



Faculty of Language Studies and Communication Studies

**The Validity of the Advanced Educational Program English Test at a
Public University in Vietnam**

Nong Thi Hien Huong

**Doctor of Philosophy
2018**



1000282682

The Validity of the Advanced Educational Program English Test at a Public University in Vietnam

Nong Thi Hien Huong

A thesis submitted

In fulfilment of the requirements for the degree of Doctor of Philosophy

(Applied Linguistics)

Faculty of Language and Communication
UNIVERSITI MALAYSIA SARAWAK
2018

CHAPTER 1

INTRODUCTION

1.1 Chapter Overview

This chapter begins with introduction to the study and explains aims of the present study. Then, the background to the study, statement of the problem and research questions are presented. Finally, it outlines chapters in the thesis.

1.1 Introduction

In the globalized world, all employers on a national as well as on an international scale always pay much attention to the foreign language skills of their future employees, thus it can be noted that being able to speak one or more foreign languages is a prerequisite for each individual who expects to get a good job in the future. The most dominantly popular foreign language is English which is increasing gaining an important position in several countries all over the world. English is not only a means but also an important key to gain access to the latest scientific and technological achievements for developing countries such as Vietnam, Laos, Cambodia, Thailand and Malaysia.

Furthermore, it can be observed that the number of native English speakers is from 400 million to 500 million; more than one billion people who have been using some forms of English. Many people have used English for different purposes, English has become the official or second language of a majority of the countries around the world including India, Thailand, Singapore, Malaysia and Vietnam (Kluitmann, 2008; Roengpitya & Saelim, 2015).

In Vietnam, the Vietnamese Government has identified the urgent social – political, commercial and educational need for Vietnamese people to be able to better communicate in English. In line with this aspiration, all Vietnamese tertiary institutions have accepted English as a compulsory subject as well as medium of instruction for academic purposes (Canh, 2009; Le, 2011; Nha, 2014; Tran, Griffia & Nguyen, 2010; Van, 2007). This development has given rise to the need to teach and measure students' command of English at institutional level. However, the issue that is often raised in relation to in-house language test is the validity of the test.

According to Bachman (1990), test validation is the process of producing validity and reliability evidence to support the well-foundedness of inferences concerning trait from test scores, i.e., principally, testing should be concerned with evidence-based validity. Test writers need to provide a clear argument for test validity in assessing specific traits with reliable evidence to support the plausibility of this interpretative argument (Kane, 1992). Therefore, test validation has been regarded as the most important role in test development and use and should always be examined at different educational levels (Bachman & Palmer, 1996). Similarly, Cumming, Alister and Berwick (1998) also point out that validation in language testing and assessment plays a central role in academic decisions, pedagogical practices and educational policies. In a nutshell, validation is an effort to improve the quality of education, as it reveals strengths and weaknesses in the curriculum, appropriacy of the programme, students' promotion as well as teachers' evaluation.

Validity and reliability are the two most essential components of test validation as well as two most important characteristics of a good test. Test validity measures what is intended to

measure while reliability is an assessment tool which produces consistent results (Bachman & Palmer, 1996). Although a reliable test may not be necessarily valid, a valid test is always reliable (Bachman, 1990; Henning 1989) because a test cannot be considered as valid unless its measurement is reliable. Therefore, a valid test encompasses reliability elements. Several previous test validation studies have investigated the validity of their in-house language tests and make valuable contributions to understand the concepts of test validation (Advi, 2003; Ayers, 1977; Cumming, 2004; Dooey & Oliver, 2002; Fulcher, 1997a; Sar, 2008; Huong, 2001 ; Lee & Green, 2007; Mojtaba, 2009; Moritoshi, 2001; Nakamura, 2006; O'Sullivan, Weir & Saville, 2002; Pishghadam & Khosropanah, 2011; Riari, 2013; Siddick, 2010; Wilson, 1999; Zubairi, 2001).

On the other hand, there often have been a lot of gathered concerns in relation to the validity of in-house language tests from the educators and administrators who always raise a question of whether the in-house test faithfully reflects the course objectives, expected linguistic knowledge, skills as well as measures the real linguistic competence of test-takers. Similar concerns also occur to the case of the AEPET, which is an important in-house language test of public universities in Vietnam. Bearing this in mind, the study aims to validate the validity of AEPET by focusing on three crucial validity components: concurrent, predictive and content validity. The study aims to examine the concurrent validity of the AEPET by looking at the relationship between the AEPET score and a standardized international language test score, which is International English Language Testing Services (IELTS). Furthermore, the study investigates the predictive validity of the AEPET by addressing the question to what extent the AEPET determines academic success of students. Last but not least, the study focuses on content validity of the AEPET by accessing English language lecturers' judgments about the

AEPET in order to see whether the AEPET content reflects the knowledge and skills mentioned in the Advanced Educational Program (AEP) syllabus, as well as to find out to what extent AEPET preparation adheres to crucial test preparation procedures before the examination is administrated.

1.2 Background to the Study

The background of this research covers three areas related to context where the data for this study were collected. The first area focuses on the status of teaching and learning of English language in Vietnam. The second section introduces the Advanced Educational Program (AEP) conducted at the research site, which is a public university in Vietnam. The last area presents information about the English course in the AEP.

1.2.1 English Language Teaching and Learning in Vietnam

Vietnamese Government considers education and training its top driver of socio-economic development. Thus, in recent years the Government has proposed and implemented many new policies to promote education and training and to meet the demands of industrialization, modernization, and international integration (Van, 2007).

The former Vice Chairman of the National Assembly's Committee for Culture, Education, Youth, and Children said that "in order to fundamentally and drastically reform the educational system, it is important to teach English as a second language at schools. We have to encourage innovative ideas in the content and teaching English method. If we are afraid of doing something wrong, we cannot reform" (as cited in Canh & Barnard, 2009, p. 22). He stated that English plays an important role as an international language because it is the key to

open a brilliant future for each country in general and for Vietnam in particular. Crystal (2000) also stated that English language is used in a number of countries, serves sometimes as the first language, sometimes as the second language or official language and sometimes as a foreign language. Sharing the same view, Hung (2013) and Van (2007) indicate that English ranks first with a much greater influence than other foreign languages such as French, Chinese and Russian in Vietnamese educational system. English has become an international language and is the key for success in business, commerce and technological science.

Bearing this in mind, Vietnamese Government introduced English nationally as a compulsory subject both at upper-high schools and secondary schools and as an elective subject at primary schools in Vietnamese general educational system. Vietnamese educational system consists of three levels with 12 grades: primary level (from grade 1 to 5 for children aged 6-11); upper-secondary level (from grade 6 to 9 for children aged 11-15); and upper- high school (from grade 10 to 12 for children aged 16-18).

At all Government schools and universities, English classes are divided into two semesters within 35 weeks per year. At the primary level, English is taught for 2 periods a week from Grade 3 to Grade 5 (age 8-10), making a total of 70 periods per year for each grade. Each period consists of 40 minutes. At the upper-secondary level, English is taught 3 periods per week from Grade 6 to Grade 8 (age 11-13), totaling to 105 periods per year for each grade. In contrast, it is taught 2 periods per week in Grade 9 (age 14), making the total 70 periods per year. At the upper-secondary level (age 15-17), English is taught 3 periods a week, making the total of 105 period per year for each grade. For both secondary upper-secondary levels, time for each period is 45 minutes.

At higher education level (age 18-22), students who study a normal class in which Vietnamese is the medium of instructions, English is taught with 5 periods a week, making the total of 175 periods. In contrast, students who study in advanced educational class in which English is the medium of instructions, English is taught for 15 periods a week making the total of 525 periods. Each period consists of 50 minutes. Table 1.1 presents the number of English periods which is taught at each level of education in Vietnam (Van, 2007).

Table 1.1: Number of English Periods in Vietnamese Educational System

Level of education (Grades)	Period duration	Periods per week	Weeks per year	Periods per academic year
Primary (G3- G4-G5)	40	2	35	70
Secondary (G6-G9) G6- G7 -G8		3	35	105
G9	45	2	35	70
Upper-secondary (G10- G11-12)		3	35	105
University Normal class	50	5	35	175
Advanced class		15	35	525

For schools, the English materials, textbooks designed by Vietnamese Minister of Education Training (MOET) are applied and taught in class in order to facilitate students to use English as a means of communication at certain level of proficiency in four skills: Listening, Reading, Writing and Speaking. This series of textbooks was designed and produced by a group of Vietnamese textbook writers in MOET from 1989 to 1992 and was introduced into Vietnamese secondary schools stage by stage during those four years (Nguyen & Crabbe, 2007). They are being used across the country until now. MOET emphasizes that, by the end of upper-high schools, students are expected to obtain a standard level of understanding English and other multi- cultures in the world in order to become aware of cross-cultural differences, to be better overall communicators, to introduce about Vietnam people, histories, cultures and customs to the world (Cahn et al., 2009; Huong & Pham, 2010; Van, 2007; Neha & Burns, 2014). For universities, the selection, evaluation and use of suitable English

materials, textbooks are decided and approved by *the* university administrators in order to better adapt the students' specific needs to an increasingly globalized environment.

In brief, Vietnamese educational reform in this time is totally suitable for catching up with the development taking place in the world. Command of English is crucial to meet the urgent socio-political, commercial, educational needs of Vietnamese people and enable them to have a better communication and interactions at global level.

