Complete Research Report

Investigating the Construct Validity of the Listening Component of the PSU-TEP

การศึกษาความตรงเชิงโครงสร้างทักษะการฟังของข้อสอบ PSU-TEP

Anchana Rukthong

This research project has been financially supported by Faculty of Liberal Arts, Prince of Songkla University

Of the fiscal year 2017, Research Code: LIA600796S

Abstract

One of the main concerns over the use of the PSU-TEP listening is construct validity or the potential of the test to tap into the abilities it is aimed to test. This is because there was no validation study since the test was administered. This study, therefore, analyzed test-takers' scores from 4 versions of the test administrated in the year of 2017 to identify the number of listening sub-skills assessed by the items used. Factor analysis was run for this purpose. Then to explain the sub-skills or listening abilities that were extracted by factor analysis, verbal data collected by means of stimulated recall from 24 participants on a one-on-one basis were analyzed. The extent to which the test measured its construct was then discussed.

The results showed that 3-4 components were extracted for each part of the tests. The abilities that were tapped into by the test, as revealed in stimulated recall, involved the ability to process listening texts at the local level and at the higher/global level. While the local-level processes enable the test-takers to understand independent idea units of the listening texts, the global or higher-level processes assist them to comprehend the main point or the message that the speaker intended to deliver. In addition to cognitive processes at both levels, different types of strategies, including inferencing, elaboration, prediction, and comprehension monitoring and test-wise strategies, i.e. choice deletion and lexical matching, were activated to facilitate a listening process, to bridge gaps in comprehension, and to select appropriate answers. Overall, it is possible to conclude the test assessed the three listening abilities it was aimed to –listening for specific details, listening for main idea, and inferencing. However, considering what the test-takers reported doing while completing the test, there appears a threat to construct under-representation. Stimulated recalls data revealed that taker-takers relied to a large extent on test-

wise strategies to obtain correct answers whereas some strategies that L2 listening literature suggest to be crucial in real-life listening such selective attention and real-time assessing of listening input did not appear to be assessed by this test. Given to the fact that language tests should tap into the abilities performed in real-life situations, to improve the validity of the PSU-TEP listening, this study recommends to re-conceptualize the test construct as mental processes listeners need in non-testing contexts and define the listening construct accordingly in order to tap into listening abilities required in the real world.

บทคัดย่อ

หนึ่งในข้อกังวลหลักของการทดสอบทักษะการฟังของข้อสอบ PSU-TEP คือความตรงเชิงโครงสร้างหรือศักยภาพของข้อสอบในการวัดความสามารถทางการฟังที่ผู้ออกข้อสอบตั้งใจจะวัด ทั้งนี้เพราะตั้งแต่มีการใช้ข้อสอบนี้มายังไม่เคยมีการศึกษาความตรงเชิงโครงสร้างของข้อสอบมาก่อน การศึกษาครั้งนี้จึงได้มีการวิเคราะห์คะแนนสอบจากผู้เข้ารับการทดสอบใน 4 รอบของการจัดสอบในปี 2560 เพื่อระบุจำนวนของทักษะการฟังย่อยที่ถูกประเมินโดยข้อสอบโดยใช้การวิเคราะห์องค์ประกอบ (Factor Analysis) และเพื่ออธิบายทักษะย่อยหรือความสามารถที่ใช้ในการฟัง งานวิจัยนี้ได้เก็บข้อมูลทางวาจาที่เก็บรวบรวมโดยใช้เทคนิคการกระตุ้นการเรียกคืน (Stimulated Recalls) จากผู้ให้ข้อมูลจำนวน 24 คน แบบตัวต่อตัว ข้อมูลที่ได้ถูกวิเคราะห์และใช้เพื่อการอภิปรายถึงขอบเขตความสามารถของผู้เข้าสอบที่ข้อสอบสามารถวัดได้

ผลการวิจัยพบว่าข้อสอบในแต่ละตอนได้วัดทักษะการฟังย่อยจำนวน 3-4 ทักษะ จากข้อมูลทางวาจาพบว่าทักษะย่อยนั้นประกอบด้วย การประมวลผลข้อความระดับย่อยเพื่อเข้าใจรายละเอียดของเรื่องที่ฟัง การประมวลผลข้อความระดับที่สูงขึ้นหรือระดับภาพรวมเพื่อเข้าใจประเด็นหลัก/ข้อความที่ผู้พูดต้องการจะสื่อ นอกเหนือจากนี้พบว่ามีการใช้กลยุทธ์การฟังประเภทต่าง ๆ คือ การอนุมาน การใช้ความรู้รอบตัวด้านต่างๆ เพื่อประมวลผลเรื่องที่ฟังให้ชัดขึ้น การทำนายเนื้อหาของเรื่องที่ฟังล่วงหน้า และการตรวจสอบความเข้าใจระหว่างการฟังเพื่อจัดการกระบวนการฟังให้มีประสิทธิภาพ และปะติดปะต่อเรื่องราวจากข้อมูลบางส่วนที่ได้จากการฟัง นอกจากนี้ยังพบว่ามีการใช้กลยุทธ์การทำข้อสอบอย่างชาญฉลาด เช่น การตัดตัวเลือกที่ไม่เกี่ยวข้องออกและการจับคู่คำศัพท์ในตัวเลือกกับเสียงที่ได้ยินเพื่อเลือกคำตอบที่เหมาะสม ผลจากการวิจัยโดยภาพ สามารถสรุปได้ว่าข้อสอบที่ใช้สามารถวัดความสามารถทั้ง 3 ส่วนของการฟังได้ตามวัตถุประสงค์ของข้อสอบซึ่ง ประกอบด้วย การฟังเพื่อเข้าใจรายละเอียดของเรื่อง การฟังเพื่อเข้าใจข้อความสำคัญหรือแนวคิดหลัก และการอนุมาน อย่างไรก็ตามเมื่อพิจารณาถึงกระบวนการฟังที่ผู้เข้ารับการทดสอบรายงาน พบว่ามีข้อบกพร่องในการกำหนดทักษะการฟังที่ถูกวัดโดยการใช้ข้อสอบดังกล่าวซึ่งส่งผลให้ข้อสอบไม่ได้วัดความสามารถที่จำเป็นแท้จริงสำหรับการฟังในชีวิตประจำวัน การฟังเพื่อความเข้าใจเป็นกระบวนการที่อาศัยการประมวลผลข้อมูลในหลายส่วนและหลายระดับและมีการใช้กลยุทธ์ทางการฟังเพื่อทำให้การฟังมีประสิทธิภาพ ข้อมูลจากการกระตุ้นการเรียกคืนเผยให้เห็นกระบวนการฟังและการสร้างความรู้ความเข้าใจหลายอย่างซึ่งมีความสำคัญในการฟังในบริบทที่ไม่ใช่เป็นการทดสอบแต่ไม่ได้ถูกกำหนดให้เป็นทักษะย่อยที่ข้อสอบจะวัด ดังนั้นจากหลักการที่ว่าแบทดสอบที่มีความเที่ยงตรงเชิงโครงสร้าง ควรวัดความสามารถที่มีการใช้จริงในสถานการณ์การฟังจริง งานวิจัยชิ้นนี้จึงเสนอแนะให้มีการปรับปรุงความเที่ยงเชิงโครงสร้างของแบบทดสอบการฟังของ PSU-TEPโดยกำหนดโครงสร้างการทดสอบทักษะย่อยและประมวลผลข้อความตามทักษะและกระบวนการที่เกิดขึ้นจริงในชีวิตประจำวัน

Acknowledgements

This research project would not have been completed without the help and support of various individuals and organizations to whom I would like to offer my sincerest thanks.

First of all, I would like to thank the Department of Languages and Linguistics, Faculty of Liberal Arts for granting a permission to have access to their test items and the Faculty of Liberal Arts, Prince of Songkla University for a research grant and Research Section, Faculty of Liberal Arts, for providing financial support

Additionally, I would like to thank Assistant Professor Dr Chonlada Laohawiriyanon, who has been a massive inspiration to me in academic life and a research mentor in this project.

Special thanks to 24 participants who, without hesitation, agreed to participate in this research and all the test-takers of the PSU-TEP in 2017. Without their contribution, this research would not have been completed.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background to the study

The Prince of Songkla University Test of English Proficiency (PSU-TEP), produced and administered by the Department of Languages and Linguistics, Faculty of Liberal Arts, PSU, is an English proficiency test for non-native speakers of English. The test has been developed from the Prince of Songkla University Graduate English Test (PSU-GET), formerly used as an English exit test for post-graduate students at the university. As other uses of the PSU-GET have later been identified, e.g., for professional purposes, in January, 2013, the construct of the test was revised and the test was renamed the PSU-TEP, or Prince of Songkla Test of English Proficiency. The test is administered four times a year, each to approximately 500-800 test takers. It consists of four components: 1) reading and structure, 2) listening, 3) writing, and 4) speaking, each of which has its own construct. Designed to assess test-takers' ability to understand written and spoken English and to communicate effectively in English, the test is aimed to provide test-takers with a valid and reliable measure of their English proficiency.

One major concern over the use of the PSU-TEP is its construct validity or whether the test measures what it is aimed to measure. This is since the design of the PSU-GET/TEP test, no test validation was carried out, and thus the extent to which the test taps into the targeted construct remains largely unclear. Tests, as concerned by language testers (see Alderson, Clapham & Wall, 1995; Bachman & Palmer, 2010) may not assess the construct they are aimed to for a number of reasons, e.g., the process of item writing and developing and scoring. If this

was the case, the interpretations and inferences made on the basis of test scores would not be justified and this would affect the quality and the credibility of the test (Bachman & Palmer, 2010). It also raises an issue of test fairness (Bachman & Palmer, 2010). Through construct validation, or the process of determining the extent to which tests measure what they are aimed to measure, language testing can be put on a sounder and more scientific footing (Hughes, 2003). With regard to this issue, an investigation of the construct validity is therefore crucially important.

This study is part of a large-scale research project which aims to investigate the construct validity of the PSU-TEP in four components: structure and reading, listening, speaking, and writing. It is aimed specifically to investigate the construct validity of only the listening component.

1.2 Scope of the study

Although the PSU-TEP has been administrated every year since 2013, this study only focused on the test papers administered in 2017 and only the listening component was researched. The study attempted to define the abilities that test-takers performed in order to answer the test question correctly. In line with the unified concept of test validity proposed by Messick's (1989) and cognitive framework for test development and validation presented by Weir's (2005), the study conceptualizes the construct underlying a test as the cognitive processes and strategies test-takers activated to complete the test and investigate what processes and strategies test-takers activate to perform the test successfully. Based on these data, the abilities that were actually assessed by the test were inferred.

1.3 Significance of the study

The significance of the study is twofold. **Practically**, this study provides an understanding of what listening abilities are actually assessed by listening component of the PSU-TEP. The processes/strategies that test-takers rely on to complete the test successfully, in particular, point out to abilities assessed by the test. Additionally, the study informs the test developing team of the extent to which the construct they aim to measure are actually measured by the test they used, as well as the strengths and limitations of the item types they included. These are all to assist the test developers to improve the quality of the test to best assess the intended construct. **Theoretically**, the study provides additional empirical evidence to improve understanding of cognitive processing for listening comprehension, which has been less researched compared to the other language skills – reading, writing, and speaking (Taylor & Geranpayeh, 2011; Wang & Treffers-Daller, 2017; Leonard, 2019) and as a consequence, its construct is less likely to be informed. This is particularly in an English as a foreign language (EFL) context, where the present study was carried out.

1.4 Definition of terms

A number of terms will be used throughout this research report. This section provides definitions of terms that are key to this study.

‘Test construct, conceptualized on the basis of a unified validity framework proposed by Messick’s (1989), refers to the mental processes that test-takers activate to complete test items/ task s successfully.

'Listening abilities' are the abilities that language users performed to comprehend listening texts. In this study, listening abilities are conceptualized as cognitive and strategic processing

'Cognitive processing' refers to the activation of six processes for listening comprehension. They are 1) acoustic-phonetic processing, 2) word-decoding, 3) parsing, 4) semantic processing at the local level, 5) semantic processing at the higher level, and 6) pragmatic processing.

'Strategic processing' refers to mental activities that language users perform in order to monitor their listening and fulfil gaps in their comprehension. In this study, it refers to inferencing, elaboration, prediction, the use of first language (L1), directed attention, comprehension monitoring, and note-taking.

'Test-wise strategies' are techniques that test-takers used to get a correct answer to test items, not the strategies that are actually performed in a non-testing situation. They are not part of listening construct but are strategically performed just to maximize test scores. In this study, they refer to as 1) choice deletion, 2) lexical matching, and 3) guessing.

Chapter 2 Literature Review

This section reviews the literature related to the focus of the study and data collection. First, the literature on construct validity is reviewed in order to inform aspects of test performances that need to be looked into in test validation. Next, the construct of listening comprehension, as theoretically indicated, is reviewed in order to provide guidelines for the analysis of listening processes. Towards the end of the section, a description of stimulated recall used to collect research data is provided.

2.1 Construct validity

Construct validity, as postulated by language testers, (e.g., Alderson et al., 1995; Hughes, 2003; Weir, 2005), is one important quality of valid tests. It concerns the potential of the test to measure what it is aimed to measure. When the test was not valid, the inferences of the test takers' ability drawn on the basis of their test scores would not be justified and this would raise the questions of the accurate representation and the usefulness of the test.

To ensure the quality of any test used, language testers (e.g., Messick, 1989; Alderson et al., 1995; Weir, 2005) have stressed that the construct of tests has to be clearly defined and investigated. If not, the test would probably suffer the two threats to construct validity. One is construct-underrepresentation, meaning that tests fail to measure part of the elements aimed to test. The second threat is construct-irrelevance, occurring when the test is too broad and other factors such as background knowledge, test methods, and test-wise strategies contribute to success in test performance.

Traditionally, validity has been described as a property of a test which can be divided into different components. Test validity was therefore separated into different categories and types. Cronbach and Meehl (1955) classify test validity into four types: predictive, concurrent, content and construct validities. Later, Alderson et al (1995) explain test validity into three categories and six types: internal validity, consisting of face, content, and response validities, external validity including concurrent and predictive validities, and construct validity, each of which exists independently and can be investigated in a number of indifferent ways. According to these views, construct validity is investigated by, for example, 1) using an expert to judge whether or not a test or test items measure what they are aimed to measure according to the test specification, 2) analysing test performances (scores) to identify the association between the items and the underlying construct, and 3) by exploring the relationship of test results to other external measures.

