficiency test batteries justified? An IELTS-TOEFL comparability study. *Edinburgh Working Papers in Applied Linguistics*, 5, 50-65.
- Geranpayeh, A. (2013). Scoring validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 242-272). Cambridge University Press.
- Geranpayeh, A., & Taylor, L. (2008). Examining listening: developments and issues in assessing second language listening. *Cambridge Research Notes*, 32, 2-5.
- Geranpayeh, A., & Taylor, L. (Eds.). (2013). *Examining listening: research and practice in assessing second language listening*. Cambridge University Press.

- Goh, C. (1997). Metacognitive awareness and second language listeners. *ELT Journal*, 51(4), 361–369.
- Goh, C. (1998). How ESL learners with different listening abilities use comprehension strategies and tactics. *Language Teaching Research*, 2(2), 124–147.
- Goh, C. (2008). Metacognitive instruction for second language listening development: Theory, practice and research implications. *RELC Journal*, 39(2), 188–213.
- Goh, C. (2014). Second language listening comprehension: Process and pedagogy. In M. Celce-Murcia, D.M. Brinton, & M.A. Snow (Eds.), *Teaching English as a second or foreign language* (pp. 72- 89). Cengage Learning.
- Goh, C., & Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *International Journal of Listening*, 29(3), 109–133.
- Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language and Speech*, 4(3), 171–174.
- Graham, S. (2003). Learner strategies and advanced level listening comprehension. *Language Learning Journal*, 28, 64-69.
- Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge University Press.
- Green, R. (2017). *Designing listening test*. Springer.
- Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, 26(2), 385–390.
- Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition*, 25(1), 135–155.

- Gude, K., & Duckworth, M. (2008). *Proficiency masterclass student book*. Oxford University Press.
- Güss, C. D. (2018). What is going through your mind? Thinking aloud as a method in cross-cultural psychology. *Frontiers in Psychology, 9*(1292), 1-11.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
- Hamavandy, M., & Kiany, G. R. (2014). A historical overview on the concept of validity in language testing. *Advances in Language and Literary Studies, 5*(4), 86–91.
- Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 31–49). Academic Press.
- Hasan, A. S. (2000). Learners' perceptions of listening comprehension problems. *Language, Culture and Curriculum, 13*(2), 137–153.
- Hemmati, F., & Ghaderi, E. (2014). The effect of four formats of multiple-choice questions on the listening comprehension of EFL learners. *Procedia - Social and Behavioral Sciences, 98*, 637–644.
- Henning, G. (1987). *A guide to language testing: development - evaluation - research*. Newbury House. <https://doi.org/10.1177/026553228700400208>
- Henrichsen, L. E. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning, 34*(3), 103–123.

- Hoang, V. V. (2010). The current situation and issues of the teaching of English in Vietnam. *Ritsumeikan Studies in Language and Culture*, 22(1), 7–18.
- Holzknicht, F., Eberharter, K., Kremmel, B., Zehentner, M., McCray, G., Konrad, E., & Spöttl, C. (2017). *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis listening test* (Report No. ARG/2017/3). British Council.  
[https://www.britishcouncil.org/sites/default/files/looking\\_into\\_listening.pdf](https://www.britishcouncil.org/sites/default/files/looking_into_listening.pdf)
- Hu, G., & McKay, S. L. (2012). English language education in East Asia: Some recent developments. *Journal of Multilingual and Multicultural Development*, 33(4), 345–362.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.
- Hulstijn, J. (2003). Connectionist models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning*, 16(5), 413-425.
- Jensen, C., Hansen, C., Green, S., & Akey, T. (1997). An investigation of item difficulty incorporating the structure of listening tests: a hierarchical linear modelling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suono, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 151-164). University of Jyväskylä.
- Jia, Y. (2013). *Justifying the use of a second language oral test as an exit test in Hong Kong: An application of assessment use argument framework* [Unpublished doctoral dissertation]. University of California.
- Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China standards of English: Challenges at macropolitical and micropolitical levels. *Language Testing in Asia*, 7, Article 1. <https://doi.org/10.1186/s40468-017-0032-5>

- Johnson, R. C., & Riazi, A. M. (2016). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing*, 32, 85-104.
- Jun, H. S. (2014). *A validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test* [Unpublished doctoral dissertation]. Iowa State University.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31-41.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135-170.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Prager Series on Higher Education.
- Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3-17.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17.
- Kasper, G. (1984). Pragmatic comprehension in learner-native speaker discourse. *Language Learning*, 34(4), 1-20.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading*. Cambridge University Press.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Council of Chief State Officers.
- Krashen, S. D. (1981). The fundamental pedagogical principle in second language teaching. *Studia Linguistica*, 35, 50-70.
- Kunnan, A. J., & Carr, N. (2017). A comparability study between the General English Proficiency Test - Advanced and the Internet-Based Test of English as a Foreign Language. *Language Testing in Asia*, 7, 1-16.
- Kurita, T. (2012). Issues in second language listening comprehension and the pedagogical implications. *Accents Asia*, 5(1), 30-44.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book*. McGraw-Hill.
- Le, H. (2011). *Developing a validity argument for the English Placement Listening* [Unpublished master's thesis]. Iowa State University.
- Le, V. C. (2017). English language education in Vietnam universities: National benchmarking in practice. In E. S. Park & B. Spolsky (Eds.), *English education at the tertiary level in Asia: From theory to practice* (pp. 283-292). Routledge.
- Le, V. H. (2017). Evaluating the suitability of the Project 2020 Test for EFL teachers in Vietnam. *TESOL Working Paper Series*, 15, 114-125.
- Li, Z. (2015). *An argument-based validation study of the English Placement Test (EPT) – focusing on the inferences of extrapolation and ramification* [Unpublished doctoral dissertation]. Iowa State University.