1.2.2 The Advanced Educational Program (AEP)

The study is conducted at a public university in Vietnam. The university which is the leading core university in the north of Vietnam follows the Government decision to offer the Advanced Educational Program in collaboration with different educational institutions in the United States since the year 2000. The Advanced Educational Program is an educational program which is the cooperation between public universities in Vietnam and several educational institutions in the United States since the year 2000. The programme has been endorsed by Vietnamese Government under Decision No. 14/2000/ QD-MOET entitled “Developing the Advanced Educational Program in the National Education System in Vietnam” with a common target of building, developing different educational areas and renouncing the universities in Vietnam to the world ranking (Government Decisions, 2001.)

The first specific goal of the AEP is to build up international undergraduate programmes that can attract students to study in the programme and compete with current abroad study movement both in Vietnam and overseas. Secondly, this programme provides for students with knowledge, self-study and research capacity and technical skills to meet with the

requirements of their future career or higher education related to different aspects serving for different purposes in all the country. Last but not least, this program also helps students to improve their awareness and moral attitude to ensure the loyalty with the Socialist Republic of Vietnam, being healthy and responsible for the industrialization and modernization of Vietnam.

This programme is built according to the United States standards and directly taught in English by Vietnam and overseas lecturers and professors to ensure students gain high quality learning outcomes. The AEP offers a four-year programme in which students are required to attend one-year English course before they are qualified for a 3 year of degree programme. All subjects are taught in English. For the first academic year, students are required to complete a one – year English course which prepares students English language skills and knowledge before they start their major coursework. This English course is run over two semesters. After the one - year English course, qualified students, who obtain IELTS overall band 6.5 and above and a minimum CGPA of 7.0, will be selected to continue their degree programme in the United States. Their bachelor degrees will be offered by the university in the United States. Other students will continue their degree at the home university and after the three years of degree programme, will achieve bachelor degree offered by Vietnamese universities. The structure of the AEP is summarized in Figure 1.1

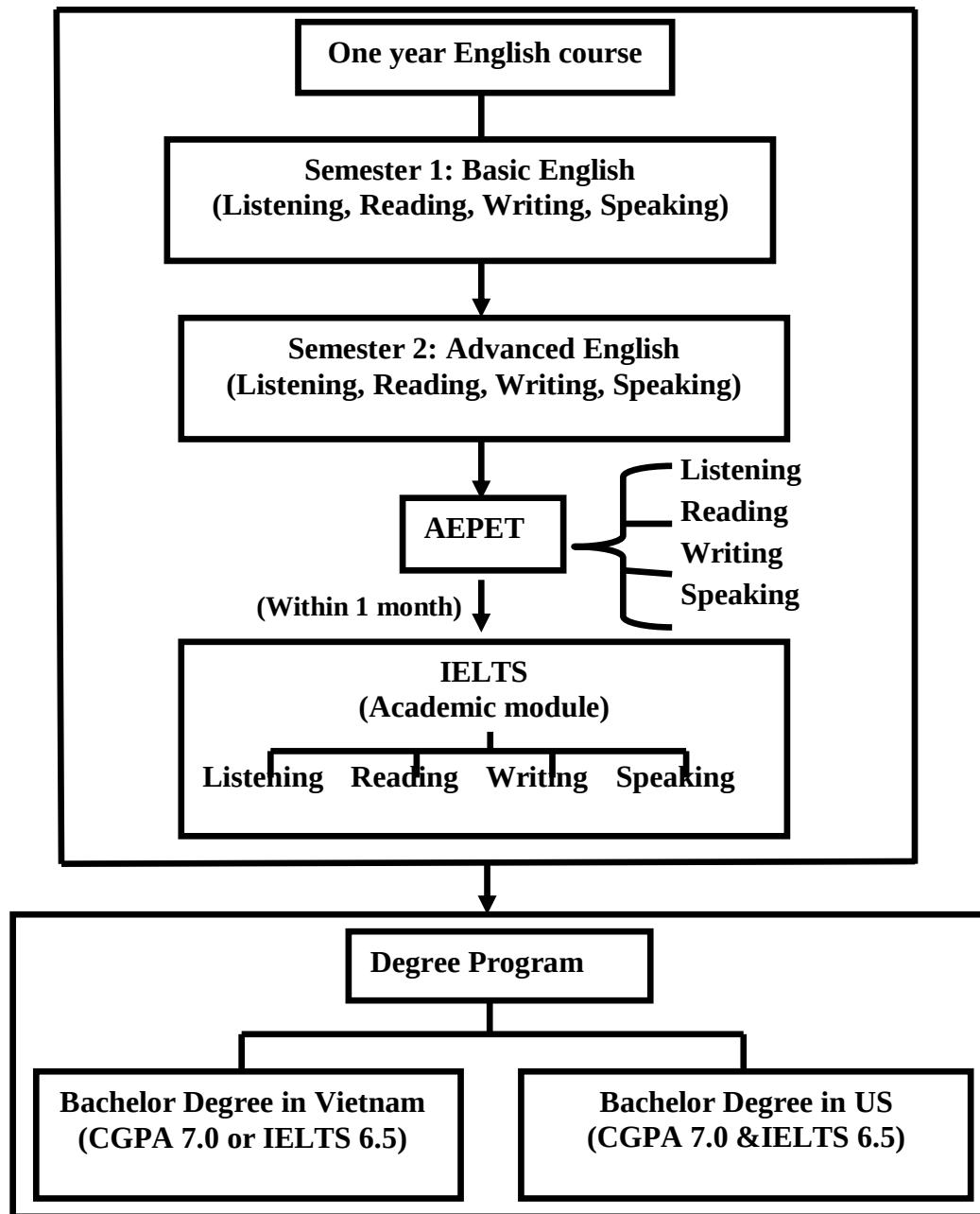


Figure 1.1: Structure of Advanced Educational Program (AEP)

In short, the AEP is the educational cooperative programme between universities in Vietnam and the United States. This programme is aimed to provide students sufficient level of English language proficiency, advanced knowledge in the educational aspects. The success of the programme helps students get more technical and advanced skills to meet the requirements of their career in the future.

1.2.3 English Course in the Advanced Educational Program

As mentioned earlier, the objectives of the AEP are to provide students not only high quality learning outcomes but also an international studying environment in which English is used as the medium of instruction. In order to meet the requirements of these objectives, the AEP requires students to attend one-year English course before they qualify for a 3 year degree programme. The one-year English course is divided into two levels: Basic English and Advanced English taught over two semesters.

In the first semester, students do four Basic English subjects: Listening, Reading, Writing and Speaking and Lang master which is an online practical English subject. Each Basic English subject takes 4 credit points. In other words, the total credit point for the first semester is 16. At the end of the first semester, students are required to take four English subtests-Listening, Reading, Writing and Speaking. In the second semester, five advanced English subjects: Listening, Reading, Writing, Speaking and Skills of IELTS are integrated in the classroom with the total credit points of 18.

At the end of the first academic year, students take the AEPET which comprises of four components: Listening, Reading, Writing and Speaking. This test is designed based on the whole content of the English course that students are taught in both semester one and semester two. Students have to sit for IELTS within one month after taking the AEPET. Students, who obtain at least 6.5 IELTS overall band scores and 7.0 CGPA and above, will have a chance to apply a full scholarship to do their degree programme at a university in the United States and their bachelor degrees will be offered by the university in the United States. On the other hand, the rest of them, who do not meet the criteria, will do their degree programme at the

home university and accordingly their bachelor degrees will be offered by the university.

Figure 1.2 presents a detailed explanation of phases that students go through in the English course