The analysis of test scores, in particular, is an important part of construct validation. Bachman (2004) explains that for any testing purpose, test designers have to define abilities that will be assessed by the test before designing it. These abilities may consist of different elements which are believed to contribute to the overall abilities. For example, overall listening abilities may be defined as the ability to decode a continuous speech, the ability to understand specific details, and the ability to understand a main idea. One way to assess construct validity, as Alderson et al. (1995) suggest, is to investigate the correlation between the test components and the items included in the test. If different components of one test are to assess abilities that contribute to the overall ability aimed to assess by the test, these components should correlate significantly at a certain level. Alderson et al. (1995), in particular, recommend a correlation of +.7 or higher. When the components are not significantly correlated or they are independent from

one another, this could mean that they measure different constructs (Alderson et al., 1995). On the other hand, when the items/components are highly correlated, e.g., at +.9, this could indicate that they measure the same construct (Alderson et al., 1995).

However, statistical analysis of test scores, as pointed out by Messick (1989), is not sufficient to investigate test's construct validity. In his proposal of a unified validity framework, Messick (1989; 1995) particularly claimed that construct validity not only accounts for the quality of a test itself, but is a characteristic of the inferences drawn on test scores and the consequences of the assessment as a whole. Test validity may not be demonstrated only through relevant content and operative processes such as the correlation between the scores from the test in question and the scores from other external measures, as traditionally conceived (Messick, 1995). On the other hand, it concerns the extent to which a test can be shown to produce data, i.e. test scores, which are accurate representation of test takers' level of language knowledge or skills (Messick, 1995). The construct underlying test tasks, according to Messick (1995), is no longer viewed as only a component of language ability theoretically indicated but cognitive processes that individuals demonstrate to achieve test tasks, and the extent to which those processes applied and the knowledge used to complete the test represent those performed in the situations where the test results are generalized to. The context where the test is designed and used therefore needs to be taken into an account in test validation. A set of construct indicators, which necessarily explain the construct underlying tests, therefore includes cognitive processes, strategies, and linguistic and non-linguistic knowledge applied to complete tests (Messick, 1995).

In line with Messick (1995), Weir (2005) in his socio-cognitive framework for test validation conceptualizes a test construct as mental processes employed by test-takers to complete test items/tasks. Weir, in particular, asks if processes adopted during test performance

resemble those normally operated in real world contexts. If test tasks tap into additional processes that the test-takers adopt just to complete a test, such as test-wise strategies, rather than necessary processes employed in the target situations, the test would yield construct irrelevance. Weir thus suggests evidence on cognitive processes that test takers rely on to complete the test is important for construct validation.

Following this line of thought, several recent construct studies have gathered data on test-taking processes to describe test construct. There are for example, Swain et al., (2009), investigating the construct validity of the TOEFL iBT speaking test, Plakans (2008) and Gebril (2010), studying the construct underlying reading and integrated reading-writing test items, Field (2013), studying the construct of IELTS listening test, and Rukthong and Brunfaut (2019) exploring the construct of listening-integrated test tasks of the PET Academic. These studies all support the importance of test-takers' processes for defining the abilities underlying the test used and justifying test validity.

The use of test-takers' processes to determine construct validity is certainly not new to construct validation. The importance of gathering information on test taking processes as part of construct validity has been recognized in several previous research in 1990s'. For instance, in Anderson, Bachman, Perkins, and Cohen (1991), data on the processes test-takers relied on to complete a multiple-choice reading comprehension test was combined with test scores to study the construct validity of a reading comprehension test. Sixty Spanish ESL students were asked to think aloud while completing two parallel multiple choice reading comprehension tests. The statistical analysis of test scores showed that of 90 items investigated, 62 items fall within acceptable range of item discrimination and 28 items falling below a cut-off point. Seventeen strategies were reported being activated by the test-takers, and nine of them were significantly

related to the item difficulty as quantitatively analysed. These results, as indicated by the researchers, enable them to understand the interactions among, test taking strategies, item content, and test scores. The verbal data revealed types of knowledge (e.g., lexical and syntactical knowledge) test-takers used to complete the items and the reasons underlying their decision to choose the answer to each item. Evidence on test-taking processes, as the researchers claimed, helped them to better determine if the test items were functioning as they were intended to function. The researchers, therefore, recommend combining test-taking processes with statistical data in construct validation.

Another study of construct validity that included test-taking processes is Storey's (1997). This study investigated the construct of a discourse cloze test by looking into cognitive processes test-takers engaged in while performing the test. Think-aloud protocols were used to collect data from 25 Hong Kong Chinese students. The participants were asked to report their test-taking processes and the reasoning underlying the selection of each answer. The researcher then inferred the test-taking processes from what the participants reported doing and compared the processes with the test construct specified at an initial stage of the test design. The results showed the processes generated by discourse cloze-test items have varying degrees of construct validity. Items consisting of deleted discourse markers engaged participant to decompose the associated arguments and analyse the structure of the text in depth. Items consisting of deleting cohesive ties were less successful since test-takers were able to rely on surface matching and answer the items correctly. The study, as the researcher contended, revealed test-taking processes could feed into the construct underlying each item. Particularly, it showed whether the testing technique and the test items used captured the processes they were aimed for.

To conclude, the literature review has suggested that different sources of data are essential for test validation. While the traditional conceptualizations of construct validity emphasize the importance of test scores and statistical procedures in construct validation, the more recent conceptualizations of the test validity claim that only statistical analyses may not be sufficient to justify the construct validity of tests. What is also important, as recommended by the recent thinking of test validity, is evidence on test-taking processes. This is because it can point out the reasoning (thought process) underlying the selection of answers, revealing what abilities test-takers performed to complete the tests. Taking into account of what the literature has suggested, this study is therefore specifically set out to investigate the construct validity of the PSU-TEP listening by combining two different sources of data, i.e., test scores and test-taking processes.

2.2. The construct of listening comprehension

Listening comprehension, as Buck (2001) has pointed out, is a complex and multidimensional process. As a result, it is not easy to define its underlying construct. Several researchers (e.g., Anderson, 1985; Rost, 2011; Field, 2013; Vandergrift & Goh, 2012) have defined abilities or sub-skills involved in listening comprehension and they seem to agree that the ability to comprehend listening texts entails two important factors: the level of knowledge listeners possess and the ability to process texts automatically in real time.

Listeners' knowledge, according to Rost (2011) and Field (2013), comprises both linguistic and non-linguistic knowledge. Linguistic or language-related knowledge is a domain of information in an individual's memory used to create and interpret discourse in language use. It includes knowledge of phonology/graphology, lexis, and syntax. These types of knowledge are

employed mainly during linguistic processing. They enable listeners to encode speech into linguistic units, detect phonetic features and recognize words in connected speech, and interpret the incoming text. Semantic and pragmatic knowledge is generally activated at a high level of processing, i.e. meaning and discourse construction (Field, 2013). It enables listeners to interpret textual discourse by relating utterances or sentences to each other, to the speaker's intentions, and to characteristics of the language use setting (Bachman & Palmer, 2010).

Another type of knowledge that affects L2 listening is the cultural and world knowledge that listeners bring to listening situations (Field, 2013). Knowledge of this kind has been found to be shaped by listeners' cultural background and experience. It is activated mainly at a high-level of text processing and is especially crucial when listeners have to make inferences or elaborations on the message being delivered (Field, 2013).

In addition to the knowledge that listeners possess, Field (2013) describes that it is important that listeners can process that knowledge automatically in real time. Field (2013) divides this processing into two level categories according to listener's cognitive operations: lower-level processes and higher-level processes (see Figure 1). The lower-level processes in listening, according to Field (2013), start from recognizing acoustic input and develop to obtaining a phonological string by input decoding, a set of words by lexical searching, and an abstract proposition by parsing. Field explains that for input decoding, proficient listeners depend on their phonological knowledge to access a sequence of speech-like sounds and convert these sounds into representations that match the phonological system of the language being spoken. This processing enables listeners to recognize strings of phonemes, some of which are marked as syllables or words. For the lexical search, listeners map sounds to spoken word forms. Based on their lexical knowledge, listeners determine word boundaries and identify words which are either

content or function words in the connected speech. At the level of parsing, listeners separate units in the connected speech and construct propositions by applying their syntactic knowledge, an understanding of standard word order, and intonation group boundaries. The output of lower-level processing is an understanding of independent idea units of the listening text.

The higher-level processes involve two levels of processing; meaning and discourse construction (Field, 2013). Listeners start to construct the actual meaning of what they have heard by relating the propositions they obtained from the lower-level processing, which are context-independent, to their own schemata or the concepts of knowledge they have developed. At the level of meaning construction, it is the listeners' task to relate the propositions to the circumstances in which they were produced to obtain their full and relevant meaning. What the speaker said is often the raw meaning of the speaker's words and insufficient to convey the complete meaning of a text (Field, 2013). Listeners, therefore, have to supply information to

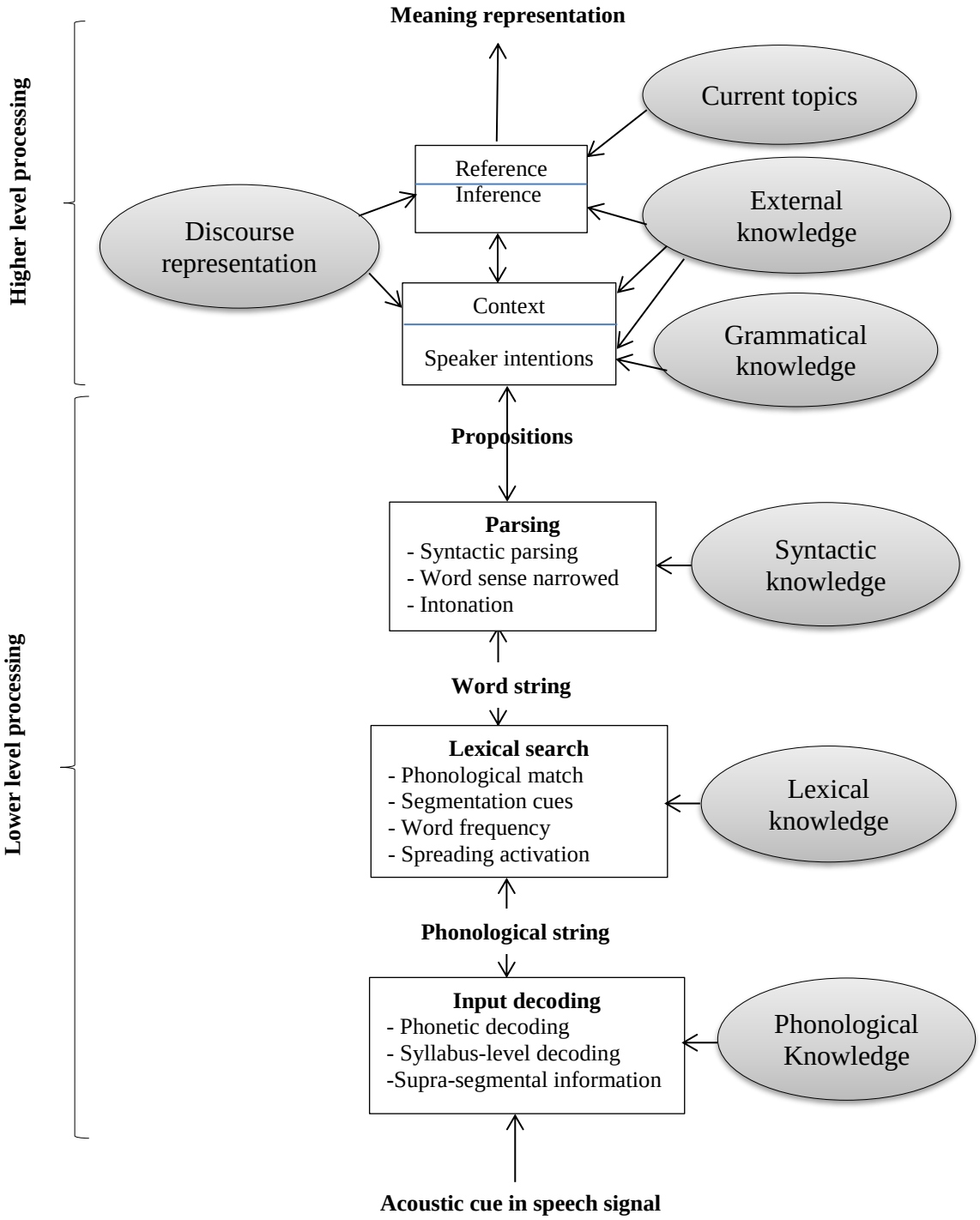

Figure 1: A cognitive processing framework adapted from Field (2005, p. 97, 101)

comprehend what is said in a number of ways. One way to do this is using pragmatic knowledge to interpret the speaker's intentions. Listeners may also have to use contextual and semantic knowledge to relate the propositions to the context in which they occur. Listeners, in addition, may have to infer what the speaker left unsaid from what they have just heard or backtrack from what is being said to what has been said earlier.

For discourse construction, Kintsch and van Dijk (1978) explain that it is related to the processes that listeners apply to construct an understanding of a spoken text. Field (2013) later separates discourse construction for listening into four processes: selecting, integrating, self-monitoring, and structure building. Selecting is deciding on the relevance of an incoming piece of information; for example, whether it is a repetition of a point made earlier or the central point of the topic being developed. On the basis of this decision, listeners may store, or ignore as irrelevant, the information being processed. Integrating is when listeners add a new piece of information to the discourse representation being developed. It involves recognizing conceptual links between the incoming information and the information processed before. Self-monitoring entails comparing whether a new piece of information is consistent with what has been processed before. If not, listeners consider whether the new judgment is correct, or question whether what they have understood and recalled earlier is correct. Structure building is prioritizing and organizing the information stored according to its importance and relevance.

In addition to Filed (2005), Vandergrift and Goh (2012) propose a cognitive model of listening comprehension to explain what is going on during a listening process and what elements contribute to listening comprehension. In line with Alderson (1995), Vandergrift and Goh (2012) described their theoretical model for listening comprehension into three different

processing components: perception, parsing, and utilization (see Figure 2). Perception is the lowest level of processing that listeners engage in. It involves converting sounds in speech stream into words or phrases. At this stage of processing, listeners can identify what words or groups of words they hear but cannot identify the points that the speaker aims to deliver. Parsing is the syntactic and semantic mapping of what was heard in the previous stage of processing, i.e. sound perception and word recognition. Listeners create a mental representation of what they listened by relying partly on the words/ groups of words they recognized in the perceptual stage combined with their existing semantic and syntactic knowledge. What is obtained at this level is 'propositional information' which is an independent idea unit segmented from a continuous speech and the listener does not yet understand how it is related to the theme of the passage. Comprehension at this level is still at what Field (2005) called a local level which takes place at the lower level processing. To understand the actual meaning of the text, listeners have to proceed to the utilization stage, which is when the listeners semantically map words, phrases, chunks of information which have been processed at the lower level and create a representation of the text they listened. Vandergrift and Goh (2012) point out that to create a meaning representation of the listening text, listeners appear to engage in not only bottom-up but top-down processing. That is in addition to the linguistic elements obtained from the lower-level processing, listeners have to rely on their discourse, pragmatic and prior knowledge of the world to interpret the real meaning of the text. Field (2013) considers processing at this stage the higher-level processing or the meaning construction stage, one of which is very important for comprehension to take place.