- Lim, G. S., & Khalifa, H. (2013). Criterion-related validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: research and practice in assessing second language listening* (pp. 303-321). Cambridge University Press.
- Linacre, J. M. (2012). *A user's guide to WINSTEPS MINISTEP Rasch-model computer program [version 3.74.0]*. <http://www.winsteps.com/index.htm>
- Littlewood, W. (1981). *Communicative language teaching: An introduction*. Cambridge University Press.
- Markham, P., & Latham, M. (1987). The influence of religion-specific background knowledge on the listening comprehension of adult second-language students. *Language Learning*, 37(2), 157–170.
- McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3(1), 31–51.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.
- Messick, S. (1990). Validity of test interpretation and use. *ETS Research Report Series*, 1990(1), 1487–1495.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Morley, J. (2001). Aural comprehension instruction: principles and practices. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed., pp. 69-85). Heinle & Heinle.
- Munby, J. (1978). *Communicative syllabus design*. Cambridge University Press.

- Murray, S (2007). Reviewing the CAE listening test. *Cambridge ESOL Research Reports*, 30, 1-28.
- Nation, I. S. P., & Newton, J. (2008). *Teaching ESL/EFL listening and speaking*. Taylor & Francis. <https://doi.org/10.4324/9780203891704>
- Newmark, L. (1981). Participatory observation: how to succeed in language learning. In H. Winitz (Ed.), *The comprehension approach to foreign language instruction*. Newbury House.
- Negishi, M., & Tono, Y. (2014, April 10-11). *An update on the CEFR-J project and its impact on the English language education in Japan*. [Conference presentation]. The 5<sup>th</sup> International Conference of the Association of Language Testers in Europe (ALTE), Paris, France.
- Nghi, P. T., & London, J. D. (2010). The higher education reform agenda: A vision for 2020. In G. Harman, M. Hayden, & P. T. Nghi (Eds.), *Reforming higher education in Vietnam: Challenges and priorities* (pp. 51–64). Springer Netherlands.
- Nguyen, T. (2017, March 26-29). *Vietnam's National Foreign Language 2020 Project after 9 years: A difficult stage* [Conference presentation]. Asian Conference on Education & International Development, Kobe, Japan.
- Nguyen, V. H., & Hamid, M. O. (2015). Educational policy borrowing in a globalized world: A case study of Common European Framework of Reference for languages in a Vietnamese University. *English Teaching: Practice & Critique*, 14(1), 60–74.
- Nguyen, V. T., & Mai, N. K. (2015). Responses to a language policy: EFL teachers' voices. *The European Journal of Social and Behavioural Sciences*, 13(2), 1830–1841.
- Nguyen, H. T. M., Nguyen, H. T., Nguyen, V. H., & Nguyen, T. T. T. (2018). 12 Local challenges to global needs in English language education in Vietnam: The perspective

- of language policy and planning. In P. C. G. Lian, C. Chua, K. Taylor-Leech, & C. Williams (Eds.), *Un(intended) language planning in a globalising world: Multiple levels of players at work* (pp. 214–233). De Gruyter.
- Nissan, S., DeVincenzi, F., & Tang, K. L. (1995). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. *ETS Research Report Series, 1995(2)*, i–42.
- Nord, J. R. (1980). Developing listening fluency before speaking: An alternative paradigm. *System, 8(1)*, 1–22.
- Nunan, D. (1997). Listening in language learning. *The Language Teacher, 21(9)*, 47-51.
- Nunan, D. (1998). Approaches to teaching listening in the language classroom. *Proceedings of the 1997 Korea TESOL Conference, Korea*, 1-10.
- O’Sullivan, B. (2012). Assessment issues in languages for specific purposes. *The Modern Language Journal, 96*, 71–88.
- O’Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O’Sullivan (Ed.), *Language testing: theories and practice* (pp. 13-32). Palgrave Macmillan.
- Oakeshott-Taylor, J. (1977). Information redundancy, and listening comprehension. In R. Dirven (Ed.), *Listening comprehension in foreign language teaching*. Scriptor. *Official Document No. 538/QLCL-QLVBCC*. (2019, May 20). Retrieved from <https://thuvienphapluat.vn/cong-van/giao-duc/Cong-van-538-QLCL-QLVBCC-2019-danh-sach-don-vi-du-dieu-kien-to-chuc-thi-cap-chung-chi-ngoai-ngu-414562.aspx>
- Otsuka, K. (2004, August 19-20). *How to determine the optimal number of listening opportunities for listening comprehension tests among Japanese high school learners*

- of English* [Conference presentation]. The 9<sup>th</sup> Conference of PanPacific Association of Applied Linguistics. Chonan, Korea.
- Pham, N. T., & Bui, L. T. P (2019). An exploration of students' voices on the English graduation benchmark policy across Northern, Central and Southern Vietnam. *Language Testing in Asia*, 9, Article 15. <https://doi.org/10.1186/s40468-019-0091-x>
- Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific Region. *LEARN Journal: Language Education and Acquisition Research Network*, 12(1), 12–18.
- Richards, J. C. (1983). Listening comprehension: approach, design, procedure. *TESOL Quarterly*, 17(2), 219–240.
- Rost, M. (1990). *Listening in language learning*. Longman.
- Rost, M. (1994). On-line summaries as representations of lecture understanding. In J. Flowerdew (Ed.), *Academic listening: research perspectives* (pp. 219-238). Cambridge University Press.
- Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-528). Lawrence Erlbaum Associates.
- Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, 78(2), 199–221.
- Ruhm, R., Leitner-Jones, C., Kulmhofer, A., Kiefer, T., Mlakar, H., & Itzlinger-Bruneforth, U. (2016). Playing the recording once or twice: Effects on listening test performances. *International Journal of Listening*, 30(1), 67-83. <http://dx.doi.org/10.1080/10904018.2015.1104252>
- Sadighi, F., & Zare, S. (2006). Is listening comprehension influenced by the background

- knowledge of the learners? A case study of Iranian EFL learners. *Linguistics Journal*, 1(3), 110-126.
- Sawaki, Y., Kim, H-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, 6(3), 190-209.
- Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal*, 78(2), 179–189.
- Shin, S. K. (2006). Construct validity of listening test items: A verbal protocol study. *English Teaching*, 61(3), 293-305.
- Shohamy, E. (1997). Testing methods, testing consequences: are they ethical? Are they fair?. *Language Testing*, 14(3), 340-349.
- Smith Jr, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2, 281-311.
- Snellings, P., Van Gelderen, A., & De Glopper, K. (2004). Validating a test of second language written lexical retrieval: a new measure of fluency in written language production. *Language Testing*, 21(2), 174–201.
- Solorzarno, W. R. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research*, 78(2). 260-329.
- Su, L. I., Weir, C. J., & Wu, J. R. W. (Eds.). (2020). *English language proficiency testing in Asia - A new paradigm bridging global and local context*. Routledge.  
<https://doi.org/10.4324/9781351254021>
- Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, 11(1), 90–105.

- Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 10(2), 89–101.
- Thompson, I., (1995). Assessment of second/foreign language listening comprehension. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 31-58). Dominie Press.
- Toan, V. (2013, March 27). English teaching in Vietnam: teacher ‘re-education’. *Tuoi tre news*. <http://tuoitrenews.vn/education/8231/english-teaching-in-vietnam-teacher-reeducation>
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.
- Underwood, M. (1989). *Teaching listening*. Longman.
- Ur, P. (1984). *Teaching listening comprehension*. Cambridge University Press.
- Vandergrift, L. (1999). Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal*, 53(3), 168–176.
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40(3), 191–210.
- Vandergrift, L. (2011). Second language listening: presage, process, product and pedagogy. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (2nd ed., pp. 455-471). Routledge.
- Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.
- Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson test of English academic: Building an assessment use argument. *Language Testing*, 29(4), 603–619.
- Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.

- Weir, C. J. (2005). *Language testing and validation*. Palgrave MacMillan.
- Weir, C., Chan, S. H. C., & Nakatsuhara, F. (2013). *Examining the criterion - related validity of the GEPT- Advanced reading and writing tests: Comparing GEPT with IELTS and real-life academic performance* (Report No. LTTC-GEPT RG-01). Language Training and Testing Centre. <https://www.ltcc.ntu.edu.tw/ltcc-gept-grants/RReport/RG01.pdf>
- Weir, C. J., & O'Sullivan, B. (2011). *Test development and validation*. Palgrave MacMillan.
- Winke, P. M., & Lim, H. (2014). *Effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation* (Report No. 2014/3). IELTS Research Report Series. <https://www.ielts.org/teaching-and-research/research-reports/online-series-2014-3>
- Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policies & Practice*, 17(2), 117-132.
- Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15(1), 21-44.
- Wu, R., Yeh, H., Dunlea, J., & Spiby, R. (2016). *Aptis-GEPT test comparison study: looking at two tests from multiple perspectives using the socio-cognitive model* (Report No. VS/2016/002). British Council. <https://www.britishcouncil.org/exam/aptis/research/publications/validation/aptis-gept-test-comparison-study>
- Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 2316–2335). Springer.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test.

*System*, 36(1), 107–122.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, 32(2), 199–225.



## Appendix A. Request for permission to access data (Vietnamese)

**ĐƠN XIN TIẾP CẬN DỮ LIỆU PHỤC VỤ NGHIÊN CỨU**

Kính gửi: PGS-TSK [REDACTED]

Đồng kính gửi: TS. [REDACTED]

Tên tôi là: Trần Thị Ngọc Diệp, giảng viên Khoa tiếng Anh, [REDACTED]

Hiện tại tôi đang làm nghiên cứu sinh chuyên ngành Ngôn ngữ học ứng dụng tại Đại học Victoria, Wellington, New Zealand. Nghiên cứu của tôi tập trung vào việc xác trị bài thi đánh giá năng lực tiếng Anh ([REDACTED]). Trong quá trình thực hiện nghiên cứu này, tôi rất mong sự hỗ trợ của Trường [REDACTED] để có thể được tiếp cận với các dữ liệu sau đây:

- Các bài thi [REDACTED] các kỳ thi đánh giá năng lực ngoại ngữ do [REDACTED] Nội tổ chức và thực hiện;
- Bản ghi âm phần thi nói của bà [REDACTED] trong các kỳ thi đánh giá năng lực ngoại ngữ do Trường [REDACTED] tổ chức và thực hiện;
- Kết quả các bài [REDACTED] trong các kỳ thi đánh giá năng lực ngoại ngữ do Trường Đại [REDACTED] tổ chức và thực hiện.

Những dữ liệu trên đóng vai trò quyết định đối với nghiên cứu của tôi. Vì vậy, tôi rất mong sẽ có được sự chấp thuận từ phía Trường [REDACTED] cho phép tôi được sử dụng các dữ liệu này. Tôi xin cam đoan, tất cả những dữ liệu tôi được phép tiếp cận sẽ chỉ được sử dụng trong nghiên cứu của tôi, không sử dụng cho bất kỳ mục đích nào khác. Đồng thời, tôi cũng sẽ bảo đảm nghiêm ngặt tính bảo mật đối với những thông tin mà tôi được tiếp cận. Ngoài tôi và giáo viên hướng dẫn thứ nhất, PGS, TS. Peter Gu, không ai khác có thể được tiếp cận với những dữ liệu trên. Tôi tin tưởng rằng những kết quả từ nghiên cứu của tôi sẽ giúp ích cho công tác kiểm tra đánh giá của Trường [REDACTED] cũng như Đề án Ngoại ngữ Quốc gia 2020 trong tương lai.

Tôi xin trân trọng cảm ơn!

Nếu Quý trường yêu cầu các thông tin cụ thể hơn, xin hãy liên lạc với tôi và giáo viên hướng dẫn qua các địa chỉ sau:

PGS-TS. Peter Gu	Trần Thị Ngọc Diệp
Email: Peter.Gu@vuw.ac.nz	Email: Diep.Tran@vuw.ac.nz

Wellington, ngày 12 tháng 1 năm 2017

Ý kiến Trường [REDACTED]

Ký tên


[REDACTED]

Trần Thị Ngọc Diệp


[REDACTED]

[REDACTED]


## Appendix B. Request for permission to access data (English)

 <p>TE WHARE WĀNANGA O TE ĀHURANGA O TE IKA A MĀUI VICTORIA UNIVERSITY OF WELLINGTON</p>	<p>SCHOOL OF LINGUISTICS &amp; APPLIED LANGUAGE STUDIES TE KURA TĀTARI REO Level 2, Von Zedlitz Building, Kelburn Parade, PO Box 600 Wellington 6140 Phone +64 4 463 5800 Fax +64 4 463 5804 Email lals@vuw.ac.nz Website www.victoria.ac.nz/lals</p>
<p>Date: 12<sup>th</sup> January, 2017</p>	
<p>To: <input type="text"/></p>	
<p><b>REQUEST FOR PERMISSION TO ACCESS DATA FOR RESEARCH PURPOSE</b></p>	
<p>I am Tran Thi Ngoc Diep, a <u>Ph.D</u> candidate in the School of Linguistics and Applied Language Studies, Victoria University of Wellington. The proposed topic of my research is: <u>Validating the Listening Component of the</u> <input type="text"/>. In order to conduct and complete my research, it is important that I have the permission from <input type="text"/> university to access related data. I am hereby seeking your consent to let me gain access to the following data:</p>	
<ul style="list-style-type: none"> <li>- <input type="text"/> test papers used in English proficiency examinations conducted by <input type="text"/></li> <li>- The recordings of the speaking component of <input type="text"/></li> <li>- The results of <input type="text"/> tests conducted by <input type="text"/></li> </ul>	
<p>The aforementioned data will play a decisive role in my research. Therefore, I would highly appreciate it if <input type="text"/> grants me the access to such data. I hereby guarantee that all the data that I can gain access to will be used only in my research and for research purposes. Once collected, all the data will be kept <u>strictly confidential</u> and available only to me and my primary supervisor, Associate Professor, <u>Dr. Peter Gu</u>. I have faith that the outcomes of my research will be useful for <input type="text"/> as well as Project 2020 in the future.</p>	
<p>Should you require any further information, please contact me and my primary supervisor. Our contact details are as follows:</p>	
<p>Associate Professor, <u>Dr. Peter Gu</u> Email: <a href="mailto:Peter.Gu@vuw.ac.nz">Peter.Gu@vuw.ac.nz</a></p>	
<p>Tran <u>Thi Ngoc Diep</u> Email: <a href="mailto:Diep.Tran@vuw.ac.nz">Diep.Tran@vuw.ac.nz</a> Telephone number: +64 220 942 379</p>	
<p>Your permission to conduct this study will be greatly appreciated.</p>	
<p><input type="text"/></p>	<p>Signature</p>
<p>Approved</p>	
<p>Signed and sealed by the President of <input type="text"/> University</p>	<p>Tran <u>Thi Ngoc Diep</u></p>

## Appendix C. Information sheet (The study on LCLT scores)

 <p><b>Validating the Listening component of the [redacted] of English Proficiency</b></p> <p><b>INFORMATION SHEET FOR PARTICIPANTS</b></p> <p>Thank you for your interest in this project. Please read this information before deciding whether or not to take part. If you decide to participate, thank you. If you decide not to take part, thank you for considering my request.</p> <p><b>Who am I?</b></p> <p>My name is Tran Thi Ngoc Diep and I am a Doctoral student in Applied Linguistics at Victoria University of Wellington. This research project is work towards my thesis entitled "Validating the Listening Component of the V [redacted]"</p> <p><b>What is the aim of the project?</b></p> <p>This project investigates various aspects [redacted] Listening test based on your test scores. This research has been approved by the Victoria University of Wellington Human Ethics Committee with the approval number 0000024025.</p> <p><b>How can you help?</b></p> <p>If you agree to take part, you will take a [redacted] ng test which will take at most 40 minutes. You can leave the test at any time, without giving a reason. You can also choose not to hand in your test papers without giving a reason. You can withdraw from the study by contacting me at any point before May 1<sup>st</sup>, 2017. If you withdraw, your test papers will be destroyed or returned to you.</p> <p><b>What will happen to your test papers and results?</b></p> <p>This research is confidential. This means that the researchers named below will be aware of your identity but the research data will be aggregated and your identity will not be disclosed in any reports, presentations, or public documentation. However, you should be aware that in small projects your identity might be obvious to others in your community.</p> <p>Only my primary supervisor, Associate Professor. Peter Gu and I will get access to your test papers and results. The test papers and results will be kept securely and destroyed 01 year after the research ends.</p> <p><b>What will the project produce?</b></p> <p>The information from my research will be used in my PhD report.</p> <p><b>If you accept this invitation, what are your rights as a research participant?</b></p> <p>You do not have to accept this invitation if you don't want to. If you do decide to participate, you have the right to:</p> <ul style="list-style-type: none"> <li>• choose not to answer any question, or not to take any part of the test;</li> <li>• leave any time during the test;</li> <li>• withdraw from the study before May 1<sup>st</sup>, 2017;</li> <li>• ask any questions about the study at any time;</li> <li>• receive the results of the test;</li> <li>• be able to read any reports of this research by emailing the researcher to request a copy.</li> </ul> <p><b>If you have any questions or problems, who can you contact?</b></p>	<p>If you have any questions, either now or in the future, please feel free to contact either:</p> <table border="0"> <tr> <td><b>Student:</b></td> <td><b>Supervisor:</b></td> </tr> <tr> <td>Name: Tran Thi Ngoc Diep</td> <td>Name: Peter Gu</td> </tr> <tr> <td>University email address: Diep.Tran@vuw.ac.nz</td> <td>Role: primary supervisor</td> </tr> <tr> <td></td> <td>School: Linguistics and Applied Language Studies</td> </tr> <tr> <td></td> <td>Phone: +64 4 4635606</td> </tr> <tr> <td></td> <td>Peter.Gu@vuw.ac.nz</td> </tr> </table>	<b>Student:</b>	<b>Supervisor:</b>	Name: Tran Thi Ngoc Diep	Name: Peter Gu	University email address: Diep.Tran@vuw.ac.nz	Role: primary supervisor		School: Linguistics and Applied Language Studies		Phone: +64 4 4635606		Peter.Gu@vuw.ac.nz
<b>Student:</b>	<b>Supervisor:</b>												
Name: Tran Thi Ngoc Diep	Name: Peter Gu												
University email address: Diep.Tran@vuw.ac.nz	Role: primary supervisor												
	School: Linguistics and Applied Language Studies												
	Phone: +64 4 4635606												
	Peter.Gu@vuw.ac.nz												

## Appendix D. Consent form (The study on LCLT scores)



TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI  
**VICTORIA**  
 UNIVERSITY OF WELLINGTON

**Validating the Listening component of the [REDACTED]  
 Test of English Proficiency**

**CONSENT TO TAKE [REDACTED] LISTENING TEST**  
 This consent form will be held for 4 years.


Researcher: Tran Thi Ngoc Diep, School of Linguistics and Applied Language Studies, Victoria University of Wellington

- I have read the information Sheet and the project has been explained to me. My questions have been answered to my satisfaction. I understand that I can ask further questions at any time.
- I agree to take [REDACTED] Listening test.

I understand that:

- I may withdraw from this study at any point before May 1<sup>st</sup>, 2017, without giving any reason, and any information that I have provided will be returned to me or destroyed.
- The information I have provided will be destroyed 01 year after the research is finished.
- The results will be used for a PhD report and a summary of the results may be used in academic reports and/or presented at conferences.
- My name will not be used in reports, nor will any information that would identify me.

I would like to receive the results of the [REDACTED] listening test      Yes  No

Signature of participant: 

Name of participant: [REDACTED]


Date: 10/09/2017

Contact details: [REDACTED]

## Appendix E. Information sheet (The study on test-takers' cognitive processes)

<div data-bbox="392 309 612 412" data-label="Image"> </div> <div data-bbox="336 468 668 508" data-label="Section-Header"> <h3>Validating the Listening Component of the</h3> </div> <div data-bbox="301 517 715 568" data-label="Text"> <p>[Redacted]</p> </div> <div data-bbox="379 591 625 618" data-label="Section-Header"> <h4>INFORMATION SHEET FOR PARTICIPANTS</h4> </div> <div data-bbox="229 651 775 745" data-label="Text"> <p>Thank you for your interest in this project. Please read this information before deciding whether or not to take part. If you decide to participate, thank you. If you decide not to take part, thank you for considering my request.</p> </div> <div data-bbox="229 779 301 808" data-label="Section-Header"> <h4>Who am I?</h4> </div> <div data-bbox="229 826 775 931" data-label="Text"> <p>My name is Tran Thi Ngoc Diep and I am a Doctoral student in Applied Linguistics at Victoria University of Wellington. This research project is work towards my thesis entitled "Validating the Listening component of the [Redacted]"</p> </div> <div data-bbox="229 943 416 974" data-label="Section-Header"> <h4>What is the aim of the project?</h4> </div> <div data-bbox="229 990 775 1131" data-label="Text"> <p>This project aims at checking whether the abilities that [Redacted] est claims to measure are actually engaged by test-takers in test task performance. This research has been approved by the Victoria University of Wellington Human Ethics Committee with the approval number 0000024025.</p> </div> <div data-bbox="229 1146 347 1176" data-label="Section-Header"> <h4>How can you help?</h4> </div> <div data-bbox="229 1191 775 1447" data-label="Text"> <p>If you agree to take part, you will take [Redacted] listening test and verbalise in as much detail as possible what exactly you will be thinking as you take the test. You will be asked to verbalize how you answer the test questions while you take the Listening test. The whole process will be audio-recorded and will take at most 2 hours. You can leave the test at any time, without giving a reason. You can also choose not to hand in your test papers without giving a reason. You can withdraw from the study by contacting me at any point before May 1<sup>st</sup>, 2017. If you withdraw, your test papers and recordings will be destroyed or returned to you.</p> </div> <div data-bbox="229 1458 617 1491" data-label="Section-Header"> <h4>What will happen to your test papers, test results and recordings?</h4> </div> <div data-bbox="229 1505 775 1650" data-label="Text"> <p>This research is confidential. This means that the researchers named below will be aware of your identity but the research data will be aggregated and your identity will not be disclosed in any reports, presentations, or public documentation. However, you should be aware that in small projects your identity might be obvious to others in your community.</p> </div> <div data-bbox="229 1664 775 1760" data-label="Text"> <p>Only my primary supervisor, Associate Professor. Peter Gu and I will get access to your test paper, test results and recordings. The test papers, test results and recordings will be kept securely and destroyed 01 year after the research ends.</p> </div> <div data-bbox="229 1792 416 1825" data-label="Section-Header"> <h4>What will the project produce?</h4> </div> <div data-bbox="229 1834 601 1868" data-label="Text"> <p>The information from my research will be used in my PhD report.</p> </div>	<div data-bbox="922 275 1361 306" data-label="Section-Header"> <h4>If you accept this invitation, what are your rights as a research participant?</h4> </div> <div data-bbox="922 322 1469 387" data-label="Text"> <p>You do not have to accept this invitation if you don't want to. If you do decide to participate, you have the right to:</p> </div> <div data-bbox="922 403 1468 732" data-label="List-Group"> <ul style="list-style-type: none"> <li>• choose not to answer any question, or not to take any part of the test;</li> <li>• leave any time during the test;</li> <li>• withdraw from the study before May 1<sup>st</sup>, 2017;</li> <li>• ask any questions about the study at any time;</li> <li>• receive the results of the test;</li> <li>• receive a copy of your recordings;</li> <li>• read over and comment on a written summary of your interview;</li> <li>• agree on another name for me to use rather than your real name;</li> <li>• be able to read any reports of this research by emailing the researcher to request a copy.</li> </ul> </div> <div data-bbox="922 808 1283 840" data-label="Section-Header"> <h4>If you have any questions or problems, who can you contact?</h4> </div> <div data-bbox="922 842 1425 873" data-label="Text"> <p>If you have any questions, either now or in the future, please feel free to contact either:</p> </div> <div data-bbox="930 907 989 936" data-label="Section-Header"> <h4>Student:</h4> </div> <div data-bbox="930 969 1086 999" data-label="Text"> <p>Name: Tran Thi Ngoc Diep</p> </div> <div data-bbox="930 1030 1082 1090" data-label="Text"> <p>University email address: Diep.Tran@vuw.ac.nz</p> </div> <div data-bbox="1184 907 1264 936" data-label="Section-Header"> <h4>Supervisor:</h4> </div> <div data-bbox="1184 969 1289 996" data-label="Text"> <p>Name: Peter Gu</p> </div> <div data-bbox="1184 1030 1337 1057" data-label="Text"> <p>Role: primary supervisor</p> </div> <div data-bbox="1184 1088 1447 1149" data-label="Text"> <p>School: Linguistics and Applied Language Studies</p> </div> <div data-bbox="1184 1180 1324 1207" data-label="Text"> <p>Phone: +64 4 4635606</p> </div> <div data-bbox="1184 1240 1318 1267" data-label="Text"> <p>Peter.Gu@vuw.ac.nz</p> </div>
1	2

## Appendix F. Consent form (The study on test-takers' cognitive processes)



TE WHARE WĀNANGA O TE ŪPORA O TE IKA A MĀUI  
**VICTORIA**  
 UNIVERSITY OF WELLINGTON

**Validating the Listening component of the [REDACTED]  
 Test of English Proficiency**

**CONSENT TO TAKE A [REDACTED] LISTENING TEST AND TAKE PART IN THINK-ALOUD PROTOCOL**


This consent form will be held for 04 years.

Researcher: Tran Thi Ngoc Diep, School of Linguistics and Applied Language Studies, Victoria University of Wellington

- I have read the information sheet and the project has been explained to me. My questions have been answered to my satisfaction. I understand that I can ask further questions at any time.
- I agree to take [REDACTED] listening test and take part in a recorded think-aloud protocol.

I understand that:

- I may withdraw from this study at any point before May 1<sup>st</sup>, 2017, without giving any reason, and any information that I have provided will be returned to me or destroyed.
- The information I have provided will be destroyed 01 year after the research is finished.
- Any information I provide will be kept confidential to the researcher and the supervisor. I understand that the results will be used for an PhD report and a summary of the results may be used in academic reports and/or presented at conferences.
- My name will not be used in reports, nor will any information that would identify me.
- I would like to receive the result of the [REDACTED] test Yes  No
- I would like a transcript the think-aloud protocol: Yes  No
- I would like to receive a copy of the final report and have added my email address below. Yes  No

Signature of participant: 

Name of participant: [REDACTED]

Date: 03/09/2017

Contact details: [REDACTED]

## Appendix G. Information sheet (The comparability study between the LCLT and IELTS)



### Validating the Listening component of the [redacted] Test of English Proficiency

#### INFORMATION SHEET FOR PARTICIPANTS

Thank you for your interest in this project. Please read this information before deciding whether or not to take part. If you decide to participate, thank you. If you decide not to take part, thank you for considering my request.

#### Who am I?

My name is Tran Thi Ngoc Diep and I am a Doctoral student in Applied Linguistics at Victoria University of Wellington. This research project is work towards my thesis entitled "Validating the Listening component of the [redacted]".

#### What is the aim of the project?

This project compares various aspects of [redacted] listening test and the IELTS listening test based on students' scores. This research has been approved by the Victoria University of Wellington Human Ethics Committee with the approval number 0000024025.

#### How can you help?

If you agree to take part, you will take [redacted] listening test and an IELTS Listening test one week later. The [redacted] listening test will take about 40 minutes. The IELTS Listening test will take about 45 minutes. You can leave the test at any time, without giving a reason. You can also choose not to hand in your test papers without giving a reason. You can withdraw from the study by contacting me at any point before May 1<sup>st</sup>, 2017. If you withdraw, your test papers will be destroyed or returned to you.

#### What will happen to your test papers and results?

This research is confidential. This means that the researchers named below will be aware of your identity but the research data will be aggregated and your identity will not be disclosed in any reports, presentations, or public documentation. However, you should be aware that in small projects your identity might be obvious to others in your community.

Only my primary supervisor, Associate Professor. Peter Gu and I will get access to your test papers and results. The test papers and results will be kept securely and destroyed 01 year after the research ends.

#### What will the project produce?

The information from my research will be used in my PhD thesis.

#### If you accept this invitation, what are your rights as a research participant?

You do not have to accept this invitation if you don't want to. If you do decide to participate, you have the right to:

- choose not to answer any question, or not to take any part of the test;

- leave any time during the test;
- withdraw from the study before May 1<sup>st</sup>, 2017;
- ask any questions about the study at any time;
- receive the results of the test;
- be able to read any reports of this research by emailing the researcher to request a copy.

#### If you have any questions or problems, who can you contact?

If you have any questions, either now or in the future, please feel free to contact either:

#### Student:

Name: Tran Thi Ngoc Diep

University email address:  
Diep.Tran@vuw.ac.nz

#### Supervisor:

Name: Peter Gu


Role: primary supervisor

School: Linguistics and Applied Language  
Studies

Phone: 04 4635606

Peter.Gu@vuw.ac.nz

## Appendix H. Consent form (The comparability study between the LCLT and IELTS)



TE WHAKA WĀNANGA O TE ŌPŌKO O TE IKA A MĀUI  
**VICTORIA**  
 UNIVERSITY OF WELLINGTON

**Validating the Listening component of the [redacted]  
 Test of English Proficiency**

**CONSENT TO TAKE [redacted] and AN IELTS LISTENING TEST**


This consent form will be held for 4 years.

Researcher: Tran Thi Ngoc Diep, School of Linguistics and Applied Language Studies, Victoria University of Wellington

- I have read the Information Sheet and the project has been explained to me. My questions have been answered to my satisfaction. I understand that I can ask further questions at any time.
- I agree to take [redacted] Listening test and an IELTS Listening test afterwards.

I understand that:

- I may withdraw from this study at any point before May 1<sup>st</sup>, 2017, without giving any reason, and any information that I have provided will be returned to me or destroyed.
- The information I have provided will be destroyed 01 year after the research is finished.
- Any information I provide will be kept confidential to the researcher and the supervisor. I understand that the results will be used for a PhD report and a summary of the results may be used in academic reports and/or presented at conferences.
- My name will not be used in reports, nor will any information that would identify me.
- I would like to receive the results of the [redacted] test and the IELTS test    Yes  No

Signature of participant: 

Name of participant: [redacted]

Date: 12/09/2017

Contact details: [redacted]



## Appendix I. Detailed descriptors for listening skill in CEFR-V

### Listening to conversations between native speakers

Level	Descriptors
Level 3 (B1 – CEFR)	Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect.
Level 4 (B2 – CEFR)	Can understand much of what is said around him/her but may find it difficult to understand all the details if the speakers do not modify their language.
	Can keep up with an animated conversation between native speakers.
Level 5 (C1 – CEFR)	Can understand complex conversations between native speakers even on abstract, complex unfamiliar topics.

### Listening to lectures and presentations

Level	Descriptors
Level 3 (B1 – CEFR)	Can follow in outline straightforward short talks on familiar topics provided these are clearly articulated with familiar accent.
	Can follow a lecture or talk within his/her own field, provided the subject matter is familiar and the presentation straightforward and clearly structured.
Level 4 (B2 – CEFR)	Can follow the essentials of lectures, talks and reports and other forms of academic/professional presentation which are propositionally and linguistically complex.
Level 5 (C1 – CEFR)	Can follow most lectures, discussions and debates with relative ease.

## Listening to announcements and instructions

<b>Level</b>	<b>Descriptors</b>
Level 3 (B1 – CEFR)	Can follow simple technical information, such as operating instructions for everyday equipment.
	Can follow detailed directions (e.g: traffic directions)
Level 4 (B2 – CEFR)	Can understand announcements and messages on concrete and abstract topics spoken in standard dialect at normal speed.
Level 5 (C1 – CEFR)	Can extract and understand specific information from poor quality, audibly distorted public announcements, e.g. in a station, sports stadium, etc.

## Listening to audio media and TV (CEFR-V)

<b>Level</b>	<b>Descriptors</b>
Level 3 (B1 – CEFR)	Can understand the main points of radio news bulletin as well as interviews, new stories with visual illustrations provided speech is clearly articulated with simple language.
	Can understand the main points of the radio and TV programs provided speech is clearly articulated at relatively slow speed.
	Can understand the content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech.
Level 4 (B2 – CEFR)	Can understand most radio documentaries and most other recorded or broadcast audio material.
	Can identify the speakers' mood or tone.
	Can understand recordings in standard dialect likely to be encountered in social, professional or academic life and identify speaker viewpoints and attitudes as well as the information content.
Level 5 (C1 – CEFR)	Can understand a wide range of recorded and broadcast audio material, including some non-standard usage, and identify finer points of detail including implicit attitudes and relationships between speakers.

## Appendix J. Evaluation template for content analysis

Language competence	Item	Note
<b>Grammatical knowledge: understanding short utterances on a literal semantic level</b>		
Pre-existing knowledge of grammar, vocabulary, syntax, phonology		
Factual information such as names, numbers, prices, dates ...		
<b>Discourse knowledge: understanding longer utterances or interactive discourse between two or more speakers</b>		
Discourse markers		
Cohesion		
Foregrounding		
Rhetorical schemata		
Story grammars		
Structure of unplanned discourse		
<b>Pragmatic knowledge: understanding the function or the illocutionary force of an utterance or longer text, and interpreting the intended meaning in terms of that</b>		
Indirect meaning/hints		
Pragmatic implications		
Text-based inferences		
<b>Sociolinguistics knowledge: understanding the language of particular socio-cultural settings and interpreting utterances in terms of the context of situation</b>		
Slang		
Idiomatic expressions		
Dialects		
Cultural references		
Figures of speech (comparison, metaphor, euphemism, understatement, sarcasm...)		
Levels of formality and registers		

**Appendix K. Fit statistics of the LCLT and IELTS items**

LCLT			IELTS		
Item	Infit MNSQ	Infit ZSTD	Item	Infit MNSQ	Infit ZSTD
34	1.06	.3	38	1.02	.2
22	1.07	.3	17	1.19	1.5
29	.85	-.3	7	.98	-.2
11	1.15	.6	34	1.00	.1
17	.93	-.2	21	.97	.3
10	.92	-.3	23	.98	-.1
20	.85	-.9	15	.92	-.9
24	.94	-.5	31	1.04	.6
18	1.11	1.0	16	1.23	3.0
9	.84	-1.7	22	1.03	.4
13	1.05	.5	10	.93	-.9
16	.88	-1.2	20	.99	-.2
28	1.07	.7	29	.95	-.6
35	1.05	.6	24	1.13	1.7
19	1.02	.2	37	1.07	.9
31	.94	-.9	19	1.09	1.1
6	.89	-1.6	40	.96	-.5
33	1.05	.6	6	.97	-.3
5	.88	-1.3	30	.91	-1.0
26	.98	-.2	3	.97	-.3
32	1.32	3.0	12	.96	-.4
15	.86	-1.2	11	.97	-.3
12	1.19	1.4	14	.96	-.3
3	.98	-.1	26	.99	-.1
23	.96	-.2	28	.95	-.4
14	1.05	.3	9	.91	-.7
8	.97	-.1	25	1.03	.3
27	1.00	.1	36	1.03	.2

25	1.05	.3	8	.91	-.6
21	.99	.0	32	1.07	.5
30	1.00	.1	35	.96	-.1
7	.97	.1	5	.95	-.2
2	.98	.2	27	.94	-.2
4	1.00	.2	18	.98	.0
1	.99	.3	39	1.02	.2
			13	1.02	.2
			1	1.04	.2
			4	1.00	.2
			33	.94	0
			2	1.01	.3