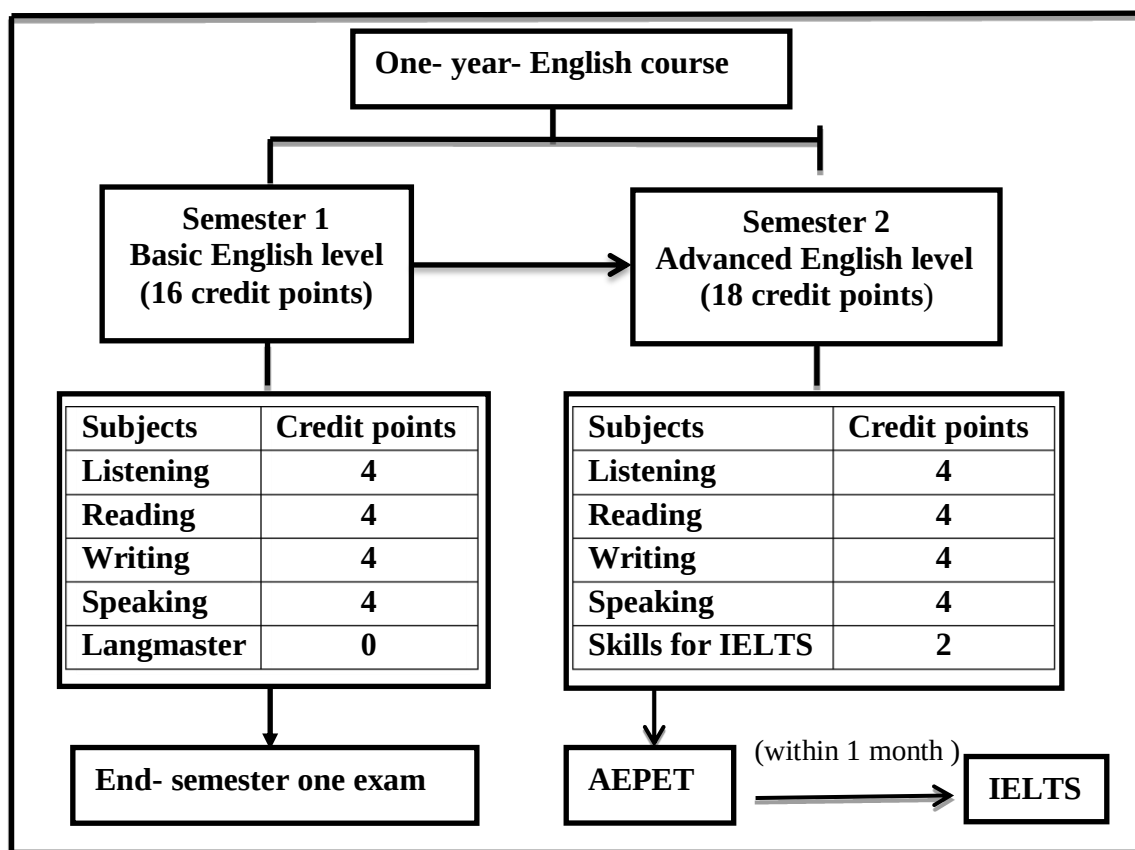


Figure 1.2: Overview of AEP English Course

In brief, the English course at the AEP is designed to provide specific and appropriate language instruction for AEP students who intend to undertake their advanced studies at the university. Successful completion of the course means that the students have a sufficient level of English language proficiency to do their degree in English as well as to provide them opportunities to undertake their studies at another university in the United States

1.3 Statement of the Problem

English has become a global language of technological, business and academic endeavors. The world-wide globalization process has confirmed English as the most widely-used means of international communications and an empowering tool which assists learners to achieve specialized knowledge, which helps them to be prepared primarily for entering the working world. More precisely, English is taught at global level as the second language in India, Indonesia, Korean, Japan, Malaysia, Singapore, Philippines, Thailand, Vietnam and so on. The educators use standardized international language tests such as IELTS, TOEFL, TOEIC, KET, PET, FCE as a global language testing standards. Although these standardized international language tests are always available, they are rather costly and may not always be appropriate for the specific needs. Therefore, at present, many public universities all over the world have taken English language proficiency seriously into consideration when making decisions on different academic purposes by designing their own language tests which are cheaper to the administrators and appropriate to address their specific needs. Various in-house language tests are designed and used by several institutions such as Malaysian University English Test (MUET) in Malaysia; Qualifying English Test (QET) in Singapore; College English Test (CET) in China; General English Proficiency Test (GEPT) in Taiwan, Iran, Ethiopia, Nigeria. However, the main issue that is often raised in relation to the validity of in-house language test is validation.

According to Bachman and Palmer (1996), test validation plays the most important role in test development and use because test validation helps to produce validity evidence for a test, show how much the educators and administrators know how much knowledge their student achieved during a course or which subjects the student are not good at, and then make

immediate chances in syllabus. Several researchers highlight that a valid test helps to reach the desired objectives of teaching, improve the quality of education by reflecting strengths and weaknesses in the curriculum, program appropriations, students' true language abilities (Bachman, 1990; Bachman & Palmer, 1996; Cumming & Berwick, 1998, Huong, 2001; D'Este, 2012; Wilson, 2016).

In Vietnam, the AEP has been introduced in the national educational system since the year 2000 (Hung, 2013; Minh, 2008) and the AEPET which is an in-house language test is approved as an English language proficiency indicator for the students who expect to enroll the advanced programs at university. However, the validity of the AEPET has been an open question due to the fact that no empirical test validation studies have been published, thus the validity evidence of the AEPET is not available. The main concern is that if the AEPET is not valid or in other words AEPET components : Listening, Reading, Writing and Speaking cannot measure students' English skills appropriately, then how could students' true English abilities be improved?. Will students think that they are wasting time and money to study the AEP at university?. Another concern is that university language lectures seem to assume that testing and assessment is not their concern, that their job is to teach well, and that testing and assessment is something to be taken care of by a special person within their university who is responsible for testing or by external, invisible means. It should be noted that if the test is designed by the educators who do not participate in teaching and are not be familiar with the test content, test techniques as well as students' individual influences such as race, gender, ethnicity, culture, language background, level of education in the first and the second languages, stage of cognitive development, learning style, the designed test will not measure what it is intended to measure. Consequently, students' true abilities are not always reflected

in the test scores that they obtain. These problems might lead to the lack of validity of the AEPET. Last but not least, several researchers highlight that a less valid test cannot reach the course objectives, and as a result, will spoil the whole process of educational system (Lee & Greene, 2007; Messick, 1996; Siddiek, 2010), thus it is suggested that the validity of the in-house language test should always be examined (Bachman & Palmer, 1996; Henning, 1998; Messick, 1996; Siddiek, 2010). The more evidence about the validity of the in-house language test in use, the better and the more reliable the test becomes.

In the light of these considerations above, the researcher has been inspired and motivated to investigate the validity of the AEPET at a public university in Vietnam with the belief to gather more validity evidence of the in-house language test, provide students with true English abilities and put the forms of teaching and learning English in the right track of the educational process.

1.4 Research Objectives

The aim of this study is to investigate the validity of the AEPET at a public university in Vietnam by using Weir's (2005) test validation framework as a guideline. The study focuses on two crucial validity components: criterion-related validity (concurrent validity and predictive validity) and content validity. Further discussion on Weir's (2005) framework is presented in Chapter 2 and Chapter 3. The specific aims of this research are

- 1) To examine the extent to which the AEPET correlates with a standardized international language test (IELTS)

- 2) To investigate to what extent the AEPET score determines the academic success of AEP students
- 3) To determine lecturers' judgments about the AEPET in order to see whether the AEPET content reflects the knowledge and skills mentioned in the AEP syllabus as well as to find out to what extent AEPET preparation adheres to crucial test preparation procedures before the examination is administrated.

1.5 Research Questions

The study aims to determine the validity of the AEPET at a public university in Vietnam in terms of the concurrent validity, predictive validity and content validity. The study intends to answer the following research questions:

- 1) What is the relationship between the students' AEPET score and IELTS score?
(Concurrent Validity)
- 2) What is the relationship between the students' AEPET score and academic achievement (CGPA), in comparison with the relationship between IELTS score and CGPA? (Predictive Validity)
- 3) What is the Content Validity of the AEPET?

1.6 Significance of the Study

With the constant use of a locally designed language test for its different purposes, it is emphasized that validity becomes a property of the in-house language tests (Bachman, 1990;

Davies, 1989; Henning, 1998; McNarama, 2000; Messick, 1990, Cope, 2011; Siddick, 2010).

Therefore, this study can be valuable in a number of ways

For academic context, this study contributes useful insights to English language teaching and learning, especially in-house English test validation. Similarly, the present study sheds light on the review of literature on language testing practices, provide educators with more information related to test validation to make sure the process of test construction and test use reflects teaching methods among the teachers and true learning abilities among students. Another significance is that the present study is valuable for other educational institutions in their endeavor to validate in-house language tests, to justify the correctness of their interpretations, help test designers and educational decision makers to check to what extent the course content can be adequately represented in the test content by observing the distribution of the frequencies among the content areas for future exam construction. Therefore, the educators may take this study as a guideline to examine the quality of their locally-designed assessment tools.

In Vietnamese context, this study is undertaken with the hope to be the representative that provides information to Vietnam policymakers for future planning and innovation with regard to the advanced education at university level, especially helps foreign language policymakers, administrators, and test writers gain a better understanding of the development and challenges for the advanced educational program which has been introduced in the national educational system since the year 2000.

As mentioned earlier, the AEP is built according to United States standards, thus the study expects to provide some information about the appropriateness of applying and adapting the curriculum theories of American educators to the Vietnamese context through research evidence of the importance of the test validation implementation on local English language tests in order to strengthen the international relationship between two educational systems in the two countries. In case of public universities in Vietnam, this study may be providing the validity evidence for the AEPET at university level. If the AEPET is found to be valid, this could be the potential for the university to venture into the test validation, encourage students to improve their English skills and competencies which are required to succeed in the respective program and more importantly this could save the university's and students' time and budget.

1.7 Outlines of the thesis

The thesis is organized into five major chapters. Each chapter begins with an overview of what is included in that particular chapter.

Chapter 1 presents some basic information about introduction, background to the study, and then explains the purpose and the significance of the present study.

Chapter 2 consists of some theory evidence providing an overview of language testing. Then, it narrows to the qualities of good tests, there view of related previous studies and ends with research gaps.

Chapter 3 describes the methodology of the study. This chapter indicates the general direction of the methodology and then provides detailed information related to the stage methodology of the study.

Chapter 4 presents the results of the Main Study. Chapter 4 describes three main parts of the results section of the thesis: Concurrent, Predictive and Content Validity of the AEPET.

Chapter 5 presents a general discussion on the results of the study, answers the research questions. It also highlights some significant implications of the findings of the study on language testing, followed by a section which addresses some recognized limitations in this study, and also some suggestions for future research. The chapter ends with a concluding remark of the study.

1.8 Chapter Review

This chapter presented the background, research objectives and addressed the research questions of the study. It also elaborated the significance of the study. The next chapter, i.e. Chapter 2 presents the review of literature to this study.

CHAPTER 2

LITERATURE REVIEW

2.0 Chapter Overview

This chapter begins with a broad perspective of language testing definition; then continues with the relationship between teaching, learning and testing. The chapter presents the definition of test and then focuses on the qualities of good tests, concepts related to validity, theoretical framework for gathering evidence for validity and some previous test validation studies. The chapter ends with the rationale for investigating the validity of the AEPET

2.1 Language Testing

In the field of education, testing which is one form of measurements that plays crucial role in educational system because it is the final step in educational progress which helps the educators to measure the educational qualities. Language testing plays very important role in teaching and learning process because it helps the learners to grasp the knowledge that they have missed previously and the teachers to understand what can be done in subsequent lessons to improve their teaching methods. Well-designed in- house language tests can help the students learn the language by asking them to study hard, emphasizing learning outcomes, and showing them what levels and what knowledge they need to improve.

In the same vein, Huong (2001) and McNamara (2000) highlight some more important roles of language testing which have been applied popularly in educational system and in other related fields are to assist in revealing the strength and weakness in academic development,

creating positive washback for learning through providing the learners the feeling of competition as well as a sense that the teachers' assessment coincides with what has been taught to them, reflecting the students' true abilities, classifying student's language proficiency from other students as well as placing them in a suitable course. Therefore, several researchers suggest that the administrators and educators should build important and qualified testing strategies which assist in evaluating learners' performance, teaching methods, materials and other conditions in order to set up educational training objectives (Bachman, 1990; Davies, 1984; Fulcher, 1997a; Giri, 2011; Hancock, 2006; Hughes, 1989; Parhizgar, 2012)

In the light of literature review above, there are several number of overview texts on language testing providing a more comprehensive overview of the theory and practice of language testing. It can be said that language testing has assumed a prominent measurement in efforts to improve the quality of education because testing sets meaningful standards to educational systems, teachers, students, administrators and researchers with different purposes. In particular, this is an important part of every teaching and learning experience which helps to reflect overall teaching process and learning objectives. Furthermore, language testing has enriched the learning and teaching process by revealing strengths and weaknesses of the educational program, curriculum, teacher's teaching methods and evaluation as well as students' students' learning motivations. Through testing, when the administrators use language testing appropriately, they successfully measure the quality of education and then can make important decisions about the course, syllabus, course book, teachers, learners and administration

2.2 Relationship between Teaching, Learning and Testing

In the past, there was a common view that tests were separate from theory and practice; accordingly, teaching and testing; learning and testing were also separated from each other. This was due to the fact that testing was more concerned with competition rather than gaining learning abilities. It means that testing at that time tended to classify winners and losers in the test and drove students in attempting to get high marks rather than to gain knowledge. Furthermore, testing was observed as necessary but held an unpleasant location inside the educational programme because it helped to set up standards, but used up the valuable class time, therefore testing at that time did not relate to what students were taught. To solve these problems, Rudman (1989) suggested that educators should move test –driven competition in hands-on cooperation to make students become movers, thinkers and innovators who are able to develop higher-level and thinking skills, even in difficult situations.

On the other hand, when testing enters the age of scientific and technological development, it combines with teaching and contributes more positive attitudes towards the development of teaching process. Rudman (1989) emphasizes that “testing and teaching are not separate entities. Testing has always been a process of helping learners to revise the old knowledge and discover “new” ideas and “new” ways of organizing that which they learned”. (p. 3). In other words, testing has remained as an integral part of teaching. Similarly, Zucker (2003) also indicates that language testing is an important process of teaching and learning. The researcher explains that it is not too difficult to teach a course to the students, but it is very crucial to know how fruitful and successful the teaching is. The success of a language course can be observed through accurate measurement. An appropriate teaching method consequently will produce good testing and good testing will yield positive washback on both

learning and teaching process. For illustration, if the in-house English language tests are frequently constructed related to the analysis of the English language needs and include tasks that are similar to those which students normally perform in the class, the number of students who reach standards in English proficiency will be more than had ever been achieved before. Therefore, it can be said that language testing is a useful measurement device to assess learners' true language abilities and create positive or negative attitudes toward teaching and learning processes (Huong, 2001).

2.2.1 Testing and Teaching

In the field of education, knowing what the students have achieved from the teaching can be seen as a very important part in the evaluation of the objectives of educational programmes. Teaching is an activity that helps students to achieve objectives of the lesson, course or programme while testing is a procedure that helps students check what knowledge and what objectives they have already achieved. Therefore, it can be noted that teaching and testing have an association of partnership because teaching cannot work without being concerned constantly with testing.

Sharing the same view, Rudman (1989) emphasizes that testing has remained as an integral part of teaching because testing helps the teachers locate the strengths and weaknesses in the students' learning abilities. Hence, the teachers will have more chances to understand about the students' levels to design suitable tests. Secondly, test scores results early in class can aid the teachers make a plan to revisit their material resources and classify the issues to be confronted. Furthermore, investigating the results of a test can help teachers find the best methods as well as good materials to teach. Constantly, it helps the teachers in placing

students to right groups which are suitable to their levels, and then apply more new teaching and testing methods. In other words, testing supports in placing students in the class or groups for different learning activities.

Last but not least, the teachers will be able to control the leap of classroom instruction by preparing different types of tests in order to observe what the students have previously learnt. With the test scores, the teachers can assess the effectiveness of a course. For instance, if the test scores are not high, this suggests that the need to review the syllabus, materials and perhaps the teaching methods at institutional levels. Thus, having improved these concerns, the administrators and teachers should find the best ways to develop the tests which are appropriate for the students' levels and meet their specific needs.

2.2.2 Testing and Learning

Testing is a measurement device to classify the strengths and weaknesses in their learning abilities (Henning, 1987). Through testing, the learners can see at which level they are standing and what difficulties they face with. Accordingly, they can adjust their learning methods or explore more effective ways of learning. At the same time, the teachers can rely on the test results to understand better their learners' abilities and then can improve their teaching methods or update their new knowledge. In addition, testing is a self-assessment tool which helps students to identify their true learning abilities. Through testing, the students can get feedback immediately from the teachers and find out what levels they have been at, what knowledge they need to improve and what important knowledge they need to focus on. Thus, students can adjust for improving their learning styles and get more motivated to study harder to get good results.

In short, better test means better teaching; better teaching means better learning, thus it can be concluded that teaching, testing and learning are inseparable and that the agreement of these three aspects is decisive to the effective achievement of objectives of the overall educational programmes. Although the purposes of testing in education cannot be fully listed in this study, it at least can provide an illustration for various types of tasks that tests are expected to perform. These targets will provide the teachers and administrators much knowledge in test construction to make sure that their currently used tests are useful measurement tools to measure several academic purposes.

2.3 Definition of Test

Tests play important role in educational program and provide an array of benefits for both students and teachers. However, test is defined differently depending on the researchers' points of view. According to Harrison (1983) a test is the way to reflect the result of teaching and learning process in the classroom. For illustration, if a test content cannot reflect of what knowledge has been taught or what knowledge is learned, then the students' test performance will not provide a good academic success in that subject.

Bachman (1990) broadens the definition of a test as a natural extension of classroom work which provides the teachers and the students with valuable information for study performance. In other words, a test is an essential part of teaching and learning in which the teachers use the test in order to examine their students' knowledge as well as reflect their teaching methods in the classroom. Similarly, Alderson and Hughes (1981) point out that the test is an educational procedure that has been constructed to create certain behaviour from which the teachers can make implications to build individual characteristics. In other words,

the test is a useful tool which brings out an exact sample of the students' learning behaviours..

Henning (1987) highlights six purposes of language tests which are presented as follows

(1) Diagnosis and feedback help to explore strengths and weaknesses of the learners. (2) Screening and selection help to give the decision of who should be allowed to be participants in a particular program. (3) Placement helps to find out students' performance levels and to place them at an appropriate level at university. (4) Program evaluation helps to give information on the effectiveness of programs (5) Providing research criteria helps to provide a standard of judgment in a variety of other research contexts based on language test scores. (6) Evaluation of attitudes and socio-psychological differences help to establish the nature, direction, and intensity of attitudes related to language acquisition (Henning, 1987, p.1)

In the light of these research views above, it should be concluded that test is a measurement tool and is a necessary part of classroom instruction. Test content gives an accurate picture of students' knowledge and skills in the subject area or domain being tested in order to contribute in assessing what students have learned and what students can do in order to help them broaden their knowledge, achieve higher standards and improve academic performance. Through testing, the teachers do not only examine and inspire students' learning mastery, but also prepare and improve the following lessons perfectly in the teaching and testing methods. Besides different qualities of a good test, the process of preparation before test construction is an essential part which helps to create an effective and well-designed test. A carefully planned and prepared test during the test development promotes quality, gives all students an opportunity to fully demonstrate their learning abilities, and reduces the factors that will influence students' test scores. The reparation process for designing tests is one of the most

challenging tasks confronting the teachers because this is the initial step in testing which helps to determine what the teachers are attempting to measure (Anyanwu & Onwauka, 2015). This process requires a lot of steps when planning and preparing the test. For planning, the teachers are required to outline the content to be considered as the basis for the test, identify learning outcomes to be measured by the test, and choose appropriate types of test items for evaluation of learning outcomes. For preparing, the teachers write test items according to rules of test construction, select the items to be included in the test according to test specification, course objective, determine difficulty level, review and edit items, decide on method of scoring. When test preparation are planned and prepared carefully, it will create well-constructed tests which help motivate students and reinforce their learning while enabling instructors to assess students' mastery of course objectives (Young et al., 2013). Thus there is a need for tests to be well- prepared carefully as the results obtained are valuable for effective teaching and learning.

2.4 Kinds of Test

There are many kinds of test which are used to measure students' academic achievement. Hughes (1989) highlights 4 kinds of test which are commonly used in the classroom: placement test, proficiency test, diagnostic test and achievement test. Each of them has specific purposes, particular criteria to be measured.

2.4.1 Placement Test

According to Hughes (1989), placement test is a test that seems not belong to the construction of the test, but to its usage purpose to place the students at an appropriate level in a program or a course. Normally, at the beginning of the school year, placement test is

used to check the academic levels of new enrolled students, and then the intuitions can place each student in classes at their right level. It also means that this type of test is also used to decide which group or class the students could be joined and assist to put the students exactly in that group that reflect their true abilities.

2.4.2 Proficiency Test

Another type of test that should be discussed is a proficiency test. Hughes (1989) holds the view that for proficiency test, training is not the matter that needs to be focused, but the language should be. This is due to the fact that the proficiency tests are constructed to measure student's mastery in a target language without any training they may have taken in that language (Siddiek, 2010). Furthermore, proficiency test content is not built on any curriculum or syllabus of what students have to do in the language learning, but this content is created to intend to check the students' language abilities until they have satisfactory commands of the language for their specific purposes. Nowadays, there are several reliable and standardized proficiency tests which are used to check the candidates' language abilities for higher educational admission, job vacancies, such as KET, PET, FCE, IELTS, TOEFL and TOEIC. The most important function of these international proficiency tests is to indicate the true language abilities or standards that the candidates have reached.

2.4.3 Diagnostic Test

Test can also be used for diagnostics purposes. Diagnostic tests are used to locate students' strengths and weaknesses in language aspects in order to help the students to move forward.. Additionally, the diagnostic tests are very beneficial because they can be used more than once throughout the study. Yin and Sims (2006) also emphasize the diagnostic test as a useful

supplementary material for a course and can be used non-limited. For example, at the beginning of the school year, it can help teachers to plan their instructions; in the mid-term, they can determine how much students have learned so far and what they need to review and update their teaching methods; at the end of the year, the diagnostic tests can provide appropriate evaluation for the whole year learning process of a student

2.4.4 Achievement Test

The last type of test to be discussed is an achievement test. An achievement test is usually a summative test that is used to measure learners mastery of the language after going through their academic program. Brown (1994) suggests a theoretical concept that an achievement test has directly influenced to classroom lessons and curriculum. This test purposely covers the course syllabus, course objectives in order to look back over a longer period of learning in the semester or during their study. In other words, the achievement test aims to launch how successful each student, each group of students achieved learning objectives. Spratt (2005) holds a view that the achievement test is a measurement device that is designed to elicit students' study progress because it identifies not only the students' weak points, but also their strong remarks in the whole academic year. Final achievement test is usually a formal examination given at the end of the school year or at the end of the course in order to measure how far students have achieved the teaching goals. It can be strongly claimed that the final achievement test content must be related to the detailed course syllabus, textbooks and other related materials in order easily administer the teaching process in the whole school year. This could be explained that if the curriculum at school is not well-designed, or the textbooks, related materials are poorly selected, then the test results cannot be satisfactory, even the results can be very misleading. On the other hand, if the test content meets all criteria in the

syllabus and objectives, it is possible for the test performance to show how far the students have achieved the knowledge or to promote a more beneficial washback on the teaching process. Figure 2.1 provides the explanations of these kinds of test

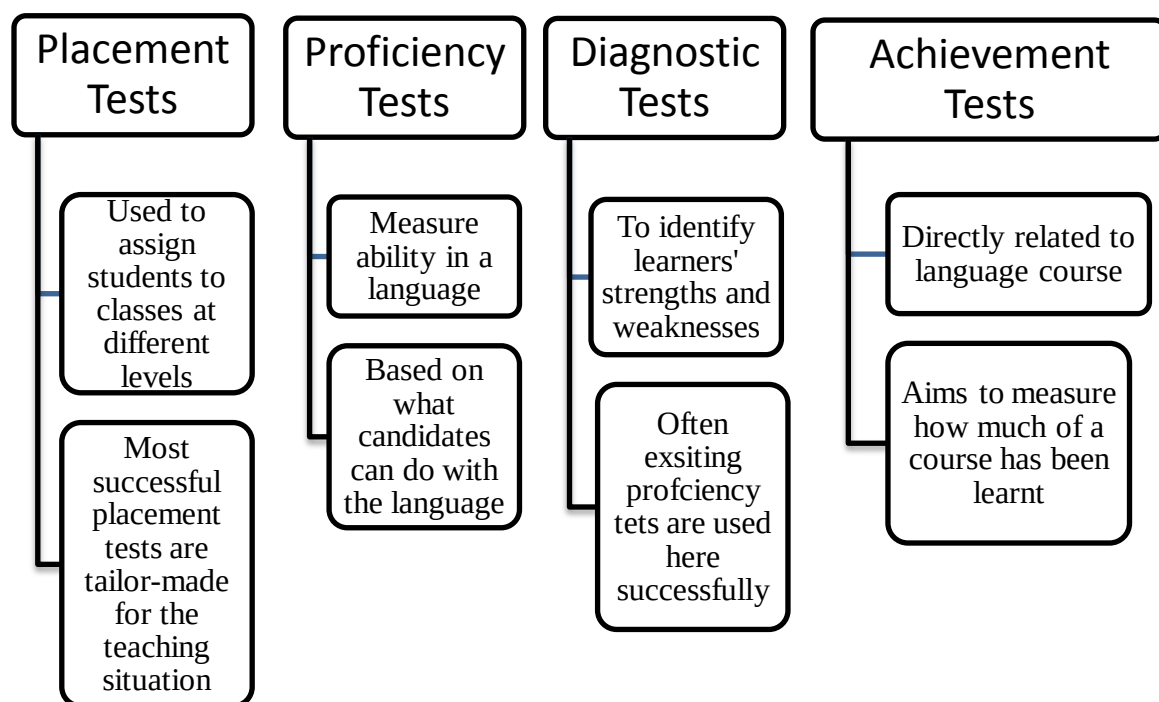


Figure 2.1: Explanations of Test Types

2.5 Qualities of a Good Language Test

In order to make a well-designed test, test writers have to take into account a variety of factors such as the purpose of the test, content of the syllabus, students' background, goals of educational program and so on. Bachman and Palmer (1996) indicate that the most important in designing and developing a language test is its usage for which it is planned to measure, thus it can be noted that the most necessary quality of a test is the test usefulness. Their idea of usefulness is the combination of reliability, validity, impact, authenticity, interactiveness and practicality. Among these ideas of usefulness, both reliability and validity are regarded as

the two most characteristics of a good test as well as two most important roles in test development and use (Bachman, 1990). The test usefulness is presented in the Figure 2.2

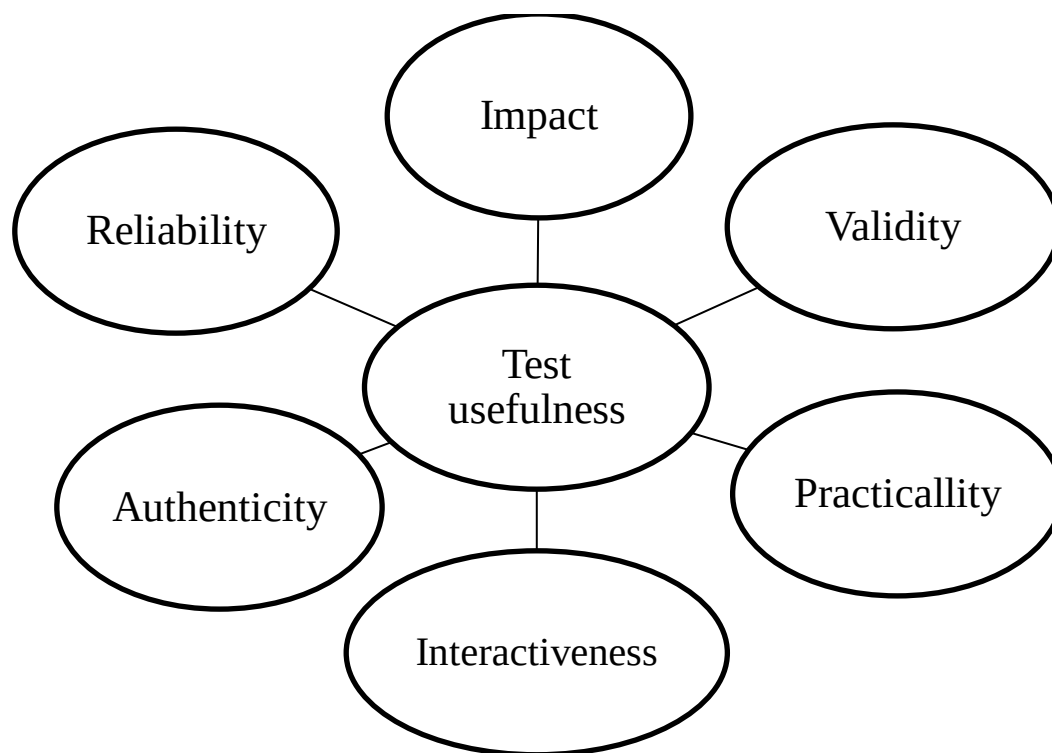


Figure 2.2: Test Usefulness

In a line with Bachman and Palmer's (1996) views, Henning (1987) defines that 'Tests, to be useful, must provide us with reliable and valid measurement of variety of purposes' (p. 87). He describes the five important characteristics of a good test which reliability, validity, practicality, discrimination and washback. In order to deeply understand what contributes a good test, these qualities are explained in the following subsections

2.5.1 Test Validity

In language testing and assessment, the ideal of test validity has been employed in for a long time. As a result, the concept of validity has been reviewed many times and has various definitions and the question whether a given test is valid or not is a central concern to all

researchers and educators. In the traditional view of test validity, some researchers defined validity is that of what a test measures what is intended to measure (Campbell & Stanley, 1966; Cronbach & Meehl, 1955; Lado, 1961). This definition seems a little too general and unclear. Therefore, when Messick's (1996) view was revolutionary because it contrasted with the traditional definitions of validity in which the research indicated that a test is considered to be either valid or not as evidenced by the correlations between the test and some other external criterion measure.

Weir (2005) holds a similar and specific definition of validity as the extent to which test scores reflect students' true level of language knowledge and skills. Sharing the same view, D'Este (2012) states that test validation involves gathering evidence and consequence of test validity outside a specific context in which the test is used, and 'test use' refers to the real use of the test in a well-defined context. In other words, test validity or test validation means evaluating theoretically and empirically the use of a test in a specific setting such as university admission, course placement and class or group classification. Therefore, when investigating the test validity, it is crucial to validate the results of the test in the environment where they are used (Henning, 1987).

Crocker and Algina (1986) highlight three kinds of test validity: Construct validity, Face validity and Criterion validity. On the other hand, Hughes (1989) categorizes test validity under some main headings which are content validity; construct validity, face validity and criterion-related validity (concurrent and predictive validity). Each validity evidence is presented separately with different researchers' points of view.

2.5.1.1 Face Validity

In the early days of language testing, face validity is used by testers and is probably considered as the most commonly discussed type of test validity because it was primarily dealt with the question of whether a test measures what it is said to measure (Hughes, 1989). In a common definition, face validity is defined as “the test’s surface credibility or public acceptability” (Henning, 1987, p.89). In other words, face validation refers to the surface of a test such as behaviours, attitudes, skills, perceptions it is supposed to measure, that test is said to have face validity. For example, if a test pretends to measure students’ speaking skills, it should measure all aspects of speaking such as vocabulary, pronunciation, intonation, word and sentence stresses, but it does not check students’ pronunciation. This means the test-takers perceive the test as a task measuring an unrelated ability, thus it can be thought that this test lacks of face validity.

In recent views on face validity, there has been a debate about this concept. On the one hand, some researchers highlight that using face validity in the study seems to be a reasonable way to gain more necessary information from a large population of people because this validity encourages a large number of people from taking part in a survey, so it can be easily get quick valuable results (Brown, 2000; Henning, 1987; Messick, 1994).

On the other hand, the value of face validity has been in controversy for a long time and has not considered as a scientific concept because this validation mainly collects data from non-experts (students, parents) who give comments on the value of the test. Davies (1984) define face validity as “the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer” (p. 59). Sharing the same view,

Stevenson (1985) criticizes the reliance on face validity in language tests, claiming that scientific evidence is likely to be ignored because a test with low face validity does not necessarily lead to invalid score interpretation and use of the test. Therefore, several experts who have been confused about the importance of face validity argue that there seems no place in the discussion of face validity in language testing. In other words, for these researchers, face validity is unscientific and not important types of test validity (Bachman & Palmer, 1981; Bekuretsion, 1997; Davies, 1978, Heaton, 1988; Sato1 & Ikeda, 2015; Stevenson,1985)

In brief, face validity is examining the appearance of test validity which helps the researchers gain more necessary information from a large population of people as well get quicker perceptions about the value of the test. However, the perceptions about the value of the test are mainly gathered from non- experts, thus the results of this validity seem to be viewed as an unreliable and unscientific in test validation. Furthermore, although the benefits of investigating face validity have been acknowledged by some researchers, there are only a few empirical studies on non-experts' judgments of the language tests (Kucuk, 2007; Liauh, 201; Sato1 & Ikeda, 2015; Wang, 2006). For these reason, the study pays much attention to other validity components: content and criterion validity to investigate the validity of the AEPET rather than face validity of the test.

2.5.1.2 Content Validity

Test quality can be affected by a number of factors. One of these factors is the sufficiency of sampling test contents from a defined syllabus, textbooks, curriculum documents and other materials. While face validity gathers relies on the judgments collected from non-experts, administrators, and students, content validity is built upon the judgments from by experts who

give comments on the test by analyzing the content, then comparing the content with the syllabus or the test specification (Alderson et al., 1995).

Hughes (1989) expresses a similar view stating that “a test is said to have content validity if its content constitutes a representative sample of the language skills” (p.11). If a test lacks content validity, it results in a harmful washback effect. For example, if content areas are less covered in the test, the students cannot present the skills that are not included in the test when the need arises; some unrelated question may be included in the test that is answered wrongly. As a result, the test will receive low attention from the students during their study. On the other hand, if all content areas during the course are represented sufficiently in the test, there will be more valuable effect on teaching. It is due to the fact that the students can give balanced attention to all content areas of the syllabus during practice and study. In other words, the more representative samples a test contains, the more valid and reliable will be the assessment of student’s knowledge and abilities. Therefore, in content validation, the experts should investigate the purposes of the test, draw up a list known as content specification as well as check whether the knowledge and skills are represented in the test, and then compare with the agreement with the content which goes into the syllabus.

In addition, Davies (1978) emphasizes that the most relevant and important type of test validity for classroom testing is the content validity which reflects the content and objectives in the syllabus that is being followed. Sharing the same view, Siddiek (2010) indicates that content validity plays very crucial factor in the success of test because it represents a sample of the language skills and structures. He also emphasizes that “a test with no content validity would spoil the whole educational process” (p.141), therefore, when selecting and designing

test items, the educators should pay much attention to a list of content standards that state the necessary knowledge which students are expected to learn during their academic year.

In brief, content validity is the most appropriate method for evaluating the validity of the test in which test items need to reflect the knowledge and skills required for a given topic area. Investigating content validity of a test will put the students and teachers on the right track by requiring themselves to the syllabus which is especially designed by experts to secure the educational objectives of the individual and community

2.5.1.3 Construct Validity

According to Alderson et al.(1995), construct validation is the most complicated type of validity to explain because it demonstrates a relationship between the scores of a test and the prediction of theoretical features. Similarly, Ebel and Frisble (1991) also give the definition of the construct validity as follows:

The term construct refers to a psychological construct, a theoretical conceptualization about an aspect of human behaviour that cannot be measured or observed directly. Examples of the construct are intelligence, behaviours, achievement, motivation, anxiety, attitude, ability, dominance and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores or score interpretation from the test to assure that the scores mean what we expect to mean (p.108)

In order to justify a particular scores interpretation, there is a need to provide evidence that the

test scores reflect the areas of language ability which are measured. Hence, Bachman and Palmer (1996) states that the construct validity is “the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task” (p. 21). In addition, Messick (1989) highlights that construct validity is considered as an ‘unified though faceted concept of which content and criterion-related evidence are contributory factors” (p.14). It means that construct validity is regarded as an unitary concept in which there are many way of accumulating evidence to support this validity. In the same token, Loevinger (1957) indicates that the content, predictive and concurrent validity are the possible supporting evidence for construct validity, which subsumed them and much more” (p. 471). In other words, to prove the construct validity of a test, it requires using a series of validity evidence such as content, concurrent and predictive validity.

In the light of literature review given above, it can be clearly noted that how complicated and difficult the construct validity is and its exact meaning still remains incomprehensible by many test experts, as a result, construct validation is less useful for classroom-based or in-house test evaluation (Alderson et al., 1995). However, another way which contribures evidence for construct validity is to gather series of validity evidence such as content relevance and coverage as well as criterion-relatedness because such information about the content domain of reference and about specific criterion behaviours predicted by the test scores clearly contributes to scores interpretation

2.5.1.4 Concurrent Validity

Alderson et al. (1995) defines that concurrtent validity is the degree to which a test correlates with other standarlized tests to measure the same thing. In other words, a test is considered to

have concurrent validity is its test scores give similar results in compared with other tests which are valid and have the same purpose.

According to Bachman (1990), when evaluating concurrent validity, there are many concerns. First, the measurement that is being used for comparison of the test in question must be valid. If the measurement is not valid, there is no point in testing another test's validity against it. Thus, it is suggested that when conducting the concurrent validity of a language test, it is better to compare this language test with one of other standardized language tests such as IELTS, TOEFL, KET, PET and TOEIC. Furthermore, the purposes and components of the two tests are the same. For example, the concurrent validity of a speaking test scores can not be used to compare with a grammar test scores

To sum up, while the different kinds of test validation mostly focus on the test content, concurrent validation focuses on test scores, and the test is only valid when there is a high correlation between test scores with other standardized test scores which are supposed to measure the same things as well as manage at the same time.

2.5.1.5 Predictive Validity

Predictive validity is one type of criterion validity and has a close relationship with concurrent validity. However, according to Alderson et al. (1995), predictive validity is said to be slightly different from concurrent validity due to the fact that concurrent validity "measures at the same time as the administration of the experimental test, the predictive validity will only be gathered some time after the test has been given" (p.180). In other words, predictive validity refers to how accurately a test score predicts performance on the criterion measure at some

future point in time. When conducting the predictive validity of a test, it always raises a question of how well the English proficiency test scores can accurately predict students' academic success at university. Carol (1968) and Harley (2001) state that there are a number of problems in trying to answer such question because there are many factors which can predict academic success rather than English proficiency. However, the findings in most previous studies which examined the correlation between English proficiency and performance indicate that English language proficiency scores are significantly predictive of academic achievement (Maleki & Zangani, 2007; Sivaraman et al., 2014; Sadeghi et al., 2013; Roche & Harrington, 2013). Therefore, it suggests that a high correlation between language proficiency scores and academic achievement would indicate a high degree of predictive validity (Alderson et al., 1995).

On balance, as language testing has been gone up and down through various stages of development in the past century, test validation and types of validity have also undergone changes. Therefore, a special attention must be paid to the test validity when one constructs it because the most important consideration in test evaluation is validity which refers to the appropriateness, meaningfulness, and usefulness of a given test as a measure of what it is supposed to measure. Table 2.1 shows the summary of the main aspects of validity in language testing context.

Table 2.1: Common Aspects of Validity in Language Testing Context

Types of evidence	What is it?	Why important?	How analyze?
Face validity	Judgments about the test content by people who judgment are not necessarily experts	Students' performance is much better if they are aware of the face validity of the test (Alderson et al.,1995)	Interview or questionnaire to students or teachers to ask about the test content after taking or administering the test
Content validity	Judgements from experts who give comments on the test by analyzing the content, then comparing the content with the syllabus or the test specification	Knowing better the content of test items helps test developers to measure course syllabus, course objectives and students' ability (Hughes, 1989)	Expert judgements compare test items with test specifications, course syllabus
Construct validity	Comparisons of test scores of the same students with the prediction of a theoretical trait: behaviours, intelligence	It is very important for psychological test (Ebel & Frisble,1991)	Find the relationship between students' test scores with psychological traits: intelligence, attitudes, behaviors...
Concurrent validity	Comparisons of test scores of the same students with other measures taken at the same time	To check students' ability intensively when comparing with a standardized test (Bachman, 1990)	Find the relationship between students' test scores and a standardized international language test : TOEFL, IELTS,
Predictive validity	Investigating how well scores in the test taken predict future performance	To determine the prediction of academic achievement and utility of a test (Alderson et al.,1995)	Correlate same students' test scores at the beginning or the end of course

Of the five types of test validity involved in the language testing context, face validity seems to be viewed as an unreliable and unscientific in test validation because the perceptions about the value of the test are mainly gathered from non- experts. Construct validity is considered as the most complicated type of validity to explain because it demonstrates a relationship between the scores of a test and the prediction of a theoretical features such intelligence, behaviours, achievement, motivation, anxiety, attitude, ability and so on. For these reasons,

these two components are less practical for test evaluation. On the other hand, several researchers suggest three more important types of validity evidence which should be paid more attention when investigating the validity of a test: content relevance and coverage, concurrent and predictive validity (Bachman, 1990; D'Este, 2012; Wang, 1998; Wier, 2005). Sharing the same view, Yen and Kuma (2009) indicates that the crucial validity considerations in language performance assessment are criterion-related validity (concurrent and predictive validity) and content relevance and coverage. Since the major purpose of tests is to provide predictive information about how well the students will perform their language proficiency under specific target conditions, predictive validity has been one of the primary concerns in performance assessment.

Taking these aspects into consideration, this study intends to give more emphasis to investigate the content, concurrent and predictive validity of the AEPET at a public university in Vietnam in order to see to what extent the representativeness of language knowledge and skills related to the course content is covered in the test as well as to find out how significant predictors of students' performance on the test in the context of in-house language test in comparison with an established standardized international test.

2.5.2 Test Reliability

Bachman (1990) describes test reliability as the measurement tool to measure the quality of test scores. He illustrates that if a student receives a low score on a test one day and a high score on the same test two days later, the test does not yield consistent results, and the score can not be regarded as a reliable indicator of that student's true abilities. Therefore, it is said that this test is not reliable. Sharing the same view, Surface, Poncheri, and Bhavsar (2008)

state that the test reliability is an essential quality of a good test which can produce the same results if it is administered to the same students at the different point of time.

Similarly, test reliability is defined by Bernstein, Moere and Cheng (2010) as “a measurement device and its consistency is the dependability and trustworthiness of that device” (p.110). For example, if a ruler is used to measure a piece of paper, it gets the same result every time. Therefore, this ruler is a reliable measure and the reliable test needs to get a truthful measurement like the ruler. In contrast, a multiple-choice test would possibly yield different scores from one administration to another and would thus be absolutely considered as unreliable. Another common view from Yu (2005) has been defined that reliability means the stability of the test scores. A test can not measure anything well unless it measures the consistency in scores because score reliability refers to the stability of the results of test performance. There are a number of ways to collect reliability data for a test such as test-retest reliability, alternate forms, split half reliability, inter-rater reliability, and internal consistency (Campbell & Stanley, 1963; D'Este, 2012; Wells & Wollack, 2003; Yu, 2005).

To sum up, reliability is one of the important characteristics of a good test. If the test result is not reliable, the assessment of this test is not reliable, either. In order to prove the test to be reliable, it is important for testers to consider influential factors such as test administration, timing, testing conditions, observation or control of doing the test, the size of the test, test instructions, and scoring methods right from the outset of the test construction process.

2.5.3 Practicality

In addition to reliability and validity, practicality plays a crucial role in deciding whether a test

is good or not. This characteristics of a test includes considering financial restrictions, time constrains, administration, scoring and interpretation. According to Brown (2004), “a test is impractical if it is taken in 10 hours and if it is prohibitively expensive” (p.253). This view is also shared by Harrison (1983) who emphasized that the validity and the reliability of a test might not be use frequently if this test does not show the to be a practical one to the administration. These definitions refer to the economical terms including administrating, scoring, reporting, storing and interpreting scores. A test should be as economical as possible in time (preparation, sittings and marking), and in cost (hidden cost of time spent and materials). In short, to make sure to be useful and efficient, tests should be as economical as possible in terms of time and costs on the one hand and should include well-written instructions on the other hand.

2.5.4 Washback

It would be incomplete in constructing a test without adding an essential quality named washback. Messick (1996) defines washback as one of the main concepts in applied linguistics because it refers to the degree that the test development and use has been influenced on language teachers and learners.

In a different way, washback is normally described as the feedback of testing on teaching and learning process (Alderson et al., 1995). Similarly, the term washback or backwash has been used in Hughes’ (1989) article as the way a test affects teaching materials and classroom management. There are two kinds of washback: negative (harmful) and positive (beneficial). Negative washback occurs when there is a mismatch between the stated objectives of instruction and the focus of assessment (Messick, 1996). For instance, if the skill of writing

is tested only by multiple choice items then there is great pressure to practice such items rather than to practice the skill of writing itself. Positive washback only occurs when there is a match between what has been taught and what has been tested (Bailey,1999; Messick, 1996). For example, if the teachers teach speaking skill and the test content relates to what they have taught, they will get a good feedback about their teaching process.

In brief, washback plays as a powerful tool for all educational institutions with the objective of controlling the quality of what teachers teach and what learners learn at school because only through washback, the educators better understand the relationship between teaching and learning for specific tests.

2.5.5 Discrimination

Apart from validity, reliability, practicality and washback, discrimination contributes a quite important role in designing qualities of a test. Heaton (1988) identifies discrimination of a test as “the capacity to discriminate the different candidates and to reflect the difference in the performance of the individual in the group”(p. 165). It means that if a test is either too easy or too difficult, it can not realize its purpose of discrimination between candidates. Therefore, the test items should be in a wide difficulty scale, ranging from easy to difficult items. In the same vein, (Harrison, 1983) defines discrimination as the degree that a test clarifies the students from each other .

However, the need to discriminate is not the same for every test. In fact, it varies due to the objectives of the test. For example, it will be much easier to divide students into suitable groups and similarly to an achievement or a diagnostic test, the level of each individual will

clearly be shown. All in all, designing a good test is considered as a complex art, so in order to avoid making mistakes in the process of designing, the educators should understand what makes a good test, the characteristics of a good test.

2.6 Relationship between Validity and Reliability

Test researchers and designers have admitted that validity and reliability are the two most crucial measurement qualities for a good test because these qualities provide major justification for using test scores, numbers as a basis for making inferences or decisions (Bachman & Palmer, 1996). In contrast, the relationship between validity and reliability is quite complicated and requires the test designers to take long time when constructing an ideal test.

On the one hand, a reliable test may not be necessarily valid. The test which gives the same consistent results at different times, but that does not mean that the test always measures what it is supposed to do (Johnson & Svingby 2007). For example, if the strength of grip is very reliably at different time, but that does not make it a valid measurement of intelligence or someone's ability. In this case, reliability is found, but not sufficient. On the other hand, a valid test is always reliable. In other words, reliability is a part of validity because the test cannot be considered as valid unless its measurement is reliable (Bachman, 1990; Hughes, 1989; Messick, 1996). Consequently, it can be said that both reliability and validity always tend to co-exist in the valid test.

Harrison (1983) claims that there is some tension between reliability and validity in which the test writers have to confront a difficulty in making a choice whether to attempt to increase

the validity of a test known to be reliable or else or to increase the reliability of a test known to be valid. The test writers will be faced with an impossible task because the features always make the test valid. As a result, it is essential to device a valid test first of all and then to establish ways of increasing its reliability (Heaton, 1988). The relationship between reliability and validity is presented in Figure 2.3 taken from social research method in 2006



Figure 2.3:The relationship between reliability and validity

(Source: Social Research Method, 2006)

In brief, it should be noted that a test could be reliable without possessing validity. However, reliability is clearly inadequate by itself if a test does not succeed in measuring what it is intended to measure. For this reason, the test writers should balance and constitute the two chief criteria, the reliability and validity, for designing idea tests.

2.7 Theoretical Framework

Test validation methods are at the heart of language testing research and validity has long been acknowledged as the most crucial aspect of language testing and has different types such as face, content, construct, concurrent, and predictive validity (Bachman, 1990; Hughes,

1989). Hence, the way to determine the scope and nature of validity investigations is crucial. Furthermore, test stakeholders (test takers, educators) and other test score users (university administrators, policy makers) always expect to be provided with the evidence of how test writers can determine and control criteria distinctions between proficiency tests applied with different levels. Therefore, there is a need among these stakeholders of the value of having not only a clear theoretical framework to support for the test but also a means of generating explicit evidence on how that model is used and taken in practice. A theoretical framework can be thought of as a useful map at the start of any research study because the framework helps to specify the procedure used to prioritize, integrate, and evaluate evidence collected using various methods (Xi, 2008). In the field of language testing, test validation framework helps collect validity evidence in a systematic way and expose threats to validity and propose ways to reduce or remove these threats” (Haladyna, 2006, p. 739) Therefore, it is likely that without a validation framework, the full range of validity questions and threats will not be identified (Ferrara, 2007).

In the light of the above literature review, several test validation frameworks have been developed in order to enable researchers to provide comprehensive supporting test validity. For example, Messick (1989) presents a unified model, and shows construct validity as a central underlying component, with content and criterion validity as aspects of construct validity. The framework focuses on two main interconnected facets in order to highlight the inferences and decisions made from test scores. One facet is “the source of justification of the testing, being based on appraisal of either evidence or consequence”, whilst the other facet is represented “by function or the outcome of the testing, being either interpretation or use” (Messick, 1989, p. 20). However, Messick’s framework has been criticised for shifting

validity to test score inferences, mainly because it seems natural and instinctive to consider validity to be a feature of a test (Van de Walt & Steyn, 2008). Furthermore, his framework is considered to be influential in testing; however, it is complex, difficult and challenging to operationalise (McNamara & Roever 2006; Van der Walt & Steyn, 2008).

In recent validation framework, Kane (2006)'s proposes a validation framework which mainly focuses on the interpretations and uses of test scores. While this framework is more manageable model for a placement test, it was criticised for removing the test itself from being the focus of validation and instead placed the focus on the score interpretation and use (Messick, 1995). This approach was also criticized by McNamara (2006) for “not paying adequate attention to issues of the policy contexts of tests and their social consequences or providing ways to study them” (cited in Tran, Griffina & Nguyen, 2010, p. 1298). Similarly, Shaw and Weir (2007) provide a framework which outlines six factors that need to be considered in test validation: test-taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity. However, Shaw and Weir's framework suggests within the literature on writing skill in which important questions that anyone intending to take a particular test or to use scores from that test would be advised to ask of the test developers in order to be confident that the nature and quality of the test matches up to their requirements. These questions represented a comprehensive approach to validate a writing test, however were not the source of all the evidence to be collected on other English skills: Listening, Reading and Speaking in order to improve the whole picture of validity of the test.

The socio-cognitive framework for developing and validating English language tests of

Listening, Reading, Writing and Speaking in Weir's (2005) model of conceptualizing test validity seem to meet all the demands of the validity in the test that test stakeholders want to use in the public domain. O'Sullivan and Weir (2011) describe the socio framework as "the first systematic attempt to incorporate the social, cognitive and evaluating (scoring) dimensions of language use into test development and validation" (p. 20). The framework has been widely used across a range of test development and validation in language testing and assessment. For example, one of the most important and reliable international test providers that has applied the framework to its examinations is University of Cambridge English for Speakers of Other Languages (ESOL) Examinations, in the United Kingdom. Cambridge ESOL has used this framework to conduct a comprehensive evaluation of Cambridge ESOL's current approach in order to examine the skill of English language according to a number of dimensions: test taker's cognitive abilities, the context in which the task is performed, and the scoring process. As a result, Cambridge English ESOL has successfully sought out validity evidence for its general English tests (e.g. KET, PET FCE, and CPE), tests of financial and business English, the International Legal English Certificate, and The Teaching Knowledge Test (O'Sullivan, 2006).

Additionally, several universities also have applied Weir's (2005) framework for variety of purposes in the assessment of both English language and other languages in different several universities in different countries such as UK, Malaysia, Taiwan, Saudi Arabia, and Japan. Sharing the same view, O'Sullivan (2009) emphasizes that the most significant contribution to the practical application of validity theory in recent years has been Weir's (2005) socio-cognitive frameworks which have had influenced on test development and validation. Similarly, Abidin (2006) points out that Weir's (2005) framework combines all the important

elements expected of a test that measures a particular construct in valid terms. Figure 2.4 presents an outline of the socio – cognitive framework for validating language tests

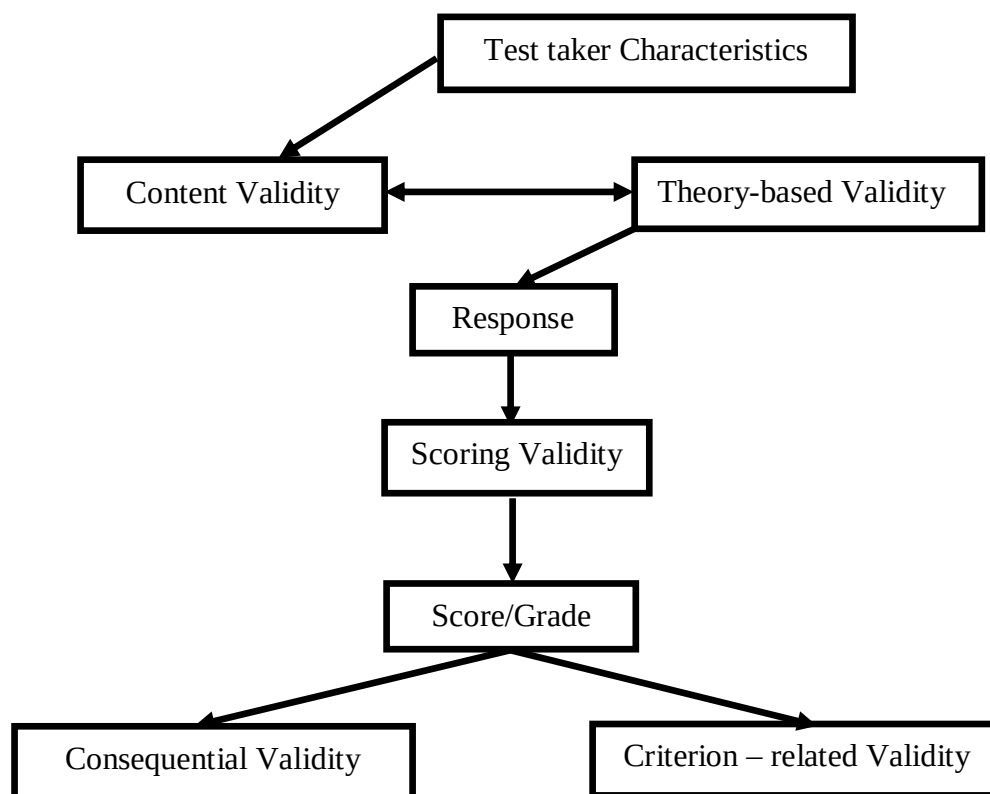


Figure 2.4: Outline of Weir's (2005) Socio-Cognitive Test Validation Framework

More specifically, Weir (2005) proposed four frameworks to validate four English language skills: Listening, Reading, Writing and Speaking. In each framework, Weir (2005) put emphasis on validating context validity, theory-based validity, scoring validity, criterion-validity (concurrent and predictive validity) and consequential validity (washback). He also provides essentially with complete explanation of each process model. These models in the four domains of language skills have been in an effort to help the test-designers reflect on their own test construction. In general, the framework consists of a number of components, each of which must be attended to by the test writers at one or more points of the test development, implementation and validation cycle.

2.7.1 Test-taker Characteristics

At the first stage of design and development of the test, the research is concerned with the components in Test-taker characteristics, which represent for candidates in the test event, always focus on the individual language user and their mental processing abilities since the candidate directly impacts on the way he/she processes the test task. In other words, in this stage, the important characteristics which are related to the test-takers may have potential effect on test, thus the test-developers should consider the test-takers as the central to the validation process first. Table 2.2 presents in detail the view of test taker characteristics under the headings: Physical/ Physiological; Psychological; Experiential

Table 2.2: Test-taker characteristics framework suggested by (Weir, 2005)

Physical/ Physiological	Psychological	Experiential
- Short-term ailments: Toothache, cold...	Personality	- Education
	Cognitive style	- Examination experience
	Concentration	- Communication experience
-Long term illