VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF LANGUAGES AND INTERNATIONAL STUDIES
******

NGUYỄN THỊ QUỲNH YẾN

SUMMARY OF DOCTORAL DISSERTATION

AN INVESTIGATION INTO THE CUT-SCORE VALIDITY OF THE VSTEP.3-5 LISTENING TEST

MAJOR: ENGLISH LANGUAGE TEACHING METHODOLOGY

CODE: 62140111

HANOI, 2017

The results that represent the basis for achieving this doctoral dissertation were all obtained at the University of Languages and International Studies, Vietnam National University, Hanoi.

SUPERVISORS

1. Prof. Nguyễn Hòa
2. Prof. Fred Davidson

EXAMINATION BOARD

Examiner 1:……
Examiner 2:……
Examiner 3:……

This doctoral dissertation will be defended at the national level in ……. at……

This doctoral dissertation can be found at:

- National Library of Vietnam
- Library and Information Center, Vietnam National University, Hanoi

THESIS-RELATED PUBLICATIONS

1. Nguyễn Thị Quỳnh Yến. (2015). Đảm bảo tính công bằng trong kiểm tra kỹ năng nghe. *Tạp chí Ngôn ngữ học và đời sống*, số 12 (242):36-39. ISSN 0868-3409. NXB Chính trị Quốc gia.
2. Nguyễn Thị Quỳnh Yến. (2017). Một số vấn đề cần xem xét trong đánh giá tính giá trị của các điểm cắt giữa các bậc năng lực cho bài thi đánh giá năng lực tiếng Anh theo khung năng lực ngoại ngữ 6 bậc dành cho Việt Nam. *Kỷ yếu hội thảo quốc gia 2017- Nghiên cứu và giảng dạy Ngoại ngữ-Ngôn ngữ-Quốc tế học tại Việt Nam*. Trang 612-618. ISBN: 978-604-8164-1. NXB Đại học Quốc gia Hà Nội.
3. Nguyễn Thị Quỳnh Yến. (2017). Building a validity argument for the Vietnamese Standardized Test of English Proficiency (VSTEP.3-5). *Kỷ yếu hội thảo quốc gia dành cho Học viên cao học và nghiên cứu sinh lần thứ nhất*. Trang 705-712. ISBN: 978-604-62-9306-4. NXB Đại học Quốc gia Hà Nội.

TABLE OF CONTENTS

ABSTRACT

Standard setting is an important phase in the development of an exam program, especially for a high-stakes test. Standard setting studies are designed to identify reasonable cut scores and to provide backing for this choice of cut scores. This present study is aimed at investigating the cut-score validity of the VSTEP.3-5 listening test administered at the University of Languages and International Studies, Vietnam National University, Hanoi. The study adopts the current argument-based validation approach with a focus on three main inferences constructing the validity argument. They are (1) test tasks, (2) accuracy and precision and (3) cut scores. The argument is that in order for the interpretations and uses of the cut-scores of the VSTEP.3-5 listening test to be valid, the test tasks first need to be properly designed in accordance with the characteristics specified in the specification. Second, the listening test scores must be sufficiently reliable so as to accurately reflect test-takers' listening proficiency. Third, the cut scores established for the VSTEP.3-5 listening test are useful for making decisions about test takers' English listening competency. In this study, both qualitative and quantitative methods are combined and structured to back for and against the assumptions in each of these three inferences. The results contribute both positive and negative attributes to the validity argument for the VSTEP.3-5 listening test.

CHAPTER I

INTRODUCTION

This chapter is to introduce the topic of the study and present the main reasons for choosing it. After that, the chapter takes a close look at some questions that are going to be addressed within the scope of the study. A brief overview of the organization of the study will close the chapter.

1. Statement of the problem

Establishing cut scores for a test has been considered an important and practical aspect of standard setting. If the cut scores are not appropriately set, the results of the assessment could come into question. For that reason, setting cut scores is a critical component of the test development process. In Kane's (2006) recent discussion for test validation, besides emphasizing the importance of carefully defining the selected cut scores, he highlights the evaluation of the reasonableness of the cut scores. According to Kane (2006), setting cut scores is a complex endeavor, but validating the cut scores is even more difficult.

The VSTEP.3-5 test is the first ever-standardized test of English officially released by the Ministry of Education and Training, Vietnam on 11th March 2015. The test aims at measuring English ability across a broad language proficiency continuum from level 3 to level 5 (equivalent to B1 - C1 CEFR levels). The VSTEP.3-5 test is a high-stakes test since the uses of the VSTEP.3-5 test and the decisions that are made from the test cut scores have important consequences for the stakeholders. Besides, since the VSTEP.3-5 test is a newly developed test, like other high-stakes tests such as TOEFL, IELTS, PTE, or Cambridge Tests, in order to gain credibility and defensibility, more research needs to be conducted about the test in general and the validity of the VSTEP.3-5 cut scores in particular. However, so far, there have been few studies on the VSTEP.3-5 test and there is no validation research on the cut scores of the test.

All of the reasons mentioned above have intrigued the author of this doctoral thesis to conduct a validation research on the cut scores of the VSTEP.3-5 listening test by using validity argument-based model proposed by Kane (2013). With the deeply-rooted desire to develop a good proficiency listening test in Vietnam, this research is expected to bring the author a profound insight into this specific area of interest, specifically using an argument-based validation approach in language testing, for her future professional development.

2. Purpose of the study

As mentioned, since the VSTEP.3-5 test is a newly developed high-stakes test, the need to standardize it is imperative. Thus, this doctoral research is conducted as an ongoing attempt in building a systematic,

transparent and defensible body of validity argumentation for the VSTEP.3-5 test in general and its listening component in particular. By adopting the argument-based approach recommended by Kane (2013), the study aims at investigating the cut-score validity of the VSTEP.3-5 listening test.

3. Scope of the study

Since the VSTEP.3-5 tests are administered by some other educational institutions throughout Vietnam, this study only focuses on the validation of the interpretation and uses of the cut-scores of the VSTEP.3-5 listening tests that are administered in ULIS-VNU. The validity argument in support of the interpretations and uses of the VSTEP.3-5 listening test cut-scores includes three propositions that will be taken into consideration in this study. Each proposition is labelled with a name for ease of reference in the discussion.

1. Test tasks: The test tasks are properly designed in accordance with the characteristics specified in the test specification.
2. Accuracy and Precision: The observed VSTEP.3-5 listening test scores show that test is reliable in measuring test-takers' proficiency.
3. Cut scores: The cut scores established for the VSTEP.3-5 listening test are useful for making decisions about test takers' English listening competency.

4. Statement of research questions

From the propositions above, the research questions of this dissertation are as follows:

1. To what extent were the test tasks of the VSTEP.3-5 listening test properly designed in accordance with the characteristics specified in the specification?
2. To what extent were the VSTEP.3-5 listening test scores reliable in measuring the test takers' English proficiency?
3. To what extent were the cut scores of the VSTEP.3-5 listening test properly established?

5. Organization of the study

The study consists of six chapters as follows:

Chapter I is aimed at introducing the topic of the study and presenting the main reasons for the author to implement this project.

Chapter II is to provide profound theoretical and empirical background with a critical discussion on the relevant concepts, models, or theories for the study.

Chapter III describes the context of the study with information on the VSTEP test in general and the VSTEP.3-5 listening test in particular.

Chapter IV presents how the study is conducted together with a review on each selected methodology.

Chapter V presents the results of the study and discusses these results.

Chapter VI has three aims. First, it summarizes the main findings of the study. Second, it specifies the limitations of the study. Finally, it suggests some directions for future studies.

CHAPTER II
LITERATURE REVIEW

This chapter deals with validation in language testing and assessment, testing listening and standard setting.The first part of the chapter works with validation in language testing and assessment. The second part of the chapter is on testing listening. The third part of the chapter focuses on the issues related to standard setting.

1. Validation in language testing and assessment

Validity is considered one of the most important concepts in psychometrics, but as Sireci (2009) states validity has taken on many different meanings over the years. Kane (2009, 2013) summarizes aspects of the current view as follows: First, validity is a matter of degree, and it may change over time as the interpretations/uses develop and as new evidence accumulates. Second, it is the interpretations and uses of test scores that are validated, and not the tests themselves. It can be quite reasonable to talk about the validity of a test if only an interpretation or use has already been adopted explicitly or implicitly. Third, the evidence needed for validation depends on the interpretations and uses. Therefore, different interpretations or uses will require different kinds and different amount of evidence for their validation.

According to Chapelle (1999) and Kane (2001, 2006), there are two main approaches in validation studies based on the history of the validity concept in language testing and assessment, namely evidence-based approach and argument-based approach. Evidence-based approach sees the final results of validation as an accumulation of evidence from different aspects of validity (Chapelle, 1999: 258). The main limitation in the criterion model is the difficulty in obtaining an adequate criterion and at some point, the criterion has to be validated. By the late 1980s, the argument-based approach to validation was born. The latest argument-based validation model receiving a lot of support is proposed by Kane (2006). Kane (2006) points out that the argument-based validation approach reflects the general principles inherent in construct validity without an emphasis on formal theories. The interpretive argument is to provide a clear statement of the inferences and assumptions inherent in the proposed interpretations and uses of test results, and these inferences and assumptions are to be evaluated in a series of analyses and empirical studies. The validity argument as a whole requires the integration of different kinds of evidence from different sources.

2. Testing listening

According to Buck (2001), listening comprehension is a complicated process and if we want to measure it, we must understand how that process works. The thing we are trying to measure is called a construct and our test will be useful and valid if it measures the right construct. Construct is simply defined as the thing that we are trying to measure (Buck, 2001:1). Linn and Miller (2005:78) state that "when we interpret assessment results as a measure of a particular construct, we are implying that there is such a construct, that differs from other constructs, and that the results provide a measure of the construct that is little influenced by extraneous factors".

According to Buck (2001), listening construct can be defined in terms of the tasks the listener can do when we are interested in what the test-takers can do, and under what circumstances they can do it. In order to define listening construct in terms of tasks, the first thing we have to come up with is a suitable list of target-language use tasks that the listeners should be able to perform. Then, it is necessary to produce a set of test tasks that replicate these real-world tasks. Buck (2001) suggest a list of task characteristics that can be used to compare test tasks with real-world tasks, which is adapted from the list given by Bachman and Palmer (1996) (table 2).

Characteristics of setting: the physical circumstances under which the listening takes place.
- **Physical characteristics:** the actual place, background noise, live or recorded, etc.
- **Participants:** the people around, including test administrators.
- **Time of task:** Time of day, whether the listeners are fresh or fatigued.

Characteristics of the test rubric: the characteristics that provide the structure for the test, and how the test-takers are to proceed. These have to be made explicit in the test, but are usually implicit in language use.
- **Instructions:** instructions, details or procedures, as well as purpose of listening, including the language used, and whether they are oral or written.
- **Structure:** structure of the test, how the parts of the test are put together, including the number of listening passages, whether the passages are repeated, their order, the number of items per passage, etc.
- **Time allotment:** the time allowed for each task.
- **Scoring method:** how the tasks are scored including criteria for correctness, steps used for scoring or rating, and how the item scores are combined into the test score.

Characteristics of the input: the listening passages and other accompanying material.

Format: whether passages are spoken or recorded, their length, etc.
Language of input: including phonology, grammar, lexis, textual, functional and sociolinguistic knowledge.
Topical knowledge: the cultural, schematic or general world knowledge necessary to understand the passages.

Characteristics of the expected response: the response expected from the test-taker to the task.
Format: selected or constructed, the length, the form it will take, time available.
Language of expected response: for constructed responses whether in the L1 or the L2, criteria for correctness, etc.

Relationship between the input and response:
Directness of relationship: the degree to which the response can be based primarily on the language of the passage, or whether it is necessary to use information from the context, or apply background knowledge.

Table 2: A framework for defining listening task characteristics (Buck, 2001)

The framework of task characteristics given in table 2 is intended to function as a checklist for comparing test tasks with target-language use tasks. According to Bachman and Palmer (1996), this framework is a means of investigating task authenticity, as well as an aid to the development of new task.

3. Standard setting for an English proficiency test

Kane (2001:54) states that standard setting is a complex endeavor, but to validate standard setting is even more difficult. According to the Standards (AERA et al., 1999:9), validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests". In the context of standard setting, since there are no "gold standards" and "true cut scores", to validate established cut scores means to provide evidence in support of the plausibility and appropriateness of the proposed cut scores interpretations, their credibility and defensibility (Kane et al., 1999). Evaluation of standard setting is a multifaceted endeavor with many potential sources of evaluation information. Validity evidence in standard setting tends to come from three places: procedural evidence, internal consistency evidence, and external evidence (Pitoniak, 2003; Hambleton & Pitoniak, 2006; Kane, 2001, 2006; Cizek & Bunch, 2007). Cizek & Bunch (2007) provide a complete listing of possible evaluation elements in the following table.

Evaluation Element	***Description***
Procedural	
Explicitness	The degree to which the standard setting purposes and processes were clearly and explicitly articulated a priori
Practicability	The ease of implementation of the procedures and data analysis; the degree to which procedures are credible and interpretable to relevant audiences
Implementation	The degree to which the following procedures were reasonable and systematically and rigorously conducted: selection and training of participants, definition of the performance standard, and data collection
Feedback	The extent to which participants have confidence in the process and in resulting cut score(s)
Documentation	The extent to which features of the study are reviewed and documented for evaluation and communication purposes
Internal	
Consistency within method	The precision of the estimate of the cut score(s)
Intraparticipant Consistency	The degree to which a participant is able to provide ratings that are consistent with the empirical item difficulties, and the degree to which ratings change across rounds.
Interparticipant Consistency	The consistency of items ratings and cut scores across participants
Decision Consistency	The extent to which repeated application of the identified cut score(s) would yield consistent classifications of examinees
Other measures	The consistency of cut scores across item types, content areas, and cognitive processes
External	
Comparisons to Other standard setting methods	The agreement of cut scores across replications using other standard setting methods
Comparisons to other sources of information	The relationship between decisions made using the test to other relevant criteria (e.g., grades, performance on tests measuring similar constructs, etc.)
Reasonableness of cut scores	The extent to which cut score recommendations are feasible or realistic (including pass/fail rates and differential impact on relevant subgroups)

CHAPTER III

CONTEXT OF THE STUDY

In order to bring light on the current context of the study, this chapter will briefly explain the administrative situation of the VSTEP.3-5 test in Vietnam. This is followed by a description of the administration of the VSTEP.3-5 listening test in ULIS – VNU and of the context surrounding the test setting.

1. About the VTEP.3-.5 test

The VSTEP.3-5 test is currently allowed by the Ministry of Education and Training to be administered by ten different institutions. They are the University of Foreign Language Studies, the University of Danang; Hanoi University; Hue College of Foreign Languages, Hue University; *Ho Chi Minh City University* of Education; Thai Nguyen University; Hanoi National University of Education; Vinh University; Can Tho University; SEAMEO RETRAC; and the University of Languages and International Studies, Vietnam National University. Based on the format and the test specification of the VSTEP.3-5 test released by the Ministry of Education and Training, some institutions have built their own test bank or some others choose to take the test forms supplied by NFL2020.

The VSTEP test comprises of four subtests. Scores on each component will be used to assess each individual language skill of test-takers separately. The final score will be the average of the scores of four components. The cut scores for the VSTEP.3-5 test were established by the Angoff standard setting method. The conversion table from the raw scores to the converted scores for the results of different subtests were established based on the results of the model VSTEP.3-5 test.

2. Description of the VSTEP.3-5 listening test

The VSTEP.3-5 Listening test is a component of the VSTEP.3-5 test. The VSTEP.3-5 listening test comprises of 3 parts with 35 multiple choice questions. All the recording is played once only. In the first part, test takers hear 8 short recordings. There is one question following each recording. In the second part, test takers hear three conversations. There are four questions for each conversation. In the third part, test takers hear three talks or lectures. There are five questions for each talk or lecture. At the end of the Listening test, test takers are given 7 minutes to copy their answers to an answer sheet. The language activities in the VSTEP.3-5 listening test are contextualized with domains. These may themselves be diverse, but for most practical purposes in relation to language learning, they are broadly classified as four folds: the public domain, the personal domain, the educational domain and the occupational domain.

The cut scores for the VSTEP.3-5 listening test were established by the Angoff standard setting method for the model VSTEP.3-5 listening test. These cut scores are preset and applied for all of the test forms which are supposed to be built strictly in accordance with the descriptions in the specification. The conversion table for the cut scores on the raw score scales to the converted score scales is as follows:

Level	Raw score	Converted score
Unrated	0- 11	0 – 3.5
3	12 – 19	4 – 5.5
4	20 – 27	6 – 8
5	28 – 35	8.5 – 10

Table 12: The cut scores preset for the VSTEP.3-5 listening test

3. Administration of the VSTEP.3-5 Listening test

Based on the VSTEP.3-5 test manual and information provided by the Center for Language Testing and Assessment, ULIS-VNU, some observations on the characteristics of the setting for the VSTEP.3-5 listening test are as follows:

First, in terms of physical characteristics, all of the test rooms are well-equipped with a good auditorium system for listening. Second, all of the instructors for the VSTEP.3-5 listening test are the staff and teachers of ULIS. They all have been trained on the duties and generally follow the instruction given by the chairperson. Third, all the test takers are informed of the test formats and provided with instructions on how to register and prepare for the test. The information is publicly announced on the website of ULIS

(http://vstep.edu.vn) at the beginning of the academic year for students in ULIS and at the beginning of the calendar year for test takers outside ULIS. The VSTEP.3-5 listening test is delivered in the morning after the reading test when the test-takers are fresh and not tired. Fourth, in terms of scoring, before the listening answer sheets are marked, to make sure there is no error with the keys, the keys for each of the test codes are once again carefully checked by an authorized member of the Center for Language Testing and Assessment, ULIS-VNU. Then, all of the answer sheets are marked with a special machine operated by two experienced technicians. The marking process goes through a strict series of steps, which are consistent across occasions. Each correct answer gets 1 point. The raw scores out of 35 are converted into the scale of 0 - 10 according to a preset rule.

CHAPTER IV
METHODOLOGY

This chapter first presents an interpretive argument developed for the cut-scores of the VSTEP.3-5 listening test and then continues with a description of the research questions, the data and the data collection as well as the use of the methodologies in the study.

1. Building a validity argument for the cut-scores of the VSTEP.3-5 listening test

The overview of interpretive argument for the cut scores of the VSTEP.3-5 listening test is as follows.

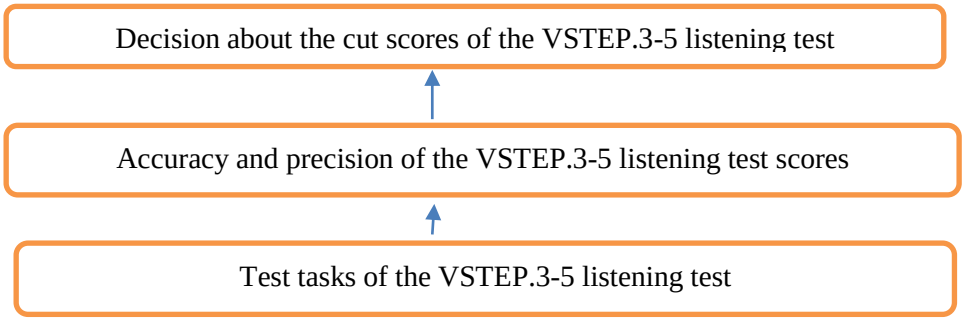

Figure 3: Overview of interpretive argument for the cut scores of the VSTEP.3-5 listening test

The three inferences involved in the validity argument of the cut-scores of the test include (1) test tasks (2) accuracy and precision, and (3) cut-scores. The warrants and assumptions are outlined for each inference as follows.

Inference 1: Test tasks

Warrant 1: The test tasks of the VSTEP.3-5 listening test are properly designed in accordance with the characteristics specified in the test specification.

Assumptions:

- The test rubrics are carefully designed so that they do not cause any linguistic problems to the test-takers.
- The input to which test-takers are exposed resembles normal everyday speech and designed in accordance with the descriptions in the test specification.
- The successful completion of the test tasks is dependent on the comprehension of the text.

Inference 2: Accuracy and precision

Warrant 2: The observed VSTEP.3-5 listening test scores show that test is reliable in measuring test-takers' proficiency.

Assumptions:

- The test reliability is in the good range for proficiency assessment.
- The test items are well-designed with the predetermined difficulty as described in the test specification and with good discrimination.

Inference 3: Cut-scores

Warrant 3: The cut scores established for the VSTEP.3-5 listening test are useful for making decisions about test takers' English listening competency.

Assumptions:

- The standard-setting procedures used for the VSTEP.3-5 listening test were properly implemented.
- The reliability of the VSTEP.3-5 listening test is within the acceptable range for 3 cut-off points.
- The cut-scores established through another standard-setting method correlate well with the current cut-scores set for the VSTEP.3-5 listening test.

As can be seen, the interpretive argument for the cut scores of the VSTEP.3-5 listening test consists of three inferences, each of which is supported by a warrant. Each warrant is in turn clarified by assumptions, which needs evidence from various sources. The research questions arising from these inferences and the methods adopted to answer these research questions are to be presented in the following parts.

2. Data collection

Within the framework of this study, several different types of data are used: (1) the VSTEP.3-5 listening test used in ULIS-VNU on 25th March 2017 and (2) the statistics of the test-takers' scores on this listening test are used for item analysis and standard setting procedure.

3. Methodology

To address the stated research questions, this study adopts complementary approaches in collecting evidence in correspondence with the three main inferences. In specific, both qualitative and quantitative methods are combined and structured to judge the plausibility of cut score interpretation and use. The evidences taken from both qualitative and quantitative methods are to back certain assumptions in each of the three inferences in the validity argument of the VSTEP.3-5 listening test. This mixed method is agreed to provide a proper insight into the validity issue and strengthen the argument made since each method has its own strengths and drawbacks.

4. Description of methods used for the study

4.1. Analyzing the test tasks

The first inference, test tasks, implies that the test tasks were appropriate and fair to the test-takers. Thus, three different aspects of test context are to be analyzed. The study basically employs the framework of listening task characteristics proposed by Buck (2001), which is then adapted by Weir (2005a). In regard to the test tasks, this study focuses on analyzing characteristics of test rubrics, linguistic demands (task input and response) and speakers.

Characteristics of the input and response

In analyzing these features, the test specification is used in order to provide the descriptions of the levels of the test intends to cover, along with a description of particular sub-skills or knowledge areas to be included.

The textual features that affect the comprehensibility and difficulty of listening tasks are analyzed through automated judgment. Advances in automated textual analysis have made it possible to examine analytically on a wider range of text characteristics to complement human judgment. In this study, the free online software tools, English Profile and Readable.io, is utilized to aid in the analysis of textual features.

4.2. Investigating accuracy and precision

Accuracy and precision addresses test reliability, item difficulty and item discrimination of the VSTEP.3-5 listening test. With regard to this inference, this study employs both quantitative and qualitative methods to provide the insight into the accuracy and precision of the VSTEP.3-5 listening test scores.

In this study, the licensed software WINSTEPS (3.92.1) sponsored by the U.S Embassy in Vietnam and other softwares such as Iteman 4.3 and SPSS 17.0 are used to construct Rash measurement and distractor analysis.

Test reliability analysis

Statistical analyses of test scores help to provide information about the test reliability. Test reliability is defined as a basic requirement of a valid test and refers to the consistency of measurement. Numerous strategies with statistical tools have been introduced in order to investigate the issue of consistency in measurement. For this study, reliability coefficients and standard error of measurement (SEM) are used to examine the reliability of the test. Reliability coefficients range from 0.00 to 1.00. Ideally, score reliability should be above 0.70. Coefficients in the range 0.80 – 0.90 are very good for proficiency assessment. Standard error of measurement (SEM) is directly related to the reliability of a test; that is, the larger the SEM, the lower the reliability of the test and the less precision there is in the measures taken and scores obtained.

Besides, statistical information is also analyzed to investigate how well the items function on a test with a group of test takers. In this study, item facility, item discrimination and distractor efficiency are taken into consideration.

Item facility analysis

Item facility analysis employs item facility (IF) which is a statistical index used to examine the percentage of students answering a given item correctly. According to Ebel (1979:267), the critical values for evaluating test item facility are as follows:

Index of Difficulty	Evaluation of Difficulty
0.80 – 1.00	Too easy
0. 60 – 0.79	Rather easy
0.40 – 0.59	Moderately difficult
0.20 – 0.39	Rather difficult
0.00 – 0.19	Too difficult

Table 14: Criteria for item selection and interpretation of item difficulty index

Item discrimination analysis

Item discrimination is the index indicating the degree to which an item separating the students who performed well from those who performed poorly.

According to Ebel (1979:267), the critical values for evaluating test item facility are as follows:

Index of Discrimination	Evaluation of Discrimination
0.60 - 1.00	Very good items
0.40 – 0.59	Good items
0.20 – 0.39	Reasonably good but possibly subject to improvement
0.10 – 0.19	Marginal items, usually need and subject to improvement
0.00 – 0.09	Poor items, to be rejected or rewritten

Table 15: Criteria for item selection and interpretation of item discrimination index

Distractor analysis

The distractor analysis provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item. In this study, the number of times each distractor is selected is noted in order to determine the effectiveness of the distractor.

4.3. Validating cut-scores

The cut-score proposition holds that the cut scores of the VSTEP.3-5 listening test are properly established so that a test taker is accurately described by the performance standards and conversely. There are 3 cut scores of the VSTEP.3-5 listening test that separate test takers into four levels (unrated, level 3, level 4 and level 5). As mentioned earlier, evaluation of standard setting is a multifaceted endeavor with many potential sources of evaluation information including procedural, internal and external evidence. In terms of this inference, the answers are sought in this study to the following questions:

- Was the standard setting of the VSTEP.3-5 listening test suitably and properly implemented?

- What are the standard errors of each cut-off points? Does the test reliability satisfy the reliability of the classification decisions based on the established cut-scores?
- Are the cut-scores calculated through Bookmark method different from the current cut-scores set by Angoff method?

4.3.1. Procedural - *Report and documentation of the standard setting of the VSTEP.3-5 listening test*

For the VSTEP.3-5 listening test, the procedural evidence was obtained by taking a close look at the report and documentation of standard setting for the test.

4.3.2. Internal - *Statistical analysis*

To collect the internal evidence for the validity of the proposed cut-scores, the precision of the cut-score estimations of the VSTEP.3-5 listening test was taken into consideration. The data about standard errors of the cut-off points (**SEc**), the standard error of the test (**SEM**) and the test reliability were calculated.

The following table is based on this index and presents what should be required level of test reliability in order to ensure a reliable separation into the desired number of proficiency levels.

Number of levels	2	3	4	5	6
Number of cut-off points	1	2	3	4	5
Test reliability	>=0.61	>=0.80	>=0.88	>=0.92	>=0.95

Table 16: Number of proficiency levels & test reliability (Wright, 1996)

The table above shows that test reliability is very important for the trustworthy classification decisions based on the proposed cut-score interpretations. The VSTEP.3-5 listening test scores classify test-takers into 4 levels with 3 cut-off points. For this reason, VSTEP.3-5 listening test reliability should be at least 0.88.

4.3.3. External - *Bookmark standard setting*

In this study, the instrument to check the external validity of cut scores of the VSTEP.3-5 listening test was through comparison between the cut-scores established for the test by the Angoff method and those set by Bookmark method. First, the current cut-scores established and published for the results of the VSTEP.3-5 listening tests were used. The Angoff standard setting method is reported to have been employed for setting the cut-scores for the test. Then, the Bookmark method was conducted as a means of external validation. The proposition is that the cut-scores from the Angoff and Bookmark standard setting method will be the same or the score regions of the mean plus and minus one standard error will be overlapping.

CHAPTER V

RESULTS OF THE STUDY

This chapter summarizes the main results of three main analyses in accordance with three main research questions or inferences of the study including test task analysis, accuracy and precision of scoring, and the correlation analysis between the cut-scores preset by Angoff method and the cut-scores set by Bookmark method.

1. Analysis of the test tasks

The analysis of the VSTEP.3-5 listening test context are based on the test task characteristic analysis given by Buck (2001). The relevant data sources for the test task characteristics analysis include the VSTEP.3-5 listening test specification and the VSTEP.3-5 listening booklet delivered on 25th March 2017 with its accompanied recordings. The test task characteristics which are taken into consideration in this study are those under the categories provided in Buck's framework (2001) which covers the following issues: (1) characteristics of the test rubric, (2) characteristics of the input and (3) the relationship between the input and response.

1.1. Characteristics of the test rubric

The investigation into the main characteristics of the test rubric on the VSTEP.3-5 listening test specification and the test booklet yields some evidence in terms of test structure, test instructions, time allotment, and scoring method as follows.

First, by taking a close look at the test specification of the VSTEP.3-5 listening test and the test booklet delivered on 25th March, it is found out that the test specification was developed in detail with all the necessary information that can help test item writers to develop and design the test.

In terms of test structure, the information in the test specification shows that the VSTEP.3-5 listening test includes 3 parts. In part 1, test-takers hear 8 short recordings and there is one question following each recording. In part 2, test-takers hear 3 conversations and there are 4 questions following each conversation. In part 3, test-takers hear 3 extended monologues and there are 5 questions following each monologue. All the recordings are played once only and the question format of the test is multiple-choice questions (MCQ) with a question stem and 4 options. A close look at the test booklet administered on 25th March 2017 reveals that the test was strictly designed in accordance with the structure described in the test specification.

With regards to instructions, the test starts with general instruction for the whole test and then there is instruction for each part. The introductions are provided both in spoken and written forms. The instructions are recorded by an American male speaker and are found to be clear and simple. The general instruction of the test provides the general information for the whole test including the brief structure of the test, number of playing, brief guides on how to do the test and time allowance for answer transferring.

Besides, each part of the test has its own instruction, which provides information on text types (conversation or talk and long or short), number of recordings and task types. There is an example given in the first part that helps test-takers to prepare and get familiar with the multiple-choice question format before they do the test.

For part 2 and 3, besides the general instructions for the whole part, there is basic contextual information in the written form for the each text so as to aid the activation of appropriate schematic information for the test-takers. The instructions include contextual features about the text that the test-takers are going to listen to such as the speakers, their location, the name of the speaker or speakers and the channel of communication (e.g. radio, podcast, face to face) as well as the topic of the talks or the conversations.

In terms of time allotment, the specifications of VSTEP.3-5 listening test reveal that on average, test takers have 5 seconds to read through a question. However, there is no information about the silent time inserted between two individual questions and after a set of questions. At the end of the test, test takers have 7 minutes to transfer their answers to the answer sheet. With the aid of a special software named cool edit pro.2.1, the silent time inserted is measured for the recordings of the VSTEP.3-5 listening test administered on 25th March, it is found out that the silent time inserted before and after the set of questions in general follows the requirement described in the test specification. For question 1-8 in part 1, though the information about the silent time between two items is not given in the test specification, 7 seconds is the time measured and this is consistent throughout the items.

In terms of scoring method, the VSTEP.3-5 listening test uses the multiple-choice format. The answer sheets will be then automatically scanned and scored by two authorized technicians. The specification states clearly how each item is cored. The listening score is based on the number of correct answers with each item equally weighted. How an item is scored is made explicit to the test-takers. However, how the raw scores are converted to the scores on the scale of 10 is not made explicit to the test-takers.

1.2. Characteristics of the input

A close look at the item writer guidelines for the VSTEP.3-5 listening test provided by the Center for Language Testing and Assessment reveals that the procedure of constructing a VSTEP.3-5 listening input materials follows the same patterns from one occasion to another.

A close look at the specifications of the VSTEP.3-5 listening test provides detailed information about the input for the test. The information includes descriptions about contextual parameters, format and level of difficulty. The contextual parameters of the test provide information about the length of the text, the domain and the speaker voice/accent. The format of the text can range from a chat between two friends at school, an announcement at the airport to an interview on the radio or a lecture at the university. The level of difficulty

of the text ranges from Level 3 to level 5 with detailed description about vocabulary and structure level, number of word per sentences, speech rate (number of words per minute/second) and readability score.

The analysis of the input of the VSTEP.3-5 listening test administered on 25th March reveals that in general the selection of the input listening materials follows the guidelines in the test specifications. In terms of speaker accent, generally speaking, the VSTEP.3-5 test takes a commendably moderate approach to the use of variety. The recordings for the tests have a familiar variety of English accent such as standard American English, British English or Australian English. However, analyzing the recordings of the VSTEP.3-5 listening test administered on 25th March 2017 points out that American English features quite extensively. According to Field (2013), usually, for low level items, the texts are recorded with standard and clear accent so that the test takers can be spared apprehension about what they are likely to hear. The range of accents increases for the higher level items and there is even quite extensive mixing of accents within recordings. However, there seems no strategies in the use varieties in the VSTEP.3-5 listening test administered on 25th March 2017. This may be because of the unavailability of voice actors on the recording days. In the following part, some more interesting remarks about the characteristics of the input will be presented.

Part 1 (Items 1-8)

Items 1-8 are separate items with separate texts. First, in terms of length, all the texts of items 1-8 satisfy the requirement of length as described in the test specifications, that is, 80 – 125 words.

In terms of speech rate, since the duration for the recorded text for items 1-8 in part 1 is short, the speech rate is analyzed in terms of words per second (wps). According to the descriptions in the test specifications for the VSTEP.3-5 listening test, on average, the speech rate for items 1-8 is 2.79 wps and it can fluctuate around 2.66-3.17 wps. This rate is even marginally higher for the level-based tests of the Cambridge ESOL suit (Field, 2013: 119). As can be seen, the speech rate for items 1-8 fluctuate from 2.05 wps to 2.67 wps, which means that the speech rate for all of the items in part 1 is slower than the standard one, especially the speech rate for item 6. The real position of difficulty for each item will be analyzed in the later parts via the statistical data on the item map.

With regards to the text content of part 1, in general, the themes for items 1-8 belong to public and personal ones, which are considered appropriate for low levels since low level language learners are more likely to deal with communicative needs within public and personal domains while higher level language users are more likely to enter the vocational and educational life of a country (Council of Europe, 2001). Besides, it can be seen that the themes and schemata of the text for these items are relevant and representative of the public and personal TLU domains. One more remark is that the topics of all 8 texts for items 1-8 do not belong to the sensitive group of topics for testing and assessment.

In terms of language levels, by using the online tool Readable.I0 (https://readable.io/text) to analyze the language level of the texts of items 1-8, it is found out that the readability index of the texts for items 1, 2, 3, 4, 5, 7 and 8 fit with the Flesh readability scores for level CEFR B1 level (70-80) with 78.8, 80.0, 70.5, 75.5, 78.1, 77.3 and 79.0 respectively while that for item 6 seem to be more difficult than expected with the readability score of 69.9 respectively.

Part 2 (Items 9-20)

There are three texts in part 2, each of which cover four related items. First, in terms of length, all the texts of items 9-20 satisfy the requirement of length as described in the test specifications.

In terms of speech rate, it can be seen that the speech rate for items 13-16 is rather slow for the text level 4 in comparison with the description from the test specifications and the results from the Cambridge ESOL suit (Field, 2013: 119). The picture is also similar for items 17-20 when the desired level of the text is level 5, the speech rate is only 2,67wps, the same as level 3.

With regards to the text content of part 2, in general, the themes for the texts in part 2 are considered appropriate, relevant and representative of TLU domains. Although the domain of the text for the items 9-12 is occupational, the topic or the schemata is an informal conversation between two friends, which is suitable

for level 3 together with the suitable level of vocabulary and structure. That for items 17-20 seems easier than the required level (level 5) both in terms of topic, vocabulary and structure.

In terms of language levels, it is found out that these three texts are generally designed with appropriate level of difficulty in terms of vocabulary and grammatical structures. The analysis reveals that the vocabulary and grammatical structures of the text for items 9-12 ranges from A1 to B2 with 5 words at B2 level (2.7%). That for items 13-16 ranges from A1-C1 with only one word at C1 level (0.42%) and that for items 17-20 ranges from A1-C2 with 4 words at C2 levels (4.64%). Besides, analysis of the readability scores show that the readability index for the text of items 9-12 is 80.4 and that for items 13-16 is 71.2, which can be considered acceptable for for B1 and B2 level of text though they are marginally higher than the required ones. However, the readability for the text of items 17-20 shows that the text is not as difficult as it should be with the readability index of 70.3 (the readability of the text for C1 level is 50-60).

Part 3 (Items 21-35)

There are three texts in part 3, each of which covers five related items. According to the specifications of the VSTEP.3-5 listening test, the expected length of the texts for items 21-25 and items 26-30 is from 560 to 740 words and the speech rate is between 3.16 wps and 3.45 wps. The data from the table 25 reveal that the text length for these two first recordings of part 3 in the VSTEP. 3-5 listening test are lower than the minimum requirement and the speech rate is also relatively slower. The shorter length of texts and the slower speech rates can lower the difficulty of the related items. For the text on which items 31-35 are built on, as can be seen, the length and the speech rate both meet the standard requirement described in the test specifications, that is, the length of the text should be between 560 and 740 words and the speech rate is supposed at 3.16 – 3.54 wps. This can contribute to the difficulty level of the related items.

With regards to the text content of part 3, the themes for the texts in part 3 are considered appropriate, relevant and representative of TLU domains. The topics range from concrete to abstract ones. The topic for text 3 is an abstract one about happiness with unfamiliar concepts like *Confucius, dignity, practice compassion, nirvana, illusions* and so on. However, according to the descriptions about the overall listening comprehension for C1 level (Council of Europe, 2001), C1 level candidates can understand enough to follow extended speech on abstract and complex topics beyond his/her own field. Thus, the topic for text 3 for items 31-35 has done its good job in terms of levels.

In terms of language levels, it is found out that in generally, the levels of the texts fit the description in the test specification though they all contain few words and grammatical structures above the targeted levels (around 3-5% of the total number of words). Besides, the analysis of the readability index of these texts are 70.1, 64.0 and 60.4 respectively, which means that the text difficulty levels relatively fit with the targeted ones.

1.3. Relationship between the input and response

The aspects of the relationship between the input and response under examination is directness and interactiveness, which in this study refer to the dependency on the content of the listening texts and the employment of listening skills and relevant academic sub-skills to succeed on the test. The subskills tested in the VSTEP.3-5 listening test include ability to listen for specific information, ability to listen for gist, ability to determine a speaker's attitude/intention, ability to relate utterances to their social and situational contexts, and ability to make inferences. In terms of directness, from the analysis of the texts and the questions in the VSTEP.3-5 listening test forms used on the 25th March 2017, it is found out that all of the 35 items of the test have high passage dependency and interactiveness, which means the successful completion of the listening tasks is dependent on the comprehension of the relevant listening texts. In other words, in order to be highly probable to choose the correct options for these items, the test-takers have to reply on their comprehension of the listening texts instead of basing on their merely background, as well as to be able to use relevant academic listening skill such as understanding main ideas, catching specific details, connecting ideas, understanding implication/inference and synthesizing information.

2. Accuracy and precision of scoring

2.1. Overall statistical result

Indices used to analyze the accuracy and precision of the VSTEP.3-5 listening test scoring include test reliability, item difficulty and item discrimination. The summary show that overall, the test is rather difficult for this group of test-takers (N=1,526) and the test in general can discriminate test-takers reasonably well.

In terms of test reliability, the test reliability on 35 scored items of the VSTEP.3-5 listening test is 0.815, which is very good for proficiency assessment. The standard error of measurement (SEM) is of an acceptable range, 2.581. With regards to each items, the statistical analysis reveals the Alpha coefficients for 35 items of the test range from 0.804 – 0.819. This show that all of the 35 items of the VSTEP.3-5 listening test administered on 25th March have high reliability.

The table 35 shows the person reliability, item reliability and separation of the test.

PERSON	1562 INPUT	1562 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	14.1	35.0	-.48	.41	1.00	.0	1.02	.0
P.SD	6.0	.0	.91	.05	.14	.9	.24	.9
REAL RMSE	.41	TRUE SD	.81	SEPARATION	2	PERSON RELIABILITY		.81
ITEM	35 INPUT	35 MEASURED			INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	627.1	1562.0	.00	.06	1.00	-.4	1.02	.0
P.SD	252.0	.0	.84	.01	.10	3.5	.15	3.4
REAL RMSE	.06	TRUE SD	.84	SEPARATION	13.69	ITEM	RELIABILITY	.99

Table 32: The person reliability and item reliability of the test

As can be seen from the table, the person separation is 2 and the person reliability is 0.815. According to explanation provided in the tutorial of the WINSTEPS, the VSTEP.3-5 listening test administered on 25th March is sensitive enough to distinguish between the low and high performers and the test contain fairly enough items to distinguish the test takers. With the item separation of 13.69 and item reliability of 0.99, it is implied that the person sample is large enough to confirm the item difficulty hierarchy (=construct validity) of the testing instrument.

In terms of difficulty, based on Ebel's table of evaluating difficulty index (1979:267), the difficulty levels are divided into five groups: too difficult, rather difficult, moderately difficult, rather easy and too easy. As can be seen, out of 35 items, 3 items are too difficult, 18 items are rather difficult, 9 items are moderately difficult, 4 items are rather easy and 1 item is too easy.

By taking a close look at the VSTEP.3-5 listening test administered on 25th March 2017 and its specifications, each item is labeled with one desired difficulty level before running the data on WINSTEPS. Figure 5 presents the item map with full item labels.

```
MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    4               +
                .   |
                    |
                    |
                    |
                    |
                .   |
    3               +
                    |
                    |
                .   |
                    |
                .   |
                    |
    2            .#  +
                    |
                .#  |T
                .#  |
                .#  |  i33.C1.3
               .## T|  i34-C1.3
                    |  i14-B1.3
    1           .##  +  i6-B1.3
```

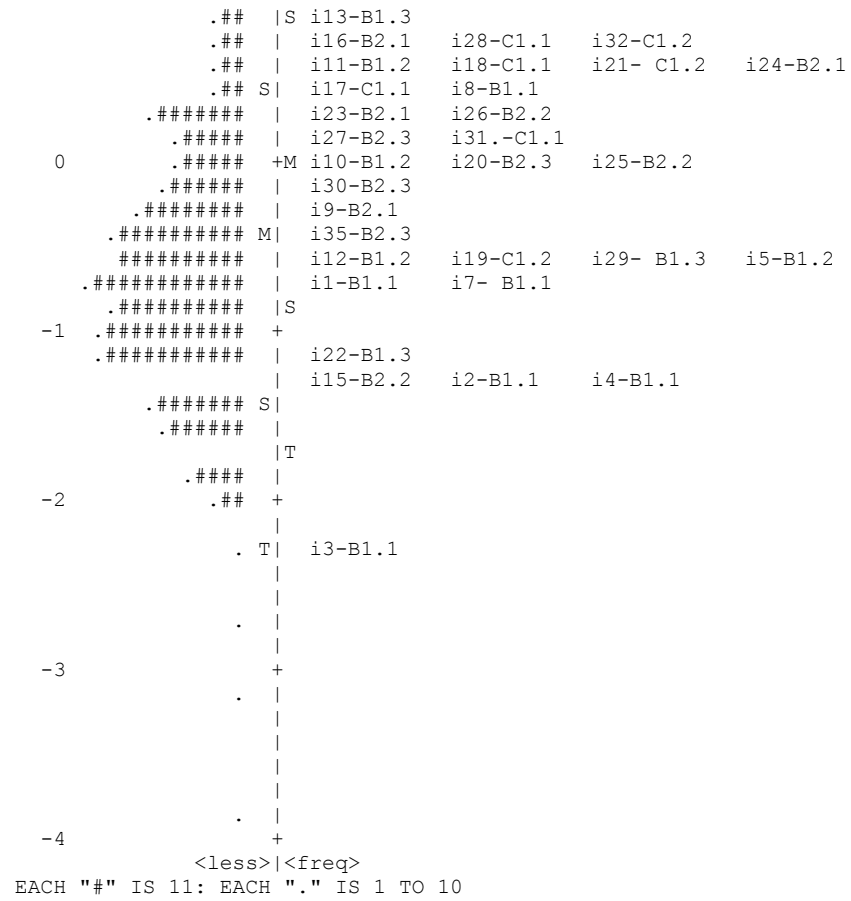

Figure 5: Item map of the VSTEP.3-5 listening test

The item map shows that in general, the VSTEP listening test administered on 25th March 2017 is rather difficult for the population of test-takers and the items are distributed fairly well on the continuum of difficulty. However, since the VSTEP.3-5 listening test is a proficiency test, the desired difficulty level of the test does not need to be dependent on the specific population of test-takers. But there needs more items in the blanks on the item map so that the test can measure the test-takers of these competency.

It is also noticed from the item map that some items are much more difficult or easier than desired. The items that are much more difficult than desired include items 14, 6 and 13 whereas items 15 and 19 seem much easier than the required level in accordance with the test specifications.

In terms of item discrimination, the overall discrimination of the test (mean discrimination of all the items) is 0.300, which means that the test in general can discriminate test-takers reasonably well. Based on Ebel's table of evaluating discrimination index (1979:267), the discrimination levels are divided into five groups: Very good items, Good items, Reasonably good but possibly subject to improvement, Marginal items, usually need and subject to improvement and Poor items, to be rejected or rewritten. As can be seen from table 37, out of 35 items, 2 items are poor with point-biserial values of under 0.09 and these two items as a general rule, these two items need to be rejected or rewritten. 4 items are marginal and need some improvement with the point-biserial values ranging from 0.1 to 0.19. To be more specific, 24 items can discriminate test-takers reasonably well and 5 items can discriminate test-takers very well. The two items that are flagged due to low discrimination are item 28 and 34. Item 34 is discriminating negatively with the index of -0.015. This means that overall, the most competent test-takers get this item wrong and the least competent test-takers get this item right. Item 28 has critically low discrimination index, which means that it has little ability to discriminate test-takers.

3. Analysis of the cut-score validity

3.1. Procedural evidence

The cut-scores were set for the VSTEP.3-5 listening model test whose specifications are described in great

detail so that the bank of listening tests are constructed. The cut-scores for the VSTEP.3-5 listening test were established by Angoff method. A close investigation into the report and documentation of setting the cut-scores for the test indicates that the standard setting were implemented with strict procedures.

The cut-score setting for the VSTEP.3-5 listening test involved 5 experts who were knowledgeable of CEFR levels and had experience in designing tests. Before the real process took place, the participants were explained about the aim of the standard setting. They had the chance to get familiarized with the performance standard descriptions of the VSTEP.3-5 listening test.

The actual process was conducted in three rounds of ratings before the final results were given. All the steps and results of each round were documented carefully.

3.2. Internal evidence

To collect the internal evidence for the validity of the proposed cut-scores, the precision of the cut-score estimations of the VSTEP.3-5 listening test was taken into consideration. The data about the test reliability was taken into consideration.

Test reliability affects strongly the reliability of the classification decisions based on the established cut-scores (Wright, 1996). Wright (1996) suggests the so called Index of Separation. The following table is based on this index and presents what should be required level of test reliability in order to ensure a reliable separation into the desired number of proficiency levels.

Number of levels	2	3	4	5	6
Number of cut-off points	1	2	3	4	5
Test reliability	>=0.61	>=0.80	>=0.88	>=0.92	>=0.95

Table 75: Number of proficiency levels & test reliability

The table above shows that test reliability is very important for the trustworthy classification decisions based on the proposed cut-score interpretations. The VSTEP.3-5 listening test scores classify test-takers into 4 levels with 3 cut-off points. For this reason, VSTEP.3-5 listening test reliability should be at least 0.88.

The statistics about the test reliability of the VSTEP.3-5 listening test administered on 25th March show that the test reliability on 35 scored items of the VSTEP.3-5 listening test is 0.815, which is very good for a normal proficiency assessment.

Score	Alpha	SEM
Scored items	0.815	2.581

Table 76: Test reliability of the VSTEP.3-5 listening test

However, based on the correlation between the number of proficiency levels and test reliability, for a proficiency test with three cut-score points like the VSTEP.3-5 listening test, the test reliability index (0.815) is marginally low for the requirement (0.88).

3.3. External evidence

In this study, the instrument to check the external validity of cut scores of the VSTEP.3-5 listening test was through comparison between the cut-scores established for the test by the Angoff method and those set by Bookmark method. In other words, the Bookmark method was conducted as a means of external validation.

The Bookmark procedure consisted of a psychometrician rank ordering all of the items for the VSTEP.3-5 listening test administered on 25th March 2017. This ranking is based on each item's difficulty. In this study, the IRT β (item difficulty) is utilized to create the ordered item booklet (OIB). The OIB of the listening test is created where the first item is the easiest, followed by the next easiest, all the way until the end where the last item is the most difficult.

Round 1

The following table summarizes the bookmarks of 20 participants for round 1 of Bookmark standard setting procedure.

	Level 3		*Level 4*		*Level 5*	
participant ID No	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*
1	4	-0.547	15	0.683	28	1.393
2	7	-0.047	18	0.783	28	1.393
3	4	-0.547	12	0.223	28	1.393
4	4	-0.547	14	0.493	28	1.393
5	4	-0.547	12	0.223	28	1.393
6	7	-0.047	16	0.703	29	1.423
7	4	-0.547	11	0.163	28	1.393
8	7	-0.047	18	0.783	29	1.423
9	5	-0.507	20	1.013	28	1.393
10	7	-0.047	22	1.083	32	1.763
11	7	-0.047	19	0.823	34	2.053
12	3	-0.587	18	0.783	27	1.313
13	4	-0.047	12	0.223	18	0.783
14	7	-0.047	19	0.823	34	2.053
15	3	-0.587	14	0.493	28	1.393
16	5	-0.507	16	0.703	28	1.393
17	4	-0.547	17	0.743	28	1.393
18	4	-0.547	20	1.013	28	1.393
19	4	-0.547	24	1.233	34	2.053
20	4	-0.547	20	1.013	29	1.423
Summary Statistics in Theta (Ability) Metric						
Mean cut		-0.372		0.700		1.481
Median cut		-0.547		0.763		1.393
SD		0.054		0.069		0.066
Minimum		-0.587		0.163		0.783
Maximum		-0.047		1.233		2.053
Mean -1SD		-0.426		0.631		1.415
Mean +1SD		-0.318		0.769		1.547

Table 78: Summary of Output from Round 1 of Bookmark standard-setting Procedure

Table 78 provides a summary of bookmark placements for each participant together with the resulting cut scores. The resulting cut scores are shown in theta (ability) metric, which is then converted into raw score metric through the conversion table run by WINSTEPS.

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-5.14E	1.84	12	-.74	.38	24	.91	.39
1	-3.89	1.03	13	-.59	.38	25	1.06	.40
2	-3.13	.75	14	-.45	.37	26	1.22	.41
3	-2.66	.63	15	-.31	.37	27	1.39	.42
4	-2.32	.56	16	-.18	.37	28	1.58	.44
5	-2.03	.51	17	-.05	.36	29	1.78	.47
6	-1.79	.48	18	.09	.36	30	2.02	.50
7	-1.58	.45	19	.22	.36	31	2.29	.55
8	-1.39	.43	20	.35	.37	32	2.62	.62
9	-1.21	.41	21	.49	.37	33	3.07	.74
10	-1.04	.40	22	.62	.37	34	3.81	1.02
11	-.88	.39	23	.76	.38	35	5.05E	1.84

Table 79: Conversion table

The following table shows the statistics in converted raw cut-score points for each level.

	Level 3		*Level 4*		*Level 5*	
Summary Statistics in Theta (Ability) Metric						
Mean cut		-0.372		0.700		1.481
Median cut		-0.547		0.763		1.393
SD		0.054		0.069		0.066
Minimum		-0.587		0.163		0.783

Maximum		-0.047		1.233		2.053
Mean -1SD		-0.426		0.631		1.415
Mean +1SD		-0.318		0.769		1.547
Summary Statistics in Raw Score Metric						
Mean cut		15.0		23.0		27.00
Median cut		13.0		23.0		27.00
Minimum		13.0		19.0		23.00
Maximum		17.0		26.0		30.00
Mean -1SD		14.0		22.0		27.00
Mean +1SD		15.0		23.0		28.00

Table 80: Summary of statistics in raw score metric for round 1

As can be seen, the mean cut and median cut in theta metric for level 3 is -0.372 and -0.547 respectively, which is equivalent to 15 and 13 in the raw scores. The cut score one standard deviation above the mean recommended cut scores for level 3 is 14 and the cut score one standard deviation below the mean recommended cut scores for level 3 are still 15 since the standard deviation is too small (0.054) and does not considerably change the converted raw cut score.

As for the cut scores for level 4, the mean cut and median cut in theta metric is 0.700 and 0.763 respectively. Both of these theta values are converted and rounded to 23 in the raw scores. The cut score one standard deviation above the mean recommended cut scores for level 4 is 22 and the cut score one standard deviation below the mean recommended cut scores for level 4 are still 23 since the standard deviation is too small (0.069) and does not considerably change the converted raw cut score.

With regard to the cut scores for level 5, the mean cut and median cut in theta metric is 1.481 and 1.393 respectively, which is both equivalent to 27 in the raw scores. The cut score one standard deviation above the mean recommended cut scores for level 5 is still 27 and the cut score one standard deviation below the mean recommended cut scores for level 5 are 28.

Round 2

Table 85 below shows the summary of output from round 2 of Bookmark standard-setting procedure.

	Level 3		*Level 4*		*Level 5*	
participant ID No	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*
1	4	-0.547	19	0.823	28	1.393
2	4	-0.547	20	1.013	28	1.393
3	4	-0.547	16	0.703	28	1.393
4	4	-0.547	19	0.823	28	1.393
5	4	-0.547	16	0.703	28	1.393
6	7	-0.047	16	0.703	29	1.423
7	4	-0.547	19	0.823	28	1.393
8	7	-0.047	23	1.133	29	1.423
9	5	-0.507	20	1.013	28	1.393
10	7	-0.047	22	1.083	32	1.763
11	7	-0.047	19	0.823	34	2.053
12	4	-0.547	18	0.783	28	1.393
13	4	-0.547	16	0.703	25	1.243
14	7	-0.047	19	0.823	34	2.053
15	3	-0.587	18	0.783	28	1.393
16	5	-0.507	18	0.783	28	1.393
17	4	-0.547	17	0.743	28	1.393
18	4	-0.547	20	1.013	28	1.393
19	4	-0.547	24	1.233	32	1.763
20	4	-0.547	20	1.013	28	1.393
Summary Statistics in Theta (Ability) Metric						
Mean cut		-0.420		0.876		1.492
Median cut		-0.547		0.823		1.393
SD		0.049		0.035		0.050
Minimum		-0.587		0.703		1.243

Maximum		-0.047		1.233		2.053
Mean -1SD		-0.469		0.841		1.442
Mean +1SD		-0.371		0.911		1.542
Summary Statistics in Raw Score Metric						
Mean cut		14.00		24.00		27.00
Median cut		13.00		24.00		27.00
Minimum		13.00		23.00		26.00
Maximum		17.00		26.00		30.00
Mean -1SD		14.00		24.00		27.00
Mean +1SD		15.00		24.00		28.00

Table 81: Summary of Output from Round 2 of Bookmark standard-setting Procedure

As can be seen, median cut in theta metric for level 3 remains the same with the theta value of -0.547, equivalent to 13 in the raw scores. Meanwhile, the mean cut for level 3 is 13, one point lower in comparison with that from round 1he cut score. One standard deviation above the mean recommended cut scores for level 3 is 14 and the cut score one standard deviation below the mean recommended cut scores for level 3 is 15. In round 2, the SD is smaller than that in round 1 with the theta value of 0.049.

As for the cut scores for level 4, the mean cut and median cut in theta metric are almost the same, 0.876 and 0,823 respectively. This results in the same raw score metric of 24 for both of the mean cut and median cut. This result implies that the ratings of the participants in round 2 are highly consistent with each other. Interestingly, since the SD for level 3 is too small, it does not change the cut score one standard deviation above and below the mean recommended cut scores for level 4.

Similarly, with regard to the cut scores for level 5, the mean cut and median cut in theta metric is very close to each other, 1.492 and 1.393 respectively, the converted raw scores are the same, 27. The cut score one standard deviation above the mean recommended cut scores for level 5 is still 27 and the cut score one standard deviation below the mean recommended cut scores for level 5 are 28.

Round 3

To begin round 3 of the Bookmark standard-setting procedure, participants again use their OIB and are provided with all of their other Round 2 materials plus the summary of the Round 2 judgments. Besides, a special version of data (table 86) is also provided to the participants. This table includes actual raw score equivalents associated with the theta values that are the recommended cut scores. This feature helps to clarify for participants the relationship between their bookmark placements, the theta values associated with those placements, and the impact that a bookmark placement will have on both the raw cut score and the percentages of test-takers classified at or above a given performance level.

Page in OIB	Original Item	IRT item difficulty /Beta (β)	Theta (Θ)	Raw cut score	% at or above
1	3	-2.27	-1.577	7	94.75%
2	15	-1.31	-0.617	13	51.79%
3	2	-1.28	-0.587	13	51.79%
4	4	-1.24	-0.547	13	51.79%
5	22	-1.2	-0.507	14	44.75%
6	1	-0.75	-0.057	17	27.40%
7	7	-0.74	-0.047	17	27.40%
8	19	-0.62	0.073	18	23.75%
9	5	-0.59	0.103	18	23.75%
10	12	-0.53	0.163	18	23.75%
11	29	-0.53	0.163	18	23.75%
12	35	-0.47	0.223	19	19.91%
13	9	-0.29	0.403	20	17.29%
14	30	-0.2	0.493	21	14.92%
15	10	-0.01	0.683	22	13.00%
16	20	0.01	0.703	23	11.33%

17	25	0.05	0.743	23	11.33%
18	31	0.09	0.783	23	11.33%
19	27	0.13	0.823	24	9.73%
20	26	0.32	1.013	25	8.19%
21	23	0.34	1.033	25	8.19%
22	17	0.39	1.083	25	8.19%
23	8	0.44	1.133	26	6.59%
24	24	0.54	1.233	26	6.59%
25	18	0.55	1.243	26	6.59%
26	11	0.6	1.293	26	6.59%
27	21	0.62	1.313	27	4.87%
28	32	0.7	1.393	27	4.87%
29	16	0.73	1.423	27	4.87%
30	28	0.77	1.463	27	4.87%
31	13	0.82	1.513	28	3.84%
32	6	1.07	1.763	29	2.69%
33	14	1.13	1.823	29	2.69%
34	34	1.36	2.053	30	1.92%
35	33	1.38	2.073	30	1.92%

Table 82: Round 3 Feedback for Bookmark Standard-setting Procedure

Table 87 is the summary of output from round 3 of bookmark standard-setting procedure.

	Level 3		*Level 4*		*Level 5*	
participant ID No	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*	*Page in OIB*	*Theta @ Cut*
1	4	-0.547	13	1.513	27	1.313
2	4	-0.547	15	0.683	25	1.243
3	5	-0.507	12	0.223	27	1.313
4	5	-0.507	12	0.223	31	1.513
5	5	-0.507	12	0.223	31	1.513
6	4	-0.547	12	0.223	29	1.423
7	4	-0.547	19	0.823	28	1.393
8	5	-0.507	12	0.223	29	1.423
9	5	-0.507	20	1.013	28	1.393
10	3	-0.587	16	0.703	32	1.763
11	7	-0.047	19	0.823	32	1.763
12	4	-0.547	18	0.783	28	1.393
13	5	-0.507	16	0.703	25	1.243
14	5	-0.507	16	0.703	27	1.313
15	3	-0.587	18	0.783	28	1.393
16	5	-0.507	18	0.783	28	1.393
17	4	-0.547	17	0.743	28	1.393
18	4	-0.547	20	1.013	28	1.393
19	4	-0.547	24	1.233	32	1.763
20	4	-0.547	15	0.683	27	1.313
Summary Statistics in Theta (Ability) Metric						
Mean cut		-0.510		0.705		1.433
Median cut		-0.547		0.723		1.393
SD		0.025		0.078		0.035
Minimum		-0.587		0.223		1.243
Maximum		-0.047		1.513		1.763
Mean -1SD		-0.535		0.627		1.398
Mean +1SD		-0.485		0.783		1.468
Summary Statistics in Raw Score Metric						
Mean cut		13.00		23.00		27.00
Median cut		13.00		23.00		27.00
Minimum		13.00		19.00		26.00
Maximum		17.00		28.00		29.00
Mean -1SD		13.00		22.00		27.00

Mean +1SD		14.00		23.00		27.00

Table 83: Summary of Output from Round 3 of Bookmark standard-setting Procedure

As can be seen, for this round, both the mean and median cut in theta metric for level 3 is very close to each other, -0,510 and -0.547 respectively, equivalent to 13 in the raw scores. Since the SD is too small, 0,025, it does not change the converted raw score of one standard deviation below the mean recommended cut score for level 3. One standard deviation below the mean recommended cut scores for level 3 is 14.

Similarly, both the mean and median cut in theta metric for level 4 are also very close to each other,0.705 and 0.723 respectively, equivalent to 23 in the raw scores. One standard deviation above the mean recommended cut score is also 23 and one standard deviation below the mean recommended cut score is 22.

Interestingly, with regard to the cut scores for level 5, the mean cut, median cut, one standard deviation above/below the mean recommended cut score in theta metric is very close to each other with the small SD. Their value converted in raw score metric is 27. And because 4.0, 6.0 and 8.5 are taken as the converted cut scores on the scale of 10, the conversion table can be set as follows.

Converted score (scale of 10)	Raw score (scale of 35)
0	0
0.5	1-2
1	3
1.5	4-5
2	6-7
2.5	8-9
3	10
3.5	11-12
4	13-15
4.5	16-17
5	18-19
5.5	20-21
6	22-23
6.5	24
7	24
7.5	25
8	26
8.5	27-29
9	30-31
9.5	32-33
10	34-35

Table 84: Conversion table for VSTEP.3-5 listening raw scores set by Bookmark method

Level	Raw score	Converted score
Unrated	0- 12	0 – 3.5
3	13 – 21	4 – 5.5
4	22 – 26	6 – 8
5	27 – 35	8.5 – 10

Table 85: The cut scores set for the VSTEP.3-5 listening test by Bookmark method

As can be seen, with the raw cut-scores recommended for the results of the VSTEP.3-5 listening test set by the Bookmark standard setting method, the classification errors can be reduced. To be more specific, this conversion table has taken into consideration the standard errors of the cut scores. For level 3 and 5, the mean cut, one standard deviation above/below the mean recommended cut score are the same point, it does not considerably lead to classification errors. For level 4, one standard deviation below the mean recommended cut score is 1 point lower than the mean cut, so 22 instead of 23 is taken as the cut point for this level.

Some comparisons can be made for the cut-scores preset for the result of the VSTEP.3-5 listening test administered on 25th March 2017. Table 88 presents the conversion table from the raw scores on the scale of

35 to the converted scores on the scale of 10 for the test. Table 80 presents the cut scores applied for the VSTEP.3-5 listening test, which was set by Angoff standard-setting method.

Converted score	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Raw score	0-1	2	3	4	5-6	7	8-9	10-11	12-13	14-15	16-17
Converted score	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	
Raw score	18-19	20-21	22	23-24	25	26-27	28-29	30-31	32-33	34-35	

Table 86: Conversion table for VSTEP.3-5 listening raw scores set by Angoff method

Level	Raw score	Converted score
Unrated	0- 11	0 – 3.5
3	12 – 19	4 – 5.5
4	20 – 27	6 – 8
5	28 – 35	8.5 – 10

Table 87: The cut scores set for the VSTEP.3-5 listening test by Angoff method

As can be seen from table 88 and 89, the raw cut-scores set by Angoff standard-setting method for level 3, 4 and 5 are 12, 20 and 28 respectively. The following table presents the comparison between the results of two methods.

Level	Converted score (scale of 10)	Raw scores by Angoff method (scale of 35)	Raw scores by Bookmark method (scale of 35)
Unrated	0 – 3.5	0- 11	0- 12
3	4 – 5.5	12 – 19	13 – 21
4	**6 – 8**	**20 – 27**	**22 – 26**
5	8.5 – 10	28 – 35	27 – 35

Table 88: Comparison between the results of two standard-setting methods

It is clearly seen from table 87 that for level 3, the raw cut score preset by Angoff standard-setting method for the results of the VSTEP.3-5 listening test is one point lower than that set by Bookmark standard-setting method. For level 4, the gap is 2 points for these two methods and for level 5, it is one point. As mentioned, the proposition is that the cut-scores from the Angoff and Bookmark standard setting method will be the same or the score regions of the mean plus and minus one standard error will be overlapping. Thus, more research and evidence are necessary before any recommendation on the change in the cut score for level 4, which is a 2-point gap between the results of two standard setting methods.

CHAPTER VI
CONCLUSION

This chapter comprises of four main parts. In the first part, some summarizing conclusions and discussion will be made in accordance with three main research questions. Then, the second part will focus on implications for the results of the current study. The third and fourth part will focus on the limitations and some suggestions for further research.

1. Findings of the study

1.1. Are the characteristics of the test tasks properly designed in accordance with the characteristics specified in the specification?

(1) The test rubrics are carefully designed so that they do not cause any linguistic problems to the test-takers.

As for the test instruction, the analysis of the test booklet together with the audio file of the VSTEP.3-5 listening test used on 25th March 2017 show that in general, the instructions provide the candidates all the information they need in order to complete the task as expected. The general instruction and the instructions for different parts in the test are short, clear and simple enough so as not to cause any significant processing

load. The vocabulary and the grammar structure in the instructions do not cause any problems for the test-takers at the proficiency levels being tested (level 3-5). One more special thing is that in part 2 and part 3 of the VSTEP.3-5 listening test, besides the general instructions for the whole part, there is basic contextual information in the written form for the each text so as to aid the activation of appropriate schematic information for the test-takers.

In terms of time allotment, an investigation into the specifications of VSTEP.3-5 listening test reveals that there are predetermined descriptions about time allotment for the whole listening test, for each part of the test and for each text of the part. Besides, there is also predetermined requirements about the silent time inserted before and after each question or a group of questions. However, there is no information about the silence time inserted between questions in part 1 of the test and thus, the suggestion is that the information about the inserted silent time between two questions should be clarified for part 1. The analysis of the audio file of the VSTEP.3-5 listening test used on 25th March 2017 shows that the time allotments strictly follow the predetermined descriptions in the test specification and the inserted silent time between questions in part 1 is 7 seconds is consistent throughout the items. From expert points of view and from the feedback to the survey questionnaire on the candidates taking the VSTEP.3-5 listening test on 25th March 2017, with the specific time allotments, the test takers have enough time to process the information and select their answers.

As for scoring method, the VSTEP.3-5 listening test uses the multiple-choice format. The listening raw score is based on the number of correct answers with each item equally weighted. How an item is scored is made explicit to the test-takers. However, how the raw scores are converted to the scores on the scale of 10 is not made explicit to the test-takers. Thus, the suggestion is that a public score conversion table for the VSTEP.3-5 listening test should be published to the test-takers and put available onto the website of the Center for Language Testing and Assessment, ULIS-VNU (vstep.vn) so that test-takers can know the relative value of each task in the test they are going to take or have taken.

(2) The input to which test-takers are exposed resembles normal everyday speech and are designed in accordance with the descriptions in the test specification.

The strengths of the VSTEP.3-5 listening input can be described as follows. First, the topics for the input are not sensitive or biased. Second, the recorded materials combine both monologue and dialogue texts, ranging from listening to announcement, instructions, voice messages to conversations, talks or lectures. These resemble the types of listening that test-takers usually have to deal with in the real-world context. Third, the recorded materials are sensitive to the difficulty of normalizing to unfamiliar L2 voices. Where there is a dialogue, speakers are of mix genders (a female voice and a male voice). Fourth, in terms of speaker accent, the VSTEP.3-5 listening test takes a commendably moderate approach to the use of variety. The recordings for the tests have a familiar variety of English accent such as standard American English, British English or Australian English. Fifth, the language used in the VSTEP.3-5 listening test texts generally follows the descriptions for intended levels predetermined in the test specification. And finally, the successful completion of the VSTEP.3-5 listening test task is dependent on comprehension of the text.

However, some weaknesses related to the input of the VSTEP.3-5 listening test can be addressed as follows. First, although as described in the specification, the VSTEP.3-5 test recordings need to employ voice actors/actresses with standard English in a variety of accent, American English features quite extensively in the recording of the VSTEP.3-5 listening test administered on 25th March 2017. Thus, the recommendation is that at the low level items of the VSTEP.3-5 tests (level 3) such as separate items in part 1, it is important to maintain standardization of accent to make sure that test-takers are spared apprehension about what they are likely to hear. The use of standard British English or American English is recommended. But for the intendedly high- level texts (level 4 and level 5), a range of accents should be utilized in the recordings since it is recognized that candidates at these levels are likely to have had greater exposure to the L2.

Second, in terms of input origin, all of the VSTEP.3-5 listening test texts are entirely scripted and semi-scripted before being given to voice actors and actresses for practice and being recorded. For low level texts

and questions in part 1, this seems a necessary method, given the need to control for linguistic content and for clarity of delivery. For higher level texts and questions in part 2 and 3, more authentic materials should be employed, especially for texts of intended level 5. If the materials cannot be recorded live, they should be made naturally with speech rates, pauses, hesitations, false starts and speaker overlaps since all of these features characterize the types of speech that candidates at higher proficiency levels should be capable of handling. In order to have these qualified products of authenticity, ULIS-VNU should employ voice actors and actresses who have experience or have been professionally trained.

Third, speech rate is a further aspect that need to be taken into consideration for the VSTEP.3-5 listening test. In generally, the speech rates for the recordings in the VSTEP.3-5 listening test administered on 25th March 2017 are slower than expected. Slowed listening audio input of the test can distort the cadences of natural speech and as result, the naturalness of the listening can be affected. As mentioned, all of the listening texts of the VSTEP.3-5 listening tests are recorded in the studio, it is important to give voice actors and actresses time for practice and trial recording before real recording. They should be provided with the samples of recordings for the level or the speech rates they need to act.

(3) The successful completion of the test tasks is dependent on the comprehension of the text.

The third assumption is addressed in relation to the relationship between the input and response. The analysis of the test booklet together with its scripts and recordings used on 25th March 2017 reveals that the successful completion of 35 questions in the test is dependent on comprehension of the texts. Although the questions of the test are in the format of traditional multiple-choice, all of the test items are well-designed with options giving no clue for the clever test-takers to respond without listening. Besides, the analysis shows that questions and the relevant texts of the test are closely independent, which means that the comprehension tasks of the test are designed in the way that they provide clear evidence of comprehension.

1.2. To what extent were the VSTEP.3-5 listening test scores reliable in measuring the test takers' English proficiency?

(1) The test reliability is in the good range for proficiency assessment.

First, for the first assumption, the recorded report history of the VSTEP.3-5 listening test provided by the Center for Language Testing and Assessment shows that the test reliability always maintains in a good range. The data analysis of the result of the VSTEP.3-5 listening test administered on 25th March 2017 has confirmed this statement (0.815), which means that the VSTEP.3-5 listening test administered on 25th March is sensitive enough to distinguish between the low and high performers and the test contain sufficient items to measure the test-takers' listening proficiency. Moreover, the item separation and item reliability of the test implies that the person sample is large enough to confirm the item difficulty hierarchy (=construct validity) of the testing instrument.

(2) The test items are well-designed with the predetermined difficulty as described in the test specification and with good discrimination.

For the second assumption, there are a number of backings that are both for and against this assumption. First, in generally, the items of the VSTEP.3-5 listening test administered on 25th March 2017 are fairly well written with the predetermined level of difficulty described in the test specification. The overall discrimination of the test show that the test in general can discriminate test-takers reasonably well.

However, the analysis of the item map reveals that the items of the test are not evenly distributed on the continuum of difficulty and this results in the fact that there is a waste of items at some level of difficulty since they measure the same thing and there is a lack of items at some level of difficulty since there is no item in that place to measure that competency. Besides, the item-by-item analysis show that the VSTEP.3-5 listening test contains some items are statistically easier or more difficult than expected and some has low discrimination indices that need to be rejected or revise. The recommendations are that (1) the specification

of the VSTEP.3-5 listening test needs improvement for better guidance. To be more specific, more details about the difficulty level and the use of language in each item need to be clearly explained enough so that the items can be designed with the right intended level of difficulty. (2) Before being used in the official test forms, the test items need to be piloted if affordable in a scale of 200 people or otherwise piloted on a small scale of 50 so as to make sure that all of the items satisfy the expected difficulty levels and can discriminate test-takers well. These two recommendations aim at ensuring the equivalent test tasks and test forms for different examinations.

1.3. To what extent were the cut scores of the VSTEP.3-5 listening test properly established?

(1) The standard-setting procedures used for the VSTEP.3-5 listening test were properly implemented.

For the first assumption, an investigation into the documentation of standard-setting procedures for the VSTEP.3-5 listening test reveals that the cut-scores for the VSTEP.3-5 listening test were established by Angoff method. The cut-scores were once set for the model test and used for other forms of VSTEP.3-5 listening tests that are supposed to be built based on the test specification of the model one. As explained in the documentation, the Angoff method was chosen to set the cut scores for the VSTEP.3-5 listening test since it is a widely accepted method which is easy to be implemented and it can be applied to MCQ tests. The process was reported to be conducted with the strict procedures.

(2) The reliability of the VSTEP.3-5 listening test is within the acceptable range for 3 cut-off points.

The second assumption for warrant 3 is that the reliability of the VSTEP.3-5 listening test is within the acceptable range for 3 cut-off points. Although the reliability of the VSTEP.3-5 listening test administered on 25th March (0.815) is judged to be very good for a normal proficiency assessment, this value is still a little bit low for a high-stakes test such as the VSTEP.3-5 listening test with 3 cut-off points. The required value of test reliability for a high-stakes test with 3 cut-off points should be 0.88. Thus in order to attain the expected reliability index of 0.88, the VSTEP.3-5 listening test forms need to be improved by revising the test items.

(3) The cut-scores established through another standard-setting method correlate well with the current cut-scores set for the VSTEP.3-5 listening test.

As for the third assumption that the cut-scores established through another standard-setting method correlate well with the current cut-scores set for the VSTEP.3-5 listening test, the Bookmark method is used to cross-validate the cut-score results of the test. The results of the cut-scores established by the Bookmark standard-setting method reveal that the Bookmark ratings are consistently higher than the results from the Angoff ratings. The final results are left open to the entity that is responsible for choosing to adjust the panel's recommended cut-scores on purely policy, political, or economic grounds.

2. Implications of the study

First, it is obvious that increasing the number of judges can reduce the variation in the cut-scores and ultimately increase the precision of the panel's cut-scores. As mentioned in the literature review, an acceptable level of dependability for the cut-scores could be reached in many contexts if 10-15 judges are employed and in educational assessments, 15-30 is generally viewed as acceptable because of the desire (1) to have both broad representations on panels setting standards and (2) to obtain stable results (Hambleton & Pitoniak, 2006). The more important a test is, the greater number of judges is needed. For a high-stakes test as a VSTEP.3-5 test, in the case of employing the Angoff method, the number of judges should have been raised since the number of judges was not sufficient.

Second, if the cut-scores are set once and used for all as in case of the VSTEP.3-5 listening, it is important to conduct a cross validation of a standard setting evaluation for the test. As Hambleton (1980), Koffler (1980), and Shepard (1980, 1984) suggest, it might be prudent to use several methods and then consider all of the

results when determining the final cut-scores. However, it would be expensive, time-consuming and effort-consuming to carry out too many different methods at the same time, it can be solved by conducting two panels of the same standard setting method and compare two results before the final cut scores are decided.

Third, in order to ensure all of the VSTEP.3-5 listening test forms are equivalent across examinations, the Center for Language Testing and Assessment, ULIS-VNU as well as other institutions which are allowed by the Ministry of Education and Training, Vietnam to develop and administer the VSTEP.3-5 test in general and the VSTEP.3-5 listening test in particular, need to establish equivalent test forms used for different examinations. In order to achieve this, first, they need to make sure that the test is aligned with the specification of the test blueprint. Second, they need to ensure statistical equivalence. Statistical equivalence can be controlled by carefully selecting items when constructing the test forms.

3. Limitations of the study

As being the very first investigation into the validity issue of the VSTEP.3-5 in general and the VSTEP.3-5 listening test in particular and although the study has reached its aims, there are some unavoidable limitations.

First, because of the time limit, this study could only be conducted on a specific size of population and with one specific VSTEP.3-5 listening test form administered for one examination on 25th March 2017. Therefore, to generalize the results, the study should have taken more examinations into investigation.

Second, this study has some statistical limitations with regards to participants invited for the Bookmark standard-setting. As advised by Hambleton & Pitoniak (2006), when setting the cut-scores on a statewide assessment of test-takers like the VSTEP.3-5 test, a panel of participants should be composed of representatives of all groups who are knowledgeable in the field and who have a legitimate stake in the outcome of the assessment and the decision that will derive from its uses such as classroom teachers, curriculum specialists, university presidents, Ministry of Education and Training, business community, parents or college admission officers. However, this study could just involve the representatives of English teachers and admission officers working in ULIS-VNU. Thus, in order to obtain a more valid result of the cut-scores for the VSTEP.3-5 listening test, this study should have involved judges from more various positions and ones from outside ULIS-VNU.

4. Suggestions for further research

First, with regards to the MCQ test format like the VSTEP.3-5 listening test, a study could be conducted to make an investigation into the processes activated and strategies employed by the test-takers when taking the VSTEP.3-5 listening test and how these listening processes and strategies are related with the test-takers' success in solving the listening comprehension items.

Second, further studies into the VSTEP.3-5 listening test may be expanded into other administrations as well as other test booklets in order to build up a comprehensive validity argument for the VSTEP.3-5 listening test in particular and the VSTEP.3-5 test in general.

Besides, in terms of the standard setting, a number of other studies could be suggested for the VSTEP.3-5 listening test as follows. First, the reliability and validity of the placement decisions based on the cut-scores of VSTEP.3-5 listening test results can be taken into consideration by making some correlation analyses with other relevant criteria such as the grades of the test-takers on a English course or the test-takers' performances on other tests of the same measurement. Besides, the agreement of the cut-scores set for the results of the VSTEP.3-5 listening test can be investigated across replications by using other standard-setting methods such as Nedelsky method, Ebel method, or Borderline group method. In addition, a study could be about how the perspectives of different participants such as academic experts, teachers and other test users influence the results of cut-scores for the VSTEP.3-5 listening test. Furthermore, other issues related to standard setting could be taken into consideration in the further research such as how much difference the choice of a rounding approach makes to the result of the cut scores and how the precision of the cut-scores

established can be ensured by improving the participant training.

Finally, since several institutions throughout Vietnam are being permitted by the Ministry of Education and Training to develop and administer the VSTEP.3-5 test, various studies could and should be conducted into the above-mentioned issues in these institutions and across institutions.

REFERENCES

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1985). *The Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

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

Alderson, J. C, Clapham, C. M & Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: Cambridge University Press.

Angoff, W.H. (1988). Validity: An evolving concept. In H. Wainer & H.I. Braun (Eds.), *Test validity (*pp. 19-32), Hillsdale, NJ: Lawrence Erlbaum.

Anastasi, A. (1938). Faculties Versus Factors: A Reply to professor Thurstone. *Psychological Bulletin, 33,* 391-395.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Antastasi, A. (1986). **Evolving concepts of test validation.** *Annual Review of Psychology*, 37 (1986), pp. 1-15

Bachman, L. F. (*1990*). Fundamental Considerations in Language Testing. Oxford: Oxford. University Press.

Bachman, L. F. (2004). Statistical Analyses for Language Assessment. Mahwah, N.J.: Cambridge University Press

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 19, 4, 453 - 476.

Bachman, L. F. and *Palmer*, A.S. (*1996). Language testing in practice: designing and developing useful language tests.* Oxford: Oxford. University Press. 384 pp.

Baker, J. (1997). Study equality. In Iain MacKenzie and Shane O'Neill (eds.), *Reconstituting Social Criticism.* London: Macmillan, pp. 51-64.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. New York: Harper.

Bloomfield et al (2011) _ trong cuon examing listening (cam)

Brown, H. D. (*1987*). Principles of Language Learning and Teaching. Englewood Cliffs, NJ: Prentice Hall.

Brown, J. D. (1989a). Criterion-referenced test reliability. *University of Hawai'I Working Papers in ESL,* 1, 79-113.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced Language Testing.* Cambridge: Cambridge University Press

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

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 98-106.

Cizek, G. J. (1996a). Setting passing scores. *Educational Measurement: Issues and Practice, 15* (2), 20-31.

Cizek, G. J. (1996b). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15* (1), 12-21.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics,* 19: 254-72.

Chapelle, C.A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a Web-based ESL test. *Language Testing*, 20(4).

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

Chapelle, C. A., Enright, M. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational measurement: Issues and practice, 20(1), 3-13.*

Cronbach, L. J. (1955). Process affecting scores on "understanding others" and "assumed similarity". *Psychological Bulletin, 52,* 177-193.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New york: Harper & Row.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New york: Harper & Row.

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

Crooks, T., Kane, M., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education, 3,* 265-285.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Ebel, R. L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement, 16,* 179-197.

Elliott and Wilson. (2013) in Examing listening Emslie & Emslie (2005:13)

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language testing, 14(2),* 113-139.

Garrett. (1937). *Statistics in psychology and education.* New York: Longmans, Green.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.

Guilliksen, H. (1950a). *Theory of mental tests.* New York: Wiley.

Guilliksen, H. (1950b). Intrinsic validity. *American Psychologist, 5,* 511-517.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K. (1998). Principles and selected applications of item response theory. In R. L. Linn (ed.), *Educational measurement* (3rd ed., pp.147-200). New York: American Council on Education and Macmillan.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines . *European Journal of Psychological Assessment 17,* 164-172 .

Hambleton, R.K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hambleton, R.K., & Powell, S. (1983). A framework for viewing the proccess of standard setting. *Evaluation & the Health Professions, 6(1),* 3–24.

Hansche, L. N. (1998). *Handbook for the development of performance standards: Meeting the requirements of Title I*. Washington: U.S. Department of Education & The Council of Chief State School Officers.

Howatt, A.P. R. (1984). *A History of English Language Teaching.* Oxford University Press.

Huynh, H. (2000). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on Bookmark and item mapping. *Educational Measurement: Issues and Practice, 25(2),* 19-20.

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp.35-71). New York: Holt, Rhinehart & Winston.

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

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health professions, 17,* 133-159.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18(2),* 5-17.

Kane, M. T. (2001). So much remain the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards. Concepts, methods, and perspectives* (pp. 53-88). Mahwah, N. J.: Lawrence Erlbaum Associates.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21(1),* 31-41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2(3),* 135-70.

Kane, M. T. (2006). Validation. In R.L. Brennan (ed.), *Educational Measurement* (4th ed). Washington, DC: American Council on Education/Praeger, pp.17-64.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R.W. Lissitz (ed), *The Concept of Validity: Revisions, New Directions and Applications.* Charlotte, NC: Information Age Publishing, pp. 39-64.

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

Kelly, T. L. (1927). *Interpretations of Educational Measurements*. New York: World Book Company.

Lee, Y. J., & Greene, J. C. (2007). The preductive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed methods Research, 1,* 366-389.

Linn, R. L., & Miller, M. D. (2005). *Measurement and Assessment in Teaching* (9th ed). Upper Saddle River, NJ: Pearson.

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

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix.* Cambridge, UK: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*

(pp. 13-103). New York: American Council on Education.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). A framework for reusing assessment components. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.), *New developments in psychometrics* (pp. 28-288). Tokyo: Springer.

Morgan, D. L., & *Michaelides*, M.P. (*2005*). *Setting cut scores for college placement* (College Board Research Report No. *2005*–9). New York,NY: The College Board.

Mosier, C. L. (1947). A critical examination of concepts of face validity. *Educational and Psychological measurement, 7,* 191-205.

Pearson, K. (1896). Mathematical contributions to the theory of evolution III: Regression, heredity, and panmixia. *Philos. Trans. Roy. Soc. London Ser. A*, 187, 253-318.

Pitoniak, M. J. (2003). Standard setting methods for complex licensure examinations. In G. J. Cizek & M. B. Bunch (2007), *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Sage Publications

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and etraining participants for standard setting. In G. J. Cizek (Ed.), *Setting*

performance standards. Concepts, methods, and perspectives. Mahwah, N. J.: Lawrence Erlbaum Associates.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives (*pp. 159-174). Mahwah, NJ: Lawrence Erlbaum.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16,* 290-296.

Schmitz, C. C., & *delMas*, R. C. (*1991*). Determining the validity of placement exams for developmental college curricula. *Applied Measurement in Education, 4*, 37-52.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammon (ed), *Review of research in Education, Vol. 19.* Washinton, DC: AERA.

Spearman, C. (1904). General intelligence: objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Thurstone, L. L. (1932). *The reliability and validity of tests.* Ann Arbor, MI: Edwards Brothers.

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

Toulmin, S. E. (2003). *The use of argument (*updated ed). Cambridge: Cambridge University Press.

Truman, W. L. (1992). College placement testing. *AMATYC Review, 13*, 58-64

Vandergrift, L. & Goh, C. C. M. (2012). *Teaching and Learning Second Language Listening: Metacognition in Action,* New York: Routledge.

Wall, D., Clapman, C., & Alderson, J.C. (1994). Evaluating a placement test. *Language Testing, 11*, 321-344.

Weir, C. J. (2005). *Language Testing and Validation: An evidence-based Approach.* Basingstoke: Palgrave Macmillan.