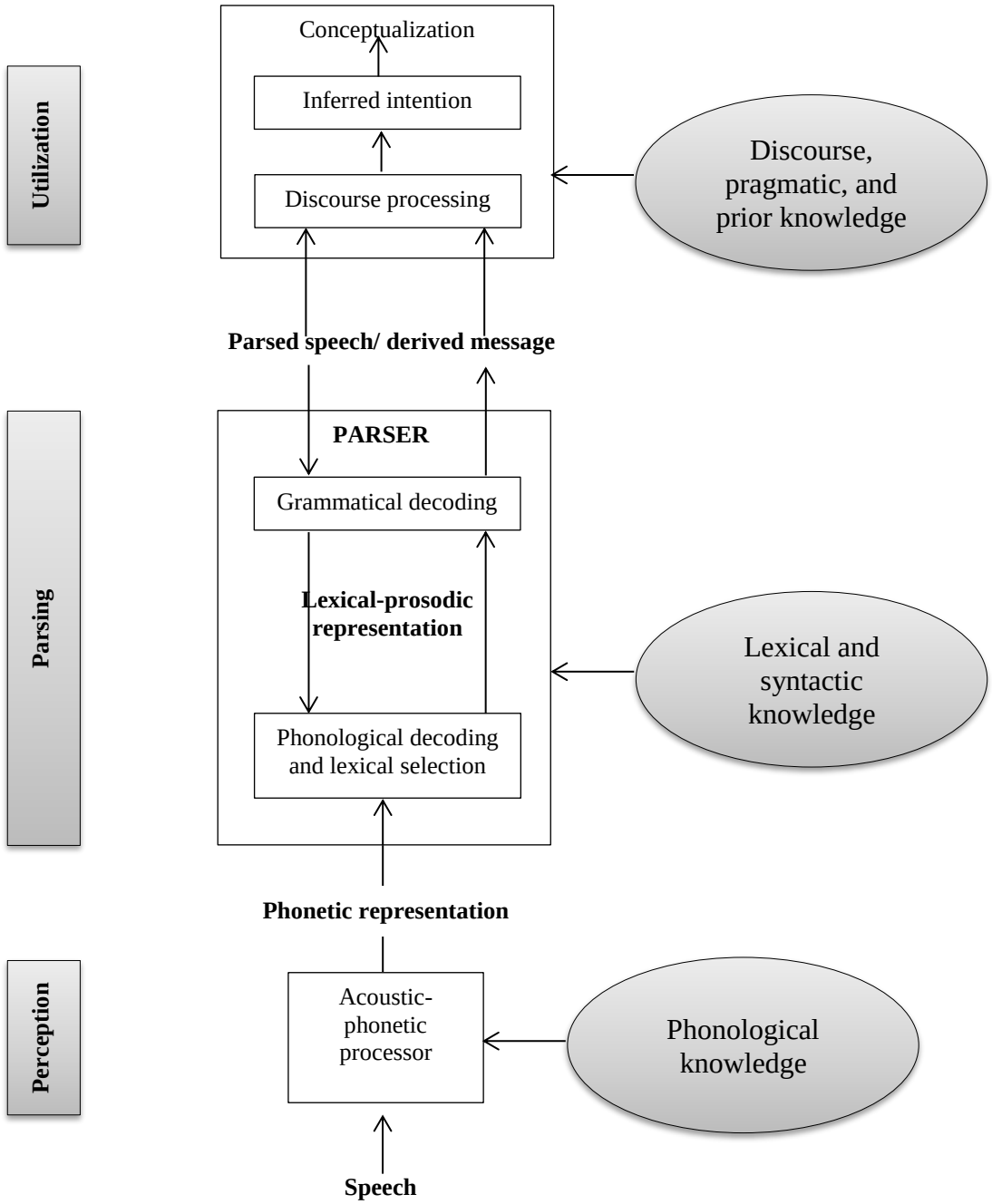

Figure 2: A cognitive model of listening comprehension by Vendergrift and Goh (2012, p.39)

In addition to cognitive processing, a considerable body of research in second language acquisition has indicated that strategic processing, which involves the use of cognitive and metacognitive strategies, plays an important role in L2 comprehension processing (Goh, 2002; Graham, Santos, & Vanderplank, 2008; O'Malley et al., 1989; Rubin, 1981; Vandergrift & Goh, 2012). This is because L2 learners have more limited L2 linguistic knowledge as well as contextual and cultural knowledge, which are crucial for comprehension to occur (Færch & Kasper, 1986). Cognitive strategies, such as inferencing and elaboration, are thereby essential to bridge gaps in the knowledge that may occur and increase comprehension of the text. However, some learners might have developed false beliefs about language learning that negatively affect listening comprehension processing (Færch & Kasper, 1986). For instance, they may think that in order to have a complete understanding of a text, they have to decode and understand every single linguistic unit in the input, which is not likely to be necessary or possible in listening situations which need rapid and online processing. To successfully understand a text, learners may thus need metacognitive strategies to manage their listening behaviours in order to catch up on what they are listening to.

The role of metacognition in language processing is, in fact, emphasized by a number of previous studies (e.g., Bachman & Palmer, 2010; Graham et al., 2008; Tanewong, 2018; Weir 2005). It helps regulate cognitive processes and enable the language learners to solve their problems in language processing. Specifically in listening, Vandergrift and Goh (2012) explain that metacognition is applied with some degree of consciousness. Listeners activate metacognitive strategies with some particular purposes, for example, planning for listening, monitoring comprehension, solving comprehension problems, and evaluating listening outcome or their understanding of the listening text.

Based on the literature reviewed, listening construct or listening abilities in this study is described in relation to two components of language processing, i.e., cognitive processing and strategic processing. Cognitive processing, which is operated on the basis of listeners' linguistic and topical knowledge, is a category of mental operations that contribute directly to text comprehension. Following Field (2013), it is sub-divided into six processing types, consisting of 1) acoustic-phonetic decoding, 2) word decoding, 3) parsing, 4) semantic processing at the local level, 5) semantic processing at the global level, and 6) pragmatic processing. Strategic processing is the use of strategies (both cognitive and metacognitive strategies) to solve problems occurring during listening and to facilitate the listening process. Strategic processing is different from cognitive processing that it involves some degree of consciousness whereas cognitive processing is automatic processing (Vandergrift & Goh, 2012). Strategic processing includes1) inferencing, or the use of linguistic information gained in listening to fill in missing information and guessing the meaning of unfamiliar words/parts, 2) elaboration, or using background knowledge or topical knowledge to make the text meaningful, 3) prediction, or anticipating listening content, 4) fixation, or stopping to think or focus attention on understanding a small part of a text, and 5) reconstruction, or using key words to recreate meaning of what is heard, 6) paying attention selectively to what a listener expects to hear, 7) re-directing attention when it is away from the incoming text.

2.3 Test-wise strategies

In a testing context, what take part in language processing in addition to cognitive processing and strategic processing, is the use of test-wise strategies. These strategies, according to Cohen (2006), occur when test-takers use the knowledge of test formats and other peripheral

information to provide an answer to the test items without relying on linguistic and cognitive processes expected to be applied. Test-wise strategies, as found in previous research (e.g., Cohen, 2007; Chang & Read, 2013; Yeldham & Gruba, 2014) include eliminating choices and lexical matching. Although research shows that higher proficiency learners used more test-wise strategies than the lower ones (Chang & Read, 2013), the use of test-wise strategies is not an indicator of test validity as it decreases the the potential of the test to assess what is aimed to assess. On the other hand, it is the evidence that shows the possible flaws of the test design (Field, 2013).

To sum up, theoretically, a listening test should only tap into listening abilities and listening construct should be composed of cognitive and strategic processing. In an actual situation, there is another category of test-taking activities that contribute to either success or failure in test performance. It is the use of test-wise strategies. The investigation of test taking processes in this study, therefore, was extended to the use of test-wise strategies in task completion (see Figure 3).

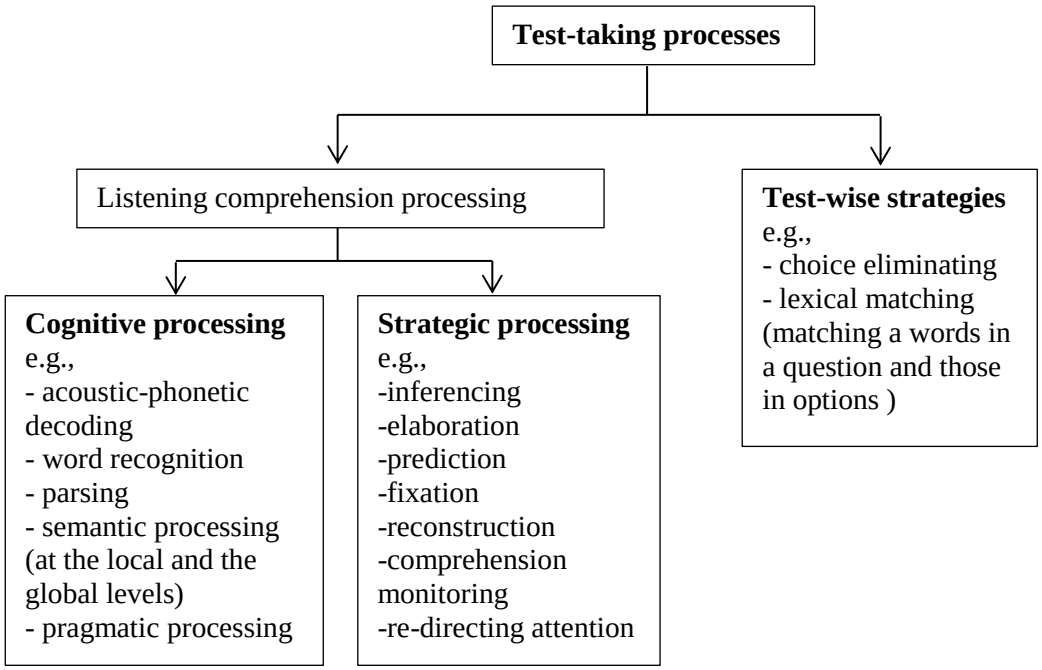

Figure 3: Components of listening test-taking processes included in the coding scheme

2.4 Stimulated recall

Stimulated recall is a verbal data collection technique which requires participants to directly verbalize the information heeded during or just after actual cognitive processing (Ericsson & Simon, 1993). Verbalizing thought processes at this period is not a description or explanation of cognitive processes one has engaged in to complete the task, but verbalization of what one was paying attention to or thinking about while generating answers to the task (Ericsson & Simon, 1993). Ericsson and Simon suggest that in tasks that take less than 10 seconds to complete, participants can recall their thought processes with supposedly high accuracy and completeness. This is because some retrieval cues are thought to remain in their short-term memory (STM). However, the longer the period between task processing and retrospective reporting, the more difficult and incomplete the recall becomes. Therefore, stimuli that can help individuals to recall

their thought processes are recommended to be included (Ericsson & Simon, 1993), such as a video recording of task performing behaviours, notes taken by the participants, or task output (Gass & Mackey, 2000).

The key advantage of this data collection technique, as indicated in previous research (e.g., Ren, 2013; Barkaoui et al., 2013) is that it reveals processes and strategies used to complete the tasks which are not otherwise directly observable by the researcher(s) or tapped into by other methods, e.g., questionnaires. Studies that have used the technique (see Ren, 2013; Barkaoui et al., 2013; Swain et al., 2009) have found that verbal protocol data evidenced strategies and processes activated successfully and unsuccessfully by the participants as well as the knowledge sources participants used to complete the tasks.

A stimulated recall was used to collect data on test-taking processes in this study for two practical reasons. First, test responses in this study were scored for the participants' language ability. Stimulated recalls, which were conducted after the participants complete each part of the test, were considered appropriate as it minimized the effect of the data collection technique (if any) on test performance. Second, as it is necessary to collect data after task completion, stimulated recalls, where some stimuli (a video recorded during the task performance and the test responses) were presented to stimulate their thought processes, were hoped to provide insightful data.

In conclusion, the investigation of the PSU-TEP listening construct validity in this study combined two types of data, i.e., test-takers' performance (test-takers' scores) and their test-taking processes, collected by the use of a stimulated recall technique. The analysis of test performances was hoped to provide data on item difficulty, discrimination index, and numbers of sub-components of listening ability tapped into by the test. The analysis of test-taking processes

was aimed to reveal sources of knowledge and the reasoning or thought process test-takers relied on to complete the test. This is in order to determine abilities test-takers performed to complete the tasks and the extent to which the listening construct aimed to measure is actually measured by the test used.

Chapter 3 Research Methodology

The purpose of this study is to uncover test-taking processes that the test-takers used to complete the PSU-TEP listening. To achieve the aim of the study both qualitative and qualitative data were collected. The quantitative data are test-scores from the test takers taking the test in one year round of the test administration. The qualitative data, on the other hand, are test-taking processes test-takers reported activating while completing the tests. Following is the detailed description of the research methodology employed in this study.

3.1 Research Questions

The following three research questions are formulated in this study.

1. What are the sub-components of listening ability captured by the PSU-TEP listening?
2. What are the test-taking processes test takers activated to complete the PSU-TEP listening test?
3. To what extent does the listening component of the PSU-TEP measure what it is aimed to measure?

3.2 Participants

The participants in this study were divided into two groups. One was the group of the test-takers taking the tests administered in the year 2017. They were postgraduate students in different programs including Sciences, Social Sciences, and Health Sciences. A total number of the

participants in this group were 607 participants. For the first round of the test administration, there were 181 test-takers, the second round 146, the third round 134, and the fourth round 146.

The other group involved 24 participants. These participants were invited to participate in stimulated recall. They were purposefully selected to suit the research purpose, which was to investigate test-taking processes test-takers engaged in to complete the tests. There were all undergraduate students at Prince of Songkla University in the academic year of 2016. Half of them (12 participants) were science students (6 participants were from the faculty of Science and 4 from the faculty of medicine and 2 from the faculty of pharmacy). The other half was the students from the faculty of Liberal Arts. There were 2 main reasons why these participants were chosen. First, as the study aimed to reveal the test-taking processes that the tests were able to elicit to describe the abilities assessed by the tests, the participants who could performed the test successfully should be included. If the participants had difficulty completing the tests, it would be almost impossible for the test-takers to explain or share what processes or strategies they engaged in to complete the test (see Rukthong, 2016). The second criteria for the selection of the participants was that the participants had to offer their willingness to participate in the study as to they had to provide at least two and a half hours for data collection. The students who had sufficient English to successfully complete the tests but were not sure whether they could provide 2.5 hours for the study were excluded.

In order to identify the potential participants, the researcher first contacted the English teachers at the faculty of Liberal Arts to identify the students who were competent in English and were likely to perform successfully in the listening test. Then the researcher contacted the study by e-mail or phone to ask if they were able to participate in the tests. They were explained what they were supposed to do and how long the data collection process would take if they agreed to

participate. In addition, they were told that they could withdraw from the study when they were uncomfortable with the data collection process. 200 baths was offered to the participants for their participation.

3.3 Research materials

Four parallel versions of PSU-TEP listening, administered in 2016, were investigated. Each version of the test consisted of 30 multiple choice items and it took one hour to be completed. The test was divided into three parts. As shown in Table 1, Part I (Items 1-10), composed of 10 short conversations, aimed to measure test takers' ability to understand everyday English. Each conversation was about 15-20 seconds long and one question was asked to check either local or global understanding of the text. The questions targeting local comprehension are, for example, what did the man buy?, what is not mentioned in the text?, and how much is 'A'?. The global comprehension questions asks the test-takers to rely on the text and anticipate what is going to happen next or what the conversation is mainly about.

Part II (Items 11-20), consisting of three longer conversations, aimed to assess test-takers' ability to understand English used in a university context. After listening to each conversation which was about 1.00-1.15 minutes long, test-takers were required to answer 3-4 items. The questions asked for each conversation targeted at both local and global understanding.

Part III (Items 21-30) was made up of two interviews/advertisements (about 1.5-2.0 minutes long) and five questions were asked after each talk. Four questions focused on specific details such as what is the advantage of 'B' or what and one asking about global understanding of the text such as what the purpose of the text is or asking to identify the purpose of the speaker

for conveying a particular message in the long text. It aimed to assess the ability to comprehend casual talks in English.

Listening test	Type of listening input	Number of listening input	Number of questions (items)	Length of each listening input
Part I	Short conversation	10	10	15-20 seconds
Part II	Longer conversation	3	10	0.40-1.15 minutes
Part III	Interview/Advertisement	2	10	1.5-2 minutes

Table 1: Listening test items

To complete the test, test takers had to first listen to a conversation/talk, and then listen to questions and choose an answer to the questions, one by one. No question preview was allowed in this test and test-takers listened only once.

3.4 Data collection

As mentioned earlier, both quantitative and qualitative data were gathered (see Table 2). To answer the first research question, qualitative data, i.e., test scores were collected from actual test takers of the PSU-TEP listening in 2017. To answer the second research question, stimulated recalls were separately organized with a different group of participants in each administration of the PSU-TEP in 2017. This was in order to collect data on test-taking processes activated to complete the test.

Research questions	Research data	Research instruments/data collection technique
1. What are the sub-components of listening ability captured by the PSU-TEP listening?	Test scores	The PSU-TEP test
2. What are the test-taking processes test takers activate to complete the PSU-TEP listening test?	Verbal data	Stimulated recalls
3. To what extent does the listening component of the PSU-TEP measure what it is aimed to measure?	Test scores and verbal data	The PSU-TEP test and stimulated recalls

Table 2: Research questions and an overview of data collection techniques

Six participants with different academic backgrounds were invited to take part in stimulated recalls in each administration of the PSU-TEP: three of them were Science students and the other three were Social Science students. As four versions of the test were investigated, a total of 24 participants were therefore included.

3.5 Stimulated recalls

Stimulated recalls were organized on a one-to-one basis with the 24 participants. In addition to the researcher, two research assistants were included to carry out stimulated recalls. A training session for stimulated recalls was provided by the researcher prior to actual data collection. In this training, the researcher first explained how stimulated recalls should be conducted and then the research assistants practiced conducting stimulated recalls by using two sample items of listening tests. As stimulated recalls were aimed at gathering data about the participants' thought processes while listening, the following questions were asked in order to help stimulate their thoughts.

1) What were you paying attention to while listening?

2) What sentences/phrases/words were you paying attention to at a particular moment in the listening?

3) How did you understand this part of the listening?

In the actual data collection, the organization of stimulated recalls followed these steps.

1. The researcher/research assistants explained to the participants what stimulated recall was, what the purpose of this data collection was, and what they were supposed to do in the data collection.

2. The participants completed a listening sample item and then stimulated recall which was organized immediately after they finished answering the questions in each conversation.

3. The actual listening test items were delivered and the participants were asked to first complete the listening test on a one-to-one basis. An audio was recorded while the participants were doing the test.

4. After the participants finish listening and answering the question(s) in each conversation, the listening test was paused and stimulated recalls was conducted. Specifically, the participants were asked to stop doing the test, listen to the audio file recorded during the task performance, and look at their answer sheet as well as their notes taken while listening, and then explain what they were thinking about or paying attention to while listening. The same process of data collection was repeated in every part of the listening test.

3.6 Data analysis

Data were analysed as follows:

1) Test scores collected from the actual test takers of the PSU-TEP in 2016 were analyzed by using SPSS. In this manner, descriptive statistics, discrimination index, and internal reliability were calculated in order to look at the performance of each individual item and the test, as a whole. Factor analysis was then carried out to investigate the construct underlying the tests. Particularly, this analysis enabled the researcher to determine the number of sub-components of listening ability captured by the test and then identify the items that performed differently from the other items under the same construct.

2) Stimulated recall data were analysed to explain sub-components of listening ability which were obtained from the statistical analysis. To be specific, the researcher inferred types of cognitive processes and strategies the participants activated to complete the test on the basis of what the participants reported doing/ thinking about while doing the test. A coding scheme drawn from the literature reviewed was used to analyze stimulated recall data. As shown in Figure 1, two main components of coding scheme were identified, i.e., listening comprehension processing and test-wise strategies. The former consists of two processing types of listening comprehension: cognitive and strategic processing, which were important for comprehension to take place. The latter (test-wise strategies) were techniques test takers adopted maximize a test score and therefore were not considered as part of listening construct.

Following is an example of the coding of the data obtained from the participant doing the tests. To begin with, what the participants reported was separated into what appeared to be plausible units of listening processes, as described in the coding scheme. After that categories and types of listening processes were assigned to the chunks.

Example 1

Stimulated recall transcription	**Analysis of the data**
I was trying to follow what he (the speaker) said…….he talked about several places like //*Bangkok, Bali, Vietnam, and Philippines*//	Cognitive processing: **word recognition**
…. Here he talked about // *'Bangkok' and 'nightlife'.*//	Cognitive processing: **word recognition**
//*I guessed he said Bangkok is good for nightlife and transportation because he said 'it is easy to travel here'*//.	Cognitive processing: **semantic processing (Local level)**
Here, I heard //*Vietnam*//,	Cognitive processing: **word recognition**
//…so I thought another place to live in Asia is Vietnam.//	Strategic processing: **inferencing**
I did not quite get it here. It was fast. I was not sure what he was talking about. I heard //*south*	Cognitive processing: **word recognition**

…..with something, and then 'Vietnam' and then 'relax'.//	
Question 3 asked 'what city does he recommend for hiring an English speaking teacher? I thought it should be (option) 2 Cebu, Philippines. Here //*I deleted Bangkok because he talked about 'nightlife'. I did not choose Vietnam because I think it is for relaxing.*//	Test-wise strategies: **choice eliminating**
Actually, I was not sure about Bali and Philippines. I did not choose Bali because //*I know that it is famous for beaches*.// I guessed the answer was Philippines.	Strategic processing: **elaboration**

After the identification of the listening processes, frequency counts of the processes were made and overall pictures of the use of listening processes were presented in the result section. As indicated in the literature review, cognitive processes are more automatically used by strategies, participants' notes taken while listening and their answers were investigated to confirm the use of the listening processes.

Taking into the consideration that the participants were able to explain what they were paying attention to or thinking about while doing the tests but not the types of listening processes that they used to complete the listening tasks the tests, the research had to infer the types of cognitive processes and strategies used based on the information obtained. To ensure coding reliability, therefore, an external coder was used to re-code 25% of the stimulated-recall transcription. Cohen's Kappa analysis was carried out to investigate inter-coder reliability and the result showed the inter-coder agreement on the overall use of listening processes was .82, indicating an acceptable level of inter-coder reliability.

Chapter 4 Findings and Discussion

To describe the overall quality of the tests, an item analysis was first carried out. As in the year of 2017, the PSU-TEP was administered four times, four parallel listening tests were studied, and these four tests are referred to in this study as Listening Test 1 (administered in January), Listening Test 2 (administered in April), Listening Test 3 (administered in June), and Listening Test 4 (administered in December). The number of the test-takers in each test administration ranged between 134 and 181. The reliability of the tests used, as presented in Table 3, falls between 0.51 and 0.67, suggesting the reliability at a moderate level.

Test	**No. of Item**	**No. of Test Takers**	**Test Results**				**Reliability** (Cronbach's Alpha)
			Min.	**Max.**	$\bar{X}$	**S.D.**	
Listening Test 1	30	181	4	23	10.66	3.51	0.51
Listening Test 2	30	146	4	22	12.13	3.74	0.58
Listening Test 3	30	134	4	27	14.4	4.31	0.67
Listening Test 4	30	146	4	26	15.59	4.11	0.67

Table 3: Descriptive data of the tests

The following sections present the results, in accordance with the research questions.

4.1 What are the sub-components of listening ability captured by the PSU-TEP listening?

The literature review suggests that several processing abilities contribute to listening comprehension. The problem with the the PSU-TEP, as shown earlier, is that it is unclear what abilities are tapped into by the test. In order to investigate the construct underlying the listening test, this study collected both quantitative data (scores from test-takers in an actual test administration) and qualitative data (verbal report on test-taking processes). Factor analysis was

run on the test scores to identify the number of listening sub-components tapped into by the test. Then, verbal data were analysed to explain what ability each sub-construct measured. To ensure that there was sufficient number of cases to run factor analysis, the the analysis was divided into three parts, according to the construct of the test, i.e. Parts I-III.

After each test administration, the scores obtained were first run by SPSS to extract the numbers of factors in each part of the listening. In order not to violate the use of factor analysis that requires N (number of items loaded) x 10 participants in each analysis, only one part (with 10 items) was analysed at a time. Then to describe the construct, the researcher and an external rater who was an English as a Foreign Language (EFL) speaker with the experience in English language teaching and test development analysed stimulated recall data and identified what abilities the test-takers performed in order to choose the correct answer to each test question.

Figures 4-6 presents the results from factor analysis of Listening Test 1 when the Eigenvaluaes of higher than 1 was taken into account and the absolute values contributing to each component lower than 0.3 were suppressed. For Part I of the test (see Figure 4), 5 components were extracted. However, if taking only the highest positive absolute values that each item positively contributes to the components, there appear 3 significant components. They are Component 1, which includes questions 1, 3, 4, 7, and 8, Component 2, which includes questions 5, 9, and 10, and Component 5, which includes questions 2 and 6. For Part II, 4 components were extracted, 2 of which contain the highest positive absolute values. They are Component 1, which are questions 11, 14, 15, 16, 17, 18, 19 and 20, and Component 4, which contains questions 12 and 13. For Part III, 5 components were extracted and only 2 components showed the highest positive absolute values. They are Component 1, which includes questions 2, 3, 4, 6, 7, 8, 9, and 10, and Component 5, which is made up of questions 1 and 5.

Component Matrix[a]

	Component				
	1	2	3	4	5
Q1	.402	.383		-.481	.340
Q2		-.375			.661
Q3	.510				
Q4	.444		.303		
Q5		.628			
Q6		.406		.308	.577
Q7	.503		.527		
Q8	.552		-.381		-.365
Q10		.588		.493	
Q9	-.415	.430		-.587	

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Figure 4: Components of Listening Test 1, Part I extracted by Factor Analysis

Component Matrix[a]

	Component			
	1	2	3	4
Q11	.587	.396		
Q12				.677
Q13	.379			.581
Q14	.451	-.360	-.430	
Q15	.444	-.388	.431	
Q16	.533	.412		
Q17	.486	-.325		
Q18	.458		.422	
Q19	.624		-.336	
Q20	.557	.462		

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 5: Components of Listening Test 1, Part II extracted by Factor Analysis

Component Matrix[a]

	Component				
	1	2	3	4	5
Q21	.500	.419			.500
Q22	.502			-.536	
Q23	.632				
Q24	.363	-.609			
Q25	.608				.566
Q26	.408		-.424		
Q27	.503				
Q28	.379		-.421		-.522
Q29	.763				
Q30		.742			

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Figure 6: Components of Listening Test 1, Part III extracted by Factor Analysis

Figures 7-9 show the results from factor analysis of Listening Test 2 when the Eigenvaluaes of higher than 1 was taken into account and the absolute values contributing to each component lower than 0.3 were suppressed. For Part I, 5 components were extracted, 2 of which contain the highest positive absolute values in of each item. Component 1 includes question 1, 2, 4, 7, 8, and 10. Component 3 are questions 3, 5, 6, and 9. For Part II, 3 components were extracted, 3 of which contain the highest positive absolute values of each item. There are Component 1, which includes questions 11, 13, 14, 16, 17, 18 and 19, Component 2 which includes only question 20, and Component 4, which includes questions 12 and 15. Part III, as shown in Figure 9, was extracted into 4 components, 3 of which contain the highest positive values of each item. They are Component 1, which includes questions 23, 26, 27, and 29, Component 3, which includes only question 28, and Component 4, which includes questions 21, 22, 24, 25, and 30.

Component Matrix[a]

	Component				
	1	2	3	4	5
Q1	.642				
Q2	.598				.442
Q3			.729		-.324
Q4	.589	.412			
Q5		-.491	.439		
Q6			.693		
Q7	.431			.379	-.641
Q8	.404	-.527			
Q9		-.341	.659		
Q10	.462			-.406	

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Figure 7: Components of Listening Test 2, Part I extracted by Factor Analysis

Component Matrix[a]

	Component				
	1	2	3	4	5
Q11	.531				-.352
Q12				.684	
Q13	.451		-.499		.414
Q14	.477	-.362	.432		
Q15			.493	.703	
Q16	.534			-.377	.525
Q17	.652			-.351	
Q18	.490				
Q19	.620				.354
Q20		.545			.473

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Figure 8: Components of Listening Test 2, Part II extracted by Factor Analysis

Component Matrix[a]

	Component			
	1	2	3	4
Q21	.409			.743
Q22				.781
Q23	.686			
Q24		-.676		.583
Q25		.519		.619
Q26	.609		-.381	
Q27	.441	.326	.314	-.447
Q28	-.343		.408	
Q29	.703			
Q30				.552

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 9: Components of Listening Test 2, Part III extracted by Factor Analysis

Figures 10-12 present the results from factor analysis of Listening Test 3 when the Eigenvaluaes of higher than 1 was taken into account and the absolute values contributing to each component lower than 0.3 were suppressed. For Part I, 4 components were extracted, 2 of which contain the highest positive value of each item. They are Component 1, which includes questions 1, 2, 4, 7, 8, and 10, and Component 4, which includes questions 3, 5, 6, and 9. For Part II, 5 components were extracted. Three components contain the highest positive value of each item. They are Component 1, which includes questions 13, 14, 17, 18, and 19, Component 2, which includes questions 11, 12, and 20, and Component 3, which includes questions 15 and 16. For part III, 4 components were extracted, all of which contain the highest positive values of the items. Component 1 includes questions 22, 24, and 27. Component 2 contains questions 21, 26, 28, and 30. Component 3 is question 25, and Component 4 contains questions 23 and 29.

Component Matrix[a]

	Component			
	1	2	3	4
Q1	.575			
Q2	.596	-.313		
Q3	.344	-.539		.386
Q4	.469			.540
Q5	.325			.517
Q6		-.585		.300
Q7	.504			
Q8	.456		-.546	
Q9	.438	.332		.566
Q10	.619	.341	.580	

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 10: Components of Listening Test 3, Part I extracted by Factor Analysis

Component Matrix[a]

	Component				
	1	2	3	4	5
Q11	.406	.531			-.352
Q12		.684			
Q13	.451		-.499		.414
Q14	.477	-.362	.432		
Q15			.793	.703	
Q16			.534	-.377	.525
Q17	.652			-.351	
Q18	.490				
Q19	.620				.354
Q20		.545			.473

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Figure 11: Components of Listening Test 3, Part II extracted by Factor Analysis

Component Matrix[a]

	Component			
	1	2	3	4
Q21		.455		
Q22	.479		-.478	
Q23	.320	-.397		.467
Q24	.604		.331	
Q25	.356		.651	
Q26		.645		-.349
Q27	.665			
Q28		.523		-.450
Q29	-.305		.431	.553
Q30	.377	.536	.360	

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 12: Components of Listening Test 3, Part III extracted by Factor Analysis

Figures 13-15 present the results from factor analysis of Listening Test 4 when the Eigenvaluaes of higher than 1 was taken into account and the absolute values contributing to each component lower than 0.3 were suppressed. For part I, 4 components were extracted, 3 of which contain the highest positive values of the items. Component 1 includes questions 1, 2, 3, 4, 6, and 7. Component 2 are includes question 5, 8 and 10. Component 3 is question 9. Part II was extracted into 4 components, 3 of which contain the highest positive value of the items. They are Component 1, which includes questions 11, 12, 15, 18, 19, and 20, Component 2, which includes questions 13 and 17, and Component 4, which includes questions 14 and 16. For Part III, 4 components were extracted, 3 of which contain the highest positive values of the items. Component 1 includes questions 24, 25, 26, 29, and 30. Component 2 includes questions 21 and 22. Component 3 includes questions 23, 27, and 28.

Component Matrix[a]

	Component			
	1	2	3	4
Q1	.521			
Q2	.525			
Q3	.377		-.660	
Q4	.451			-.560
Q5	.410	.705		
Q6	.522			.512
Q7	.538	-.362		
Q8		.518		.492
Q9			.754	-.304
Q10	-.303	.590		

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 13: Components of Listening Test 4, Part I extracted by Factor Analysis

Component Matrix[a]

	Component			
	1	2	3	4
Q11	.699		.350	
Q12	.556			
Q13		.525		
Q14		-.571		.542
Q15	.543		-.490	
Q16	.391		.422	.462
Q17		.746		
Q18	.555			
Q19	.398	-.322		-.534
Q20	.490		.447	-.452

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 14: Components of Listening Test 4, Part II extracted by Factor Analysis

Component Matrix[a]

	Component			
	1	2	3	4
Q21	.300	.553		
Q22		.508		.483
Q23		.370	.485	
Q24	.720			
Q25	.558	-.359		
Q26	.449	-.399		-.380
Q27		-.449	.663	.619
Q28			.443	-.402
Q29	.562			
Q30	.429		-.618	

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Figure 15: Components of Listening Test 4, Part III extracted by Factor Analysis

Table 4 summarizes the number of components or sub-constructs that the four versions of the PSU-TEP listening were found to assess and the items that contribute to each sub-construct. However, as factor analysis can provide the number of sub-constructs and the items tapping into the same construct but not a description of abilities that each sub-construct assesses, test-taking processes which were collected by stimulated recall were analysed to explain the underlying construct (see Section 4.3 for the findings).

Listening Sub-components	Listening Test 1	Listening Test 2	Listening Test 3	Listening Test 4
Part I	**Items No.**	**Items No.**	**Items No.**	**Items No.**
Component 1	1, 3, 4, 7, 8	1, 2, 4, 7, 8, 10	1, 2, 4, 7, 8, 10	1, 2, 3, 4, 6, 7
Component 2	5, 9, 10	3, 5, 6, 9	3, 5, 6, 9	5, 8, 10
Component 3	2, 6			9
Part II				
Component 1	11, 14, 15, 16, 17, 18, 19, 20	11, 13, 14, 16, 17, 18, 19	13, 14, 17, 18, 19	11, 12, 15, 18, 19, 20
Component 2	12, 13	20	11, 12, 20	13, 17
Component 3		12, 15	15, 16	14, 16
Part III				
Component 1	22, 23, 24, 26, 27, 28, 29, 30	23, 26, 27, 29	22, 24, 27	24, 25, 26, 29, 30
Component 2	21, 25	28	21, 26, 28, 30	21, 22
Component 3		21, 22, 24, 25, 30	25	23, 27, 28
Component 4			23, 29	

Table 4: Sub-constructs measured by the PSU-TEP listening

4.2 What are the processes test-takers activated to complete the PSU-TEP listening?

To answer the 2nd research question, which asked what test-taking processes test-takers activated to complete the PSU-TEP listening test, stimulated recall data from 24 participants were transcribed and analysed. The results are presented in three sections, according to the three parts of the tests investigated, i.e. PSU-TEP Listening Test Parts I, II, and III. Following the literature review that suggests 3 types of processing behaviours test-takers perform for a listening test – cognitive processing, strategic processing, and test-wise strategies, the findings in each part are reported accordingly.

PSU-TEP Listening Test Part I

Of all the three types of listening processing –cognitive processing, strategic processing, and use of test-wise strategies, cognitive processing was reported the most frequently by the participants in all the four versions of the listening tests. Figure 16 shows that it contributes to 50% in Listening Test 1, 63% in Listening Tests 2 and 3, and 73% in Listening Test 4. While the activation of strategic processing contributes to one-third of the whole listening processes, the use of test-wise strategies counts for less than 20%.

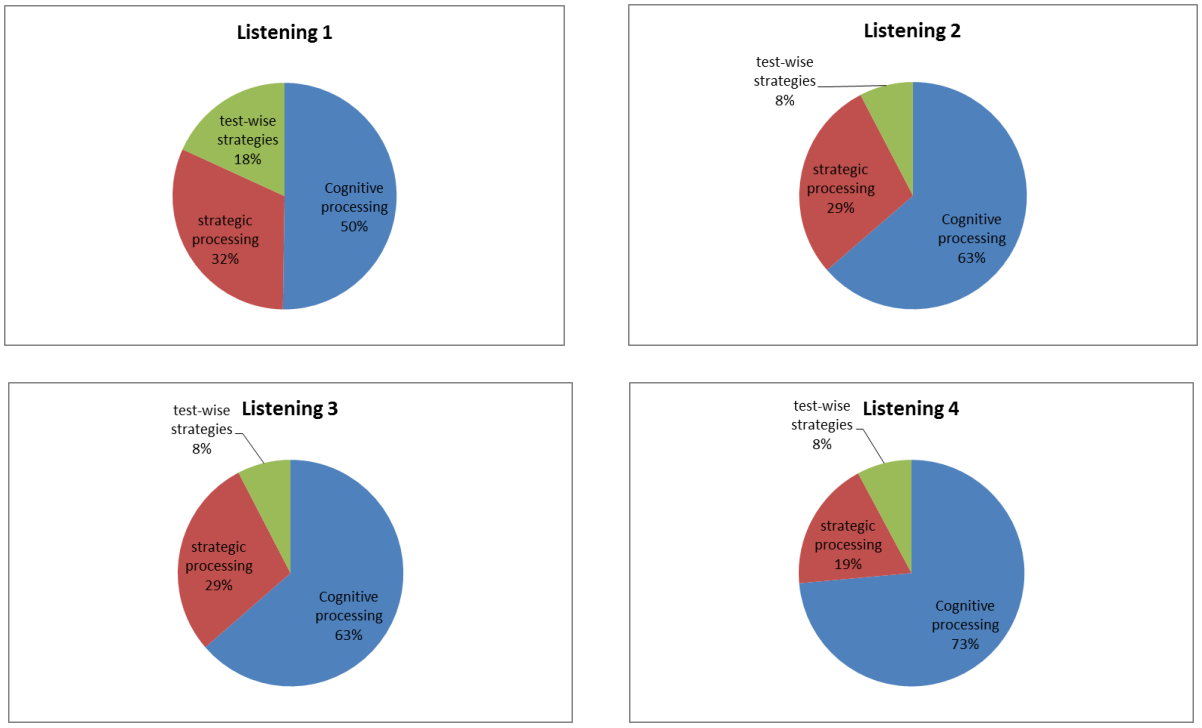

Figure 16: Test-taking processes activated to complete Part I of the 4 PSU-TEP listening tests

PSU-TEP Listening Part II

The similar pattern of test-taking processes found in the PSU-TEP Listening Test Part I was obtained for Part II (see Figure 17). That is, in order to complete the test, the participants relied mainly on cognitive processing. The participants reported activating twice as much cognitive

processing as strategic processing. Test-wise strategies, on the other hand, contribute to less than 10% of the entire processes activated.

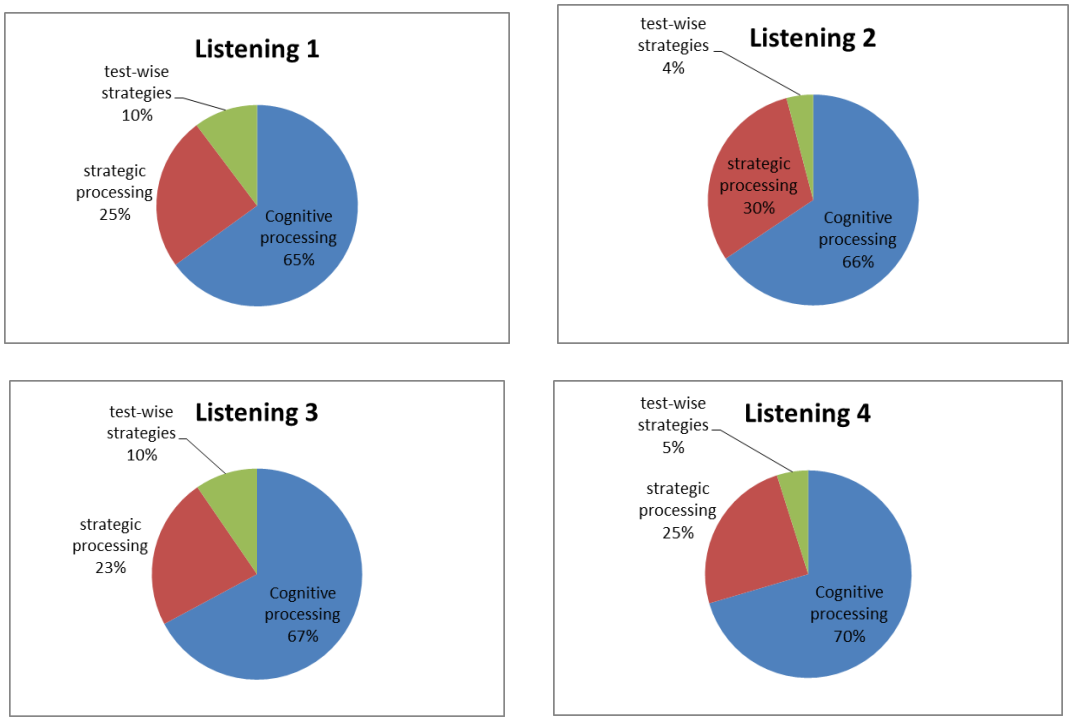

Figure 17: Test-taking processes activated to complete Part II of the 4 PSU-TEP listening tests

PSU-TEP Listening Part III

Compared to Parts I and II, Part III of the listening has offered a slightly different picture of the test-taking processes engaged in by the participants (see Figure 18). Although cognitive processing was activated the most frequently, its proportion was around half of the entire test-taking processes activated, less frequently than those activated in Parts I and II. Strategic processing was the second most frequently activated, which was similar to Parts I and II; however, the proportion of this processing for Part III was slightly higher. Test-wise strategies were found to be the least frequently used for part III, in line with what was found in Parts I and II.

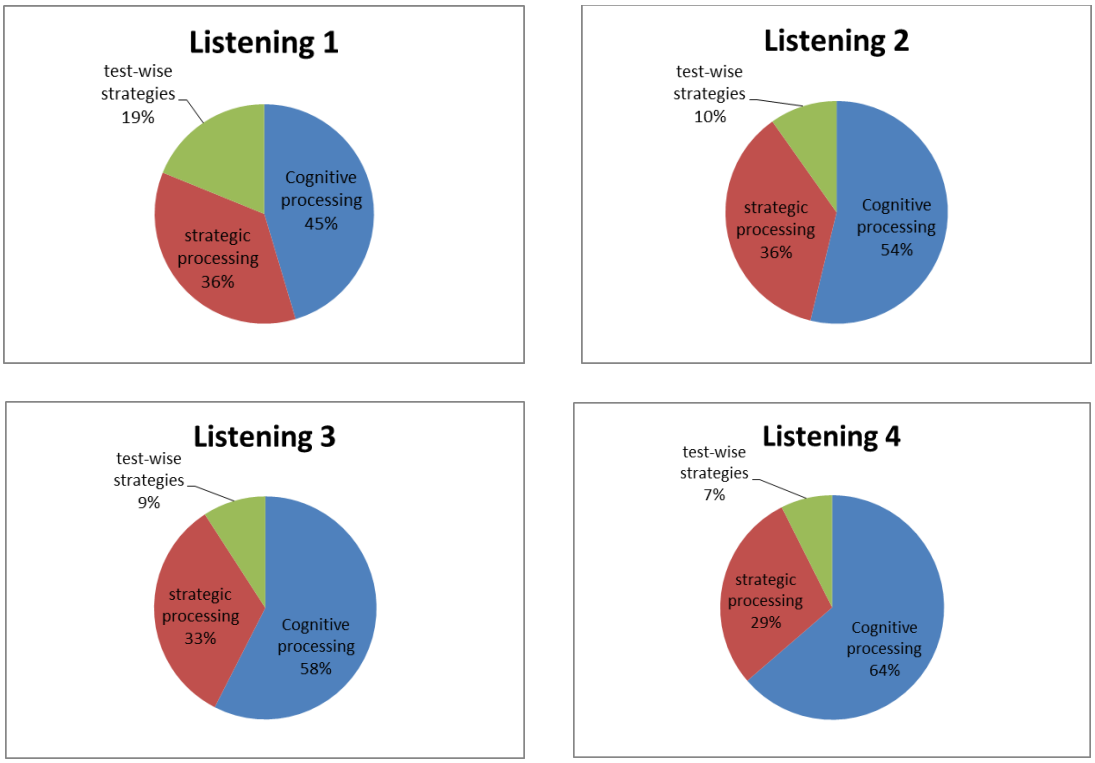

Figure 18: Test-taking processes activated to complete Part 3 of the 4 PSU-TEP listening tests

To sum up, the participants in this study reported relying on three types of test-taking processing –cognitive processing, strategic processing, and test-wise strategies. Cognitive processing was reported the most frequently activated processes in all the three parts of the four parallel versions of the listening tests, followed by strategic processing and test-wise strategies in a respective order. According to Weir (2005), test-taking processes show what abilities or source of knowledge test-takers rely on to complete the tasks. The most frequent use of cognitive processes in this study, contributing to about two-thirds of the entire processes activated, therefore, suggests that in an attempt to complete the test, the participants rely more on their their actual language abilities than their strategic processing and test-wise strategies, such as choice deletion, word matching, and blind guessing.

Cognitive processing is a key process that listeners activate to comprehend text texts in general. Field (2013) separates cognitive processing into two levels, the lower- and the higher-level processing and points out that the lower-level processing enable listeners to understand listening at a local level, meaning that they are able to understand separate ideas of the text they listen but not a global meaning or the main idea of the text. For comprehension to take place, Field states that listeners have to process listening text at the higher level. Therefore, in order to understand how listeners understand the texts they listened, cognitive processing is analysed in more details and presented in the next section.

Types of Cognitive processes activated in each listening part

As mentioned earlier, cognitive processes the test-takers activated to complete the test shows the source of knowledge and the language abilities that they used to answer the test questions. This section analyses types of cognitive processes the participants relied on to complete the tests. Following Field (2008), the cognitive processes in this study were separated into 6 groups. They are 1) acoustic-phonetic processing, 2) word decoding, 3) parsing, 4) semantic processing at the local level, 5) semantic processing at the global level, and 6) pragmatic processing. Based on the data obtained, none of the participants explicitly reported activating acoustic-phonetic processing. Therefore, only five types of cognitive processes are presented in this section. However, it is important to note that cognitive processing, according to Anderson (1985), is a linear process. What the listeners process at a lower stage is used as an input in the next stage of processing. The fact that the participants did not explicitly mention that they activated the acoustic-phonetic process, but engaged in word decoding could suggest that they all relied on the acoustic-phonetic processing with a high degree of automaticity so that they were not aware of doing it.

Literature in listening processing has classified listening processing into two different levels: the lower level and the higher level. The lower level, which involves acoustic-phonetic processing, word decoding, and parsing, is when the listeners process oral texts to understand different independent units of the texts such as words, phrases, or propositions and understand different parts of the text. However, listeners do not get the main point of the listening from the lower-level processing until they engage in the higher-level processing where they construct the meaning of what they listen. The higher level processing entails semantic processing at both the local and higher levels and pragmatic processing.

The analysis (see from Figure 19) showed that listening processing at the lower level which entails word decoding and parsing were mainly activated by the participants in all part of the tests. This suggests that while listening, the participants tried to identify word, phrases, propositions in the texts. Their activation of higher level processing, however, appeared to decrease by about one-thirds or half, especially in Listening Part III. There are two explanations to this phenomenon. One is that the majority of the test items only required text processing at the lower-level. The other is because the limitation in the test-takers' knowledge of English that prevent them from processing at the higher level in most cases. To clarify on this, point the next section compares test-taking processes operated by the test-takers at different levels of performing scores.

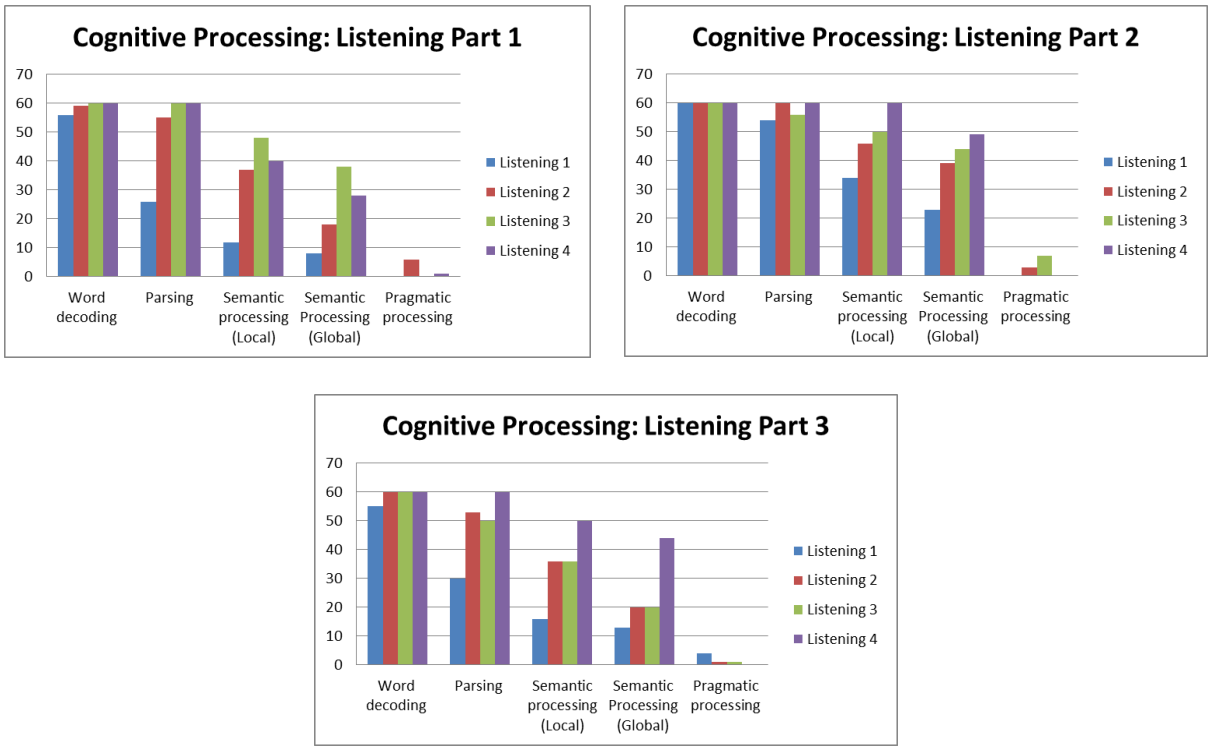

Figure 19: Cognitive processes activated to complete the 4 PSU-TEP listening tests

Test-taking processes activated by the participants with different performing levels

As suggested in the listening literature, to successfully understand listening texts, listeners have to engage in not only lower level processing but also the higher ones. In order to obtain a clearer picture, the test-taking processes activated by the participants across performing levels –low, average, and high, were compared. For this purpose, the total scores of the 24 participants participating in stimulated recall were ranked in a descending order, and the participants whose scores were at the top 5% were classified as high scoring participants, those at the bottom 5% were low scoring participants and those 5% at the middle were considered as moderate scoring participants.

The comparison (see Figure 20) shows that the participants with different performing levels activated similar types of processes and strategies. However, the proportions of processes

and strategies each group activated were found different. The high scoring participants activated cognitive processing the most frequently (40%), followed by the moderate- and low-scoring participants respectively (36% and 24%). The moderate-scoring participants relied the most heavily on strategies processing, followed by the high and low scoring participants (34% and 27%).

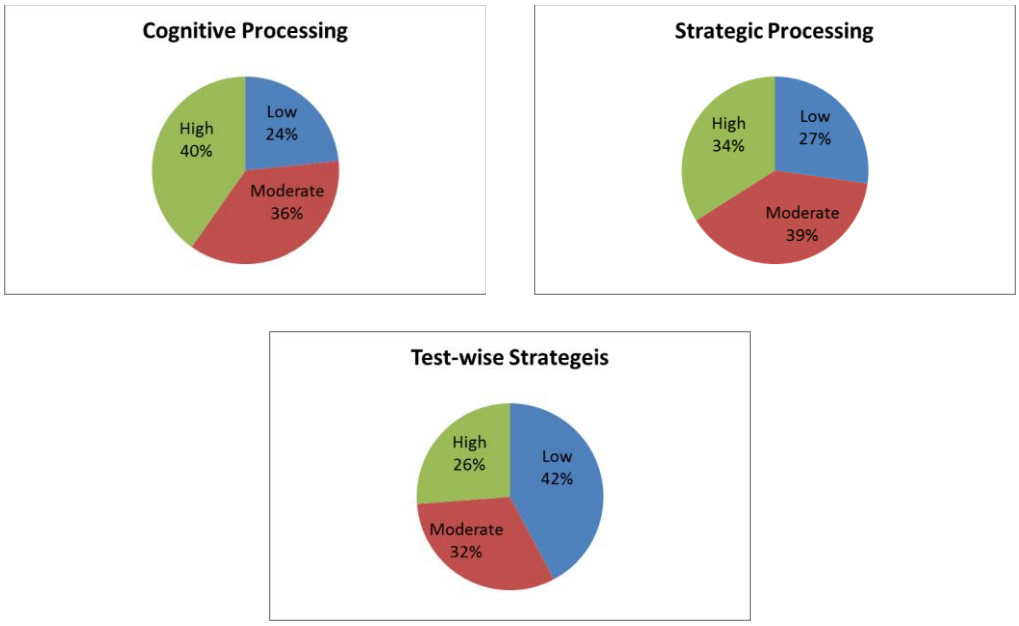

Figure 20: Proportion of test-taking processes the participants with different performing levels relied on

When considering the types of each process and strategy, it was found that the different numbers of the participants with different performing levels activated the three types of the test-taking processes at different rates (see Figure 21). The high-and moderate-scoring participants activated higher level of cognitive processing, semantic and pragmatic processing, almost three times higher than did the low-scoring participants. The fact that the higher level processing is important for the listeners to understand the main points of the texts may explain why the two former groups are more successful in test completion.

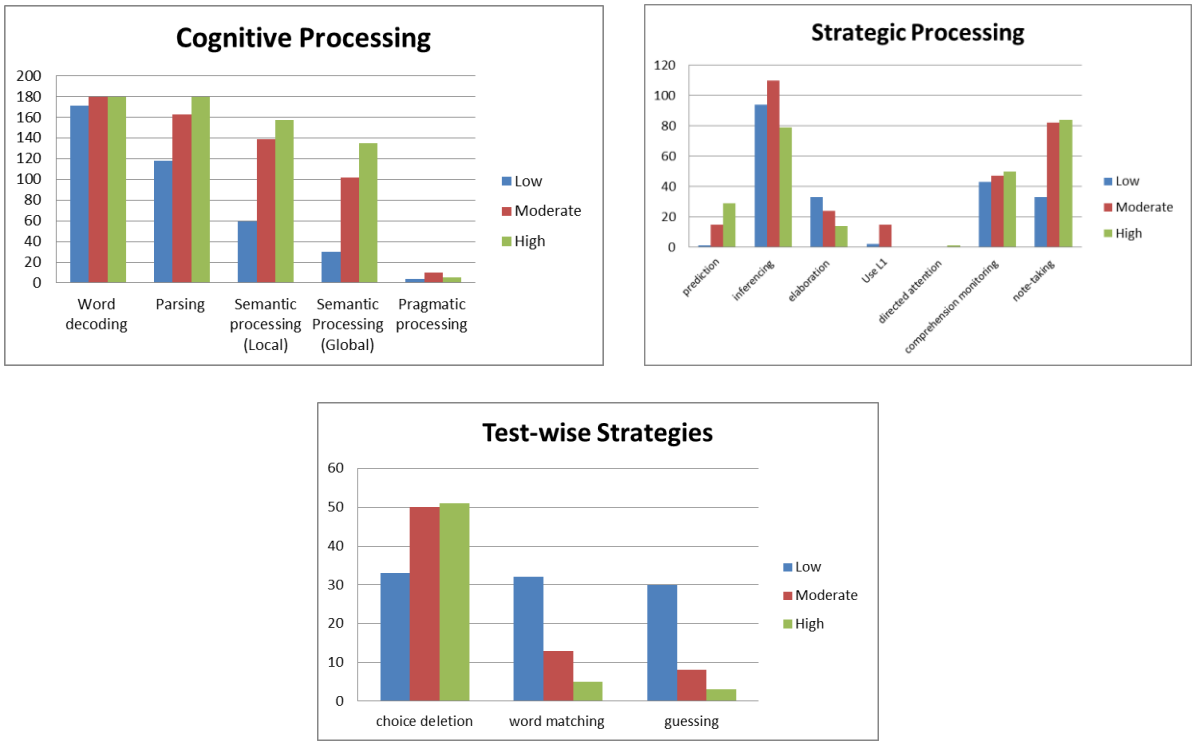

Figure 21: Test-taking processes compared between the participants with different performing levels

Regarding the activation of strategic processing, the analysis shows that three types of strategies which are popular among this group of test-takers are inferencing, note-taking, and comprehension monitoring. Among the three strategies, inferencing was used the most frequently by the moderate-scoring participants and the least by the high-scoring participants. Comprehension monitoring and note-taking, on the other hand, were used the most frequently by the high scoring group and the least by the low scoring group. The three types of test-wise strategies, choice deletion, word matching and blind guessing were used frequently by the low-scoring. Choice deletion, however, was more frequently used by the high-and moderate-scoring than did the low-scoring group. In contrast, the low-scoring group activated word matching and blind guessing about three times higher than did the moderate-and high-scoring participants.

4.3 The extent to which the PSU-TEP listening measured what it was aimed to measure?

This section addresses the 3rd research question, which asked the extent to which the listening component of the PSU-TEP measures what it is aimed to measure. To answer this question, the abilities assessed by the items are first presented and then the extent to which the test measures its construct is discussed. As presented in section 4.1, factor analysis was run to extract the number of components (sub-constructs) or types of abilities assessed by the test items. The analysis, nevertheless, did not explain the abilities performed to complete the test. Stimulated recall data on test-taking processes were, therefore, used for this purpose.

Listening abilities performed	**Listening Test 1**	**Listening Test 2**	**Listening Test 3**	**Listening Test 4**
Part I	**Items No.**	**Items No.**	**Items No.**	**Items No.**
Cognitive processing (acoustic phonetic processing, word decoding, paring) Strategic processing (inferencing)	1, 3, 4, 7, 8	1, 2, 4, 7, 8, 10	1, 2, 4, 7, 8, 10	1, 2, 3, 4, 6, 7
Cognitive processing (acoustic phonetic processing, word decoding, paring, semantic processing at local level) Strategic processing (inferencing) Test-wise strategies (choice deletion)	5, 9, 10	3, 5, 6, 9	3, 5, 6, 9	5, 8, 10
Cognitive processing (acoustic phonetic processing, word decoding) Strategic processing (inferencing) Test-wise strategies (choice deletion)	2, 6	-	-	9

Table 5: Abilities assessed by the test items in Listening Part I

For Part I of the test, the analysis (see Table 5) showed that three sets of processes and strategies were activated by the test-takers to get correct answers. First, the test-takers activated cognitive processing at the parsing level to get chunks of information and based on the

information they obtained, they inferred what the answer could be. One of the participants, for example, stated

> Excerpt 1
> The speaker said 'going to the hotel or going to buy something'. I heard 'coat'. I thought he wanted to wash it. And then I heard he asked if the hotel is near the downtown. I guess. I think it [the listening] is about asking for direction to go to the downtown to wash his coat. *[Participant 4, Listening 1]*

As can be seen from this excerpt, the participant parsed sound in speech stream to get words/chunks of information such as 'going to the hotel or going to buy something'. However, he could not tell what the speaker was talking about and based on the information he could parse, he had to make an inference as he reported that "I think it [the listening] is about asking for direction to go to the downtown to wash his coat.

The second group of abilities activated to answer the questions are cognitive processing, which includes acoustic phonetic processing, word decoding, paring, semantic processing at local level, strategic processing, which is inferencing, and test-wise strategy, which is choice deletion. This group is different from the first group in that the participants went from parsing to semantic processing at the local level and because of this, they could generate idea units in the text, making it easier for them to infer. In addition, there appears the use of test-wise strategy. That is they deleted the choices that they thought were not related to the story they were listening to. One of the participants, for example, recalled that

> Excerpt 2
> I heard the woman said 'I was looking for it in several places' 'I could not find it' and then she said 'ordered' and 'everywhere'. Here I understood that the woman was looking for something everywhere but she could not find it, so she decided to order it online. I don't think it was about looking for someone. *[Participant 5, Listening 2]*

This excerpt suggests that the text processing of this participant is quiet effective in the sense that she obtained most of key information to understand the text. However, her understanding of the text was not complete as she could not identify the item or subject in the conversation.

The third set of abilities used to answer the questions in Listening Part I are 2 cognitive processes –acoustic phonetic processing and word decoding, – strategic processing which is inferencing, and a test-wise strategy –choice deletion. In this case, it appears that with the ability to decode key words, the participants can identify some options which were not related to the listening and made an inference to get a correct answer. Some participants, for example, said

> Excerpt 3
> I didn't know clearly what it [the listening] was all about. I knew that someone was travelling. And then I heard 'from here', and 'ticket'. So, I guessed the woman was at the train station waiting to go somewhere, not at the post office or a university. *[Participant 1, Listening 1]*

> Excerpt 4
> I heard several words. I heard 'deep', 'water', 'be careful', ‘safety’, and 'swimming carefully'. So I thought this man worked as a lifeguard, not a professional swimmer or swimming instructor. *[Participant 2, Listening 2]*

Listening abilities performed	**Listening 1**	**Listening 2**	**Listening 3**	**Listening 4**
Part 2	**Items No.**	**Items No.**	**Items No.**	**Items No.**
Cognitive processing (acoustic phonetic processing, word decoding, paring) Strategic processing (inferencing, elaboration, comprehension monitoring, note-taking) Test-wise strategy (choice deletion)	11, 14, 15, 16, 17, 18, 19, 20	11, 13, 14, 16, 17, 18, 19	13, 14, 17, 18, 19	11, 12, 15, 18, 19, 20
Cognitive processing (acoustic phonetic processing, word decoding, paring, semantic processing at local and global levels) Strategic processing (inferencing, elaboration, comprehension monitoring, note-taking) Test-wise strategy (choice deletion)	12, 13	20	11, 12, 20	13, 17
Cognitive processing (acoustic phonetic processing, word decoding, paring,		12, 15	15, 16	14, 16

semantic processing at local and global levels, pragmatic processing)
Strategic processing (inferencing, note-taking)
Test-wise strategy (choice deletion)

Table 6: Abilities assessed by the test items in Listening Part II

Compared to Listening Part I, Listening Part II required more strategic processing. The participants who managed to answer the questions correctly reported engaging in more types of strategies. The strategies used were such as comprehension monitoring, elaboration, and note-taking. Overall, there appear three sets of abilities activated to answer the questions in this part. The first includes 3 cognitive processes (acoustic-phonetic processing, word-decoding, and parsing), 3 strategies (elaboration, comprehension monitoring, and note-taking), and one test-wise strategy (choice deletion). Some participants, for example, expressed

> Excerpt 5
> I heard 'vacation' and they said they are students. She said 'it is possible to work during summer'. And then they talked about ‘worked as something’. I assumed that they worked as a volunteer during their school vacation. I guess that was what they did. I spoke from my own experience- haha that was what I love to do. Um and later I heard and I noted 'the project she joined in New Mexico'. I was quite sure that it was about volunteering work during school holidays in the USA. Other choices were not related. *[Participant 1, Listening 1]*

Excerpt 5 shows that in addition to cognitive processing that enable this participant to obtain words and chunks of information, elaboration played an important role. The participant related what the speakers who are students in this context talked about to what he, who is also a student, likes to do during his school holiday to describe what the speakers did during their holiday and obtain a correct answer.

The second set of abilities include all the abilities that were activated in the first set; however, the cognitive processing goes beyond parsing to semantic processing at the local and global level which help the participants to clearly understand both the details and the main point of the text. One of the participants, for instance, explained

> Excerpt 6
> This item is quite clear. I understand almost everything. The woman said '22 April' is Earth Day', so I know that it is about Earth Day and the choice, which was 'a birthday' was wrong. I'm sure. I wrote it down. Then the woman said she should do something to save the environment and she said she would 'take a bus to work instead of driving', so I chose 'take a bus to work' for this item. Then the woman continued her conversation. She said she already bought a bus ticket and she would take a bus. She bought a bus ticket, not selling a bus ticket. The man said he did not have a ticket, but he wanted to join this campaign, so to the question 'what the man is going to do', I chose 'he is going to buy a bus ticket'. *[Participant 1, Listening 4]*

It is clear from this excerpt that this participant understood the text she was listening and she could decode the words 'Earth Day', which helped her understand the context of the talk. She could follow what the woman said to understand the purpose of her talk and also she could follow what the man said to know that he wanted to buy a bus ticket. This understanding shows she engaged in higher level of processing which includes semantic processing both at the local and the global levels. However, in addition to cognitive processing, this participant activated several strategies. She took notes in order to hold the information she could decode, and she inferred from the piece of the information she pared to understand what the man wanted to do. She used the choice deletion strategy to limit possible answers in each item.

The third set of abilities activated by the participants to get correct answers for Listening Part II is the cognitive processing at the pragmatic level and the use of inferencing, elaboration, and choice deletion. One of the participants, for example, expressed

> Excerpt 7
> Here I was trying to tell where they were talking to each other. I don't think it was at a university. It was more like they are talking about about getting an accommodation close to the university. I don't think they were talking at a supermarket. There was no such a noise that you hear when you go to supermarket. I think the woman was an office worker. From the way they talked to each other, I don't think they are friends. I heard something like the sound from a bus and then the woman said it was difficult to get the kind of the room that the man wanted with the amount that he could afford. I guess the amount was too

little. I heard 'live on my own'. I think the man did not want to share a room with anyone. *[Participant 2, Listening 2]*

Listening abilities performed	**Listening 1**	**Listening 2**	**Listening 3**	**Listening 4**
Part 3	**Items No.**	**Items No.**	**Items No.**	**Items No.**
Cognitive processing (acoustic phonetic processing, word decoding, paring, semantic processing at local and global levels) Strategic processing (inferencing, comprehension monitoring, note-taking) Test-wise strategy (choice deletion)	2, 3, 4, 6, 7, 8, 9, 10	23, 26, 27, 29	22, 24, 27	24, 25, 26, 29, 30
Cognitive processing (acoustic phonetic processing, word decoding, paring, semantic processing at local and global levels, pragmatic processing) Strategic processing (inferencing, elaboration, comprehension monitoring, note-taking) Test-wise strategy (choice deletion)	1, 5	28	21, 26, 28, 30	21, 22
Cognitive processing (acoustic phonetic processing, word decoding, paring, semantic processing at the local and global levels and pragmatic processing)		21, 22, 24, 25, 30	25	23, 27, 28
Cognitive processing (acoustic phonetic processing, word decoding)			23, 29	

Table 7: Abilities assessed by the test items in Listening Part III

For Listening Part II (see Table 7), it was found that the participants who got correct answers reported engaging in the higher level of cognitive processing which includes semantic and pragmatic processing and several types of strategies were reported being used. In addition, the participants reported using more strategies in this part than in Parts I and II. Overall, there appear 4 sets of abilities extracted by factor analysis. Stimulated recall data revealed that the first, which was performed in more than half of the items in this part, is semantic processing at the global level and the use of all types of strategies, i.e., prediction, inferencing, elaboration, use

of L1, directed attention, comprehension monitoring, and note-taking, and test-wise strategies, which is choice deletion. The activation of processes and strategies for this part was more complicated and interactive that that found for Parts I and II. One of the participants, for example, expressed

> Excerpt 8
> Here I was trying to predict what the listening was going to be about. I understand that the talk was about one company. I was listening for products that they sell. I don't know why it has to be a product –haha. But the speakers did not talk about any product. Later I heard service. I told myself that it was about a service, not a product. I was trying to catch what service. Then I heard 'we provide writing services'. So I knew that it was about writing. Then she talked about the quality of the work that they produced and she gave a code, it was a discount code, which I wrote down here. She explained that it was one-on-one service. I heard 'Paypal' and 'major credit cards', so I guessed she explained the channel for the customers to pay. Then the question asked what the listening was about. I didn't think it was about English tutoring, so I ignored this option. I was not sure between academic writing and ready-made research papers. I didn't think it was about editing and prove reading because I heard that 'we write for you'. I think they offer academic writing service, not research reporting writing because I feel like the target audience of this ad was university students. I thought I got it right. *[Participant 3, Listening 3]*

This excerpt shows that at the beginning the participant was trying to predict the content of the listening. When she heard 'a company', she predicted that it was about a product and she aimed to listen for the product that the company sold or produced. However, not any product was mentioned but a type of service. The participant then had to monitor her listening by reminding herself that the listening was about a service, not a product. Then the participant set to listen for the type of service the talk was about. She could decode writing services, so she understood that it was about writing. Then she used the words/phrases she segmented to understand that the customers can pay through Paypal and by credit cards. The data clearly showed that while the participant was engaging in cognitive processing and trying to process the text for comprehension, strategies were also activated to monitor and bridge gaps in listening.

The second set of abilities assessed by the items in part III includes all the abilities in the first set. However, in addition to those abilities, the participants appeared to rely heavily on their background knowledge to elaborate their understanding of the text. One of the participants, for example, indicated

> Excerpt 9
> To the question why teenagers like to backpack, I chose 'they want to travel before starting their career'. I heard the speaker said this. And from my own experience, I have seen many friends travelling before they applied for a job. I also want to do that. I think it is only period that we can do it because we've grown up and we have time, but because we don't have a lot of money, we have to go with backpacking. I think this is universal and it's quite true to everyone. *[Participant 3, Listening 1]*

Another set of abilities assessed by the items is the ability to process the text at the pragmatic level to identify the real purpose of the message. With the processing at the pragmatic processing, the participants appeared to have a good understanding of the text details and based on the details, they have to identify the purpose of the text in order to choose the right answers to the questions. One of the participants, for example, stated that

> Excerpt 10
> This question asked what the popular activity among the tourists in Singapore is. The speaker did not say this directly, but I was sure it was shopping. I think the purpose of this clip was to tell the tourist that there were many more to do in Singapore than shopping. Most people go to Singapore just to shop, but this clip was trying to present other activities that Singapore has offered to the tourist, such as sightseeing, and visiting a bird park. *[Participant 4, Listening 4]*

The last set of ability captured by the items in Listening Part III is the ability to decode the key words of the text. This ability was found to be useful when the question was straightforward and focused on details of the text such as 'What is not included in the list?". One of the participants, for example, explained

> Excerpt 11

> This item asked which one was not the characteristic of the service. I just followed the listening and took notes of the characteristics which were mentioned. Here the speaker presented the list quite clearly. She said first, second, another, next.... so I just followed this words. There were some details I did not understand, but I just ignored them because it was not important. *[Participant 5, Listening 3]*

PSU-TEP listening	Listening Input	No. of Items	Skills aimed to be tested
Part 1	8 short conversations on everyday topics (about 15-25 seconds each)	10	• Listening for main ideas (4-5 items) • Listening for specific details (5-6 items)
Part 2	3 longer conversations on a university life (about1 minute each)	10	• Listening for main ideas (3 items) • Listening for specific details (5 items) • Inferencing (2 items)
Part 3	2 talks (2-2.5 minutes each)	10	• Listening for main ideas (2 items) • Listening for specific details (6 items) • Inferencing (2 items)

Table 8: Specification of the PSU-TEP listening

To answer if the items used assess the test construct, Table 8 presents a rough version of the PSU-TEP listening test specification which includes the abilities that it is aimed to assess and the number of test items aimed to tap into the targeted abilities. According to the specification, the test was targeted at three listening abilities. They were 1) listening for main idea, 2) listen for specific details, and 3) inferencing. The analysis of stimulated recall data showed that overall, the construct of the tests was measured by the items used or the test measured the abilities specified in the test specification. However, stimulated recall data revealed that the ability the test-takers performed to get a correct answer to each item was not as straightforward as it was described in the specification. To get a correct answer to the items, test-takers appeared to execute several types of cognitive processes and strategies, some of which are not targeted in the test specification. Although the test, to some extent, tapped into the abilities it was aimed to,

there appeared a problem of construct underrepresentation or the test assesses the ability that is beyond what it is aimed to assess (Messick, 1995).

Listening comprehension processing, according to Field (2013) and Taylor and Geranpayeh (2011) involves two level of understand, local and global understanding. When listeners are required to provide information about the details of listening, it means that they are supposed to deal with local understanding of the text, and when they have to specify main idea, they are expected to present their global understand of the test. Cognitive processes that enable listeners to understand details are text processing at the lower level (acoustic-phonetic processing, word decoding, and parsing) and semantic processing at the local level. To determine main idea of the listening, listeners are expected to engage in semantic processing at the global level and pragmatic processing. The evidence from stimulated recall data that show the six types of cognitive processing were activated by the participants in order to choose appropriate answers to the questions in the three parts of the PSU-TEP listening test (see Tables 5-7) support that the abilities aimed to assess by the test, i.e. listening for main idea and listening for specific details) were truly assessed by the test.

Inferencing was one of the abilities that was heavily used by the participants to complete the listening test (see Tables 5-7). The participants relied on the information parsed from the listening to infer the answer to almost every question in the test. However, according to the specification, inferencing was targeted only in Parts II and III of the test. Stimulated recall data showed that to get correct answers to these questions test-takers made an inference based on the listening details that they were able to decode while listening. One of the participants, for example, reported that:

Excerpt 12

> I understood that it it was about a flight and a delayed flight. The passengers were called to go to Gate 12. So to answer the question 'where is the speaker?', I chose ‘at the airport' because of the information I had. *[Participant 4, Listening 4]*

Based on the findings presented, there appears a problem of construct underrepresentation.

Stimulated recall data, in addition, showed that the ability heavily activated in the test was listening for specific detail, even in the items that aimed to tap into listening for main idea. This is because listeners appeared to rely on listening details that they were able to segment to infer the main point of the listening text. In fact, listening for specific details was reported being activated to get an answer in all of the questions. It, in particular, supplied listening details for listeners to make inferences and to arrive at the main idea of the text. These three abilities (listening for details, listening for main idea, and inferencing), as evident in stimulated recall, were performed interactively. Most of the time, the test-takers activated them in combination in order to answer the questions in each item correctly. The attempt to separate them apart and assess them in isolation seems thus difficult. Excerpts from stimulated recall to show the interaction and interplay of these abilities are for example:

> Excerpt 13
> I thought they were at the library. I heard the man asked the lady why she came here. The woman said something about art history. I guess she studied art history. Then I thought the man said he had to write a paper but he lost his notes. So to the question that asked what they were doing. I assumed that they were discussing about how to solve the problem because later I heard the woman said ‘go to the reference room’. *[Participant 4, Listening 1]*

This excerpt shows that in order to answer the question, which was ‘what are the speakers doing?’, the test-takers performed several types of abilities. First, he had to identify where the speakers were. One piece of information that he could catch and help him to choose the answer correctly was when the man asked the lady why she came to the library. He then identified the

place where the conversation took place in the library. After that he listened for more information to answer the question which asked what the speakers were talking about. He decoded the words ‘art history’. As the words were spoken out by the woman, he assumed that the woman studied art history. Still he could not get the answer. He continued listening and was able to parse the text for more information. He got ‘I (the man) had to write a paper’ and ‘I (the man) lost his notes’. He assumed that this could be the problem of the man. To answer the question, this test-taker relied on all pieces of information he had obtained. This excerpt, in particular, shows that to understand the text and choose correct answers to the questions, the test-takers relied on more than one type of abilities. What he reported performing are 1) identifying the context of listening, 2) word decoding, 3) semantic processing, and 4) inferencing.

In conclusion, the results suggest that all the abilities aimed to assess were actually assessed by the tests used. However, these abilities were performed interactively by the test-takers, especially in the questions requiring them to infer or identify main idea. The list of an individual sub-skill that each item was aimed to assess in the test specification, may thus not well represent the ability being assessed by that particular item. The test, therefore, suffers construct underrepresentation.

In addition to the abilities specified in the test specification, it was found that the test-takers relied on test-wise strategies to get a correct answer to the items. They, for example, expressed that

> Excerpt 14
> This item asked what can be inferred about the woman. I was trying to delete some options that were not related to the listening. I heard 'New York', 'big city', 'teacher' I deleted Choice 2, which said she did not like her teacher. I thought it was wrong. It was not about her teacher, but it was about her. I deleted Choice 4 (she doesn't like teaching). I did not feel that it was about teaching. From the words I heard, I thought it was that she did not like big city because she talked about 'New York' and 'big city'. *[Participant 4, Listening 3]*

This excerpt shows that comprehension processing was performed at the lower level to decode words or chunks of information. With the information decoded and the use of test-wise strategies, this test-taker appeared to choose an answer to the question correctly. This, as Field (2013) points out, shows flaws of the test design, which could affect validity of the test. Messick (1989) considers this as a problem of construct-irrelevant because it taps into other abilities in test performance that contribute to success, the abilities which were not specified in the test specification.

Chapter 5 Conclusion

This section concludes the study by first providing its summary. This will include restating the research aims and the research questions, and summarizing the methodology and the main findings. Next, the contributions and implications of the study are discussed. The final section presents the limitations of the study and provides recommendations for future research.

5.1 Summary of the study

One major concern over the use of the PSU-TEP is its construct validity or the potential of the test to tap into the abilities it aims to tap. This is due to the fact that there appeared no test validation since the test was used. In this regard, this study set out to investigate the construct underlying the listening component of the test. Drawing upon the literature in language testing, this study conceptualized the construct or abilities assessed by a listening test as the cognitive and strategic processing test-takers engage in to complete a test and investigated the construct underlying the test. The following three research questions were asked.

1. What are the sub-components of listening ability captured by the PSU-TEP listening?
2. What are the test-taking processes test takers activated to complete the PSU-TEP listening test?
3. To what extent does the listening component of the PSU-TEP measure what it is aimed to measure?

The research data comprise 1) test scores from the test-takers of the 4 tests administered in 2017 and 2) verbal report data of 24 undergraduate students in different disciplines. These 24 participants were invited to participate in stimulated recall organized by the researcher on a one-

on-one basis. Factor analysis was run on the test scores obtained to identify the number of sub-constructs captured by the test items. To explain the abilities performed for each sub-construct stimulated data were used.

Factor analysis has identified different numbers of sub-constructs of each part of the test, For Part I, 2-3 sub-constructs have been revealed, varying according to the test versions. For Parts II and III, 3-4 sub-constructs were identified. The abilities assessed by these sub-constructs, as revealed in the stimulated recall data, include cognitive processing at the lower/local and higher/ global levels. At the lower/local level, listeners could segment pieces of information from speech stream and made an inference to understand specific details. At the higher/global levels, listeners engaged in semantic processing to understand different ideas in the text and the relation between those ideas to under the main point. Occasionally, pragmatic processing was activated to identify the real message that the speakers wanted to convey. This is especially when the purpose was not clearly mentioned. The success of cognitive processing at each level was found to facility by the use of some strategies. For Part I, for example, listeners engaged in word-decoding and parsing which are the processes at the lower level. However, to understand the details of the text, they needed inferencing. Likewise, at the higher level, they appeared to rely on inferencing, elaboration, prediction, and comprehension monitoring while engaging in semantic processing to understand the main point of the texts. The processes and strategies used by the participants who managed to answer the questions correctly were found to be highly interactive (see Excerpt 8).

To answer the research question if the PSU-TEP listening component measures its construct, the analysis showed that overall it does. According to the test specification, the PSU-TEP listening is aimed to measure three listening abilities. They are 1) listening for specific

details, 2) listening for main idea, and 3) inferencing. The results showed that the test-takers engaged in cognitive processing at the lower/local level enables. This enables them to understand specific details of the text. The test-takers, in addition, were found to activate cognitive processing at the higher/ global level, which assists them to obtain the main point of the text. For the test-takers to clearly understand both specific details and main idea of the text, they appeared to activate several types of strategies, including inferencing, elaboration, comprehension monitoring, note-taking and prediction. Two items of Listening Parts II and III were set to assess the ability to make an inference. Stimulated recall data, in fact, showed that inferencing was activated all the time when the test-takers listened to fulfill gaps in the test-takers knowledge and when the test-takers made an inference, they had to rely on other abilities, such as word decoding and parsing. Taking these findings into account, it may not be true to specify individual items to assess individual abilities as in real-time processing, these abilities work interactively to assist the test-takers to comprehend listening texts. Based on the findings, this study, therefore, recommend to conceptualize the abilities assessed by the test used. Instead of specifying what individual items are aimed to assess, the construct should be explained in terms of global understanding and local understanding. In this manner, other abilities performed to complete the test, such as comprehension monitoring, prediction, note-taking, will be described as part of the construct. Otherwise, the test can be considered construct under-representation.

5.2 Contributions of the study

This study has implications for test developers who are responsible for designing a listening test. A multiple-choice item is one of the test formats which are commonly used in a large-scale test. This is because its practicality in terms of scoring. Although it could tap into the abilities that the

test was aimed to, this item type captured other abilities which do not exist in non-taking situation, i.e. test-wise strategies. It is therefore advised to include different types of item format in a listening test.

Scholars such as Bachman and Palmer (2010), Weir (2005), Taylor and Geranpayeh, (2011) and Vandergrift and Goh, (2012) suggest that metacognitive strategies, or strategy used to manage or overlook comprehension processing are important and should be captured by a test. This is because they are important for communication beyond a testing situation. However, it appears that not many item types could tap into this ability. In this study, it was found that when listening to the listening input with different lengths, the test-takers activated different sets of processes and strategies. With the input of less than one minute the test-takers appeared to rely heavily on word decoding, parsing and inferncing. In a longer input text, i.e. 1.5 minutes the participants appeared to engage in prediction, directed attention, comprehension monitoring, note-taking, all of which are important in real-life listening. Therefore it is important for a listening test that is aimed to tap into listening abilities performed in real-life listening to include an input text of 1.5-2 minutes in length.

5.3 Implications of the study

This study sets off to investigate the construct validity of the PSU-TEP listening test by analyzing test scores from the test takers and stimulated recall data from a group of 24 test-takers completing the test of one-on-one basis. However, prior to the investigation of construct validity, the reliability of the test used was checked, the analysis showed the reliability coefficients of 0.51-0.67. The reliability of a test expresses the quality of the test in drawing conclusions about the quality of the test. The coefficient value of 0 shows the test is not reliable or it does not have

consistency in measuring test-takers' abilities and, as a result, the test is not useful. On the other hand, the coefficient value of 1 indicates is perfectly reliable with no error in measurement. Tests that are used for high-stakes purpose such as a university admission or exit-exam, as Roever and Phakiti (2018) expressed, should have the reliability coefficient of 0.90 or at least 0.80. The reliability coefficients of 0.51-0.67, as found in this study, are lower than the acceptable level. It is therefore important for the test designers and developers to improve test items and increase the reliability of the test.

Although a multiple-choice format used in the test is appropriate for a large number of test-takers, the results showed that the format provides an opportunity for test-takers to use test-wise strategies, such as choice deletion and lexical matching. The test appeared to tap into the abilities that are not actually performed in non-testing situations. In particular, it induces response processes that do not exist in non-testing contexts. It is therefore recommended for the test developers to include different types of item format. This is to reduce the problem of the use of test-wise strategies. In fact, testing scholars (Alderson et al., 1995; Hughes, 2003; Fulcher & Davidson, 2007) recommend that different test formats should be included in one language test. This is in order to fully tap into language abilities defined in the test construct and to control over the test-method effect, the influence of testing format on test scores.

The results showed that although the abilities mentioned in the test specification were performed by the test-takers to get correct answers, there appeared other abilities which were not specified but performed in an attempt to get a correct answer, such as comprehension monitoring, note-taking, and elaboration. These strategies, as indicated by Vandergrift and Goh (2012) are essential in listening in real-life situations. To better represent the abilities performed

to complete the test, it is therefore recommended to re-conceptualize the construct underlying, taking into account both cognitive and strategic processing.

5.4 Limitations and future research

Despite having been carefully designed, this study has some limitations. One is related to the item type. The study was limited to the multiple-choice listening items with no question preview. The listening test where the test-takers are allowed to see the questions before listening may provide different results. Second, to extract the components of the listening-subskills assessed by the test, factor analysis was run on the sets of quantitative data obtained in the year of 2017, when the study was carried out. Although, each part of the test was independently analyzed to guarantee the sufficient number of responses to run the analysis, for a better picture of the listening sub-components embedded in the listening test, a bigger set of test data may need to be added. In addition, the researcher organized a stimulated recall with 24 participants on one-on-one basics, and the data obtained were used to describe the sub-constructs of the test. Although the data obtained were insightful, the generalization of the findings could be limited due to a small number of the participants. To provide a clearer picture of the sub-constructs that a test assesses, after verbal data, an exploratory mixed method design where a questionnaire is used to explore processes and strategies activated by a bigger group of test-takers is recommended in future studies.

List of References

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

Anderson, J. R. (1985). *Cognitive psychology and its implications*. New York: Worth Publishers.

Anderson, N.J., Bachman, L., Perkins, K. & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing, 8*(1), 41-66.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Ernst Klett Sprachen.

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

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

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

Chang, A., & 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*, 575-586.

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

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in phychological tests. Phychological *Bulletin, 53*, 281-302.

Ericsson, K. A., & Simon, H., A. (1993). *Protocol analysis: Verbal reports as data*. USA: Massachusetts Institute of Technology.

Færch, C., & Kasper, G. (1986). The role of comprehension in second language learning. *Applied Linguistics, 7*(3), 257-274.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77-151). UK: MPG Books Group.

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

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

Goh, C. C. M. (2002). Exploring listening comprehension tactics and their interaction patterns. *System, 30*, 185-206.

Graham, S., Santos, D., & Vanderplank, R. (2008). Listening comprehension and strategy use: A longitudinal exploration. *System, 36*(1), 52-68.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Leonard, K. (2019). Examining the relationship between decoding and comprehension in L2 listening. *System, 84*, 1-12.

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

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.

O'Malley, J. M., Chamot, A. U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*(4), 418-437.

Ren, W. (2013). A longitudinal investigation into L2 learners' cognitive processes during study abroad. *Applied Linguistics*, 1-21.

Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Great Britain: Pearson Education Limited.

Rupp, A. A., and Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language testing, 23*(4), 441-474.

Rubin, J. (1981). Study of cognitive processes in second language learning. *Applied Linguistics, 11*(2), 117-131.

Rukthong, A. & Brunfaut, T. (2019). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*, (1), 31-53.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse close test. *Language Testing, 14*(2), 214-231.

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT™ (SSTiBT): Test-takers' reported strategic behaviours (TOEFL iBT™ research report)*. Retrieved from Princeton, NJ:

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.

Tanewong, S (2018). Metacognitive pedagogical sequence for less-proficient Thai EFL listeners: A comparative investigation. *RELC*, 1-18

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

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

Weir, C. J. (2005). *Language teasint and validation: An evidence-based approach*. Great Britain: Antony Rowe Ttd,.

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System, 65*, 139-150.

Yeldham, M. & Gruba, P. (2014). Toward an instructional approach to developing interactive second language listening. *Language Teaching Research, 18*(1) 33 –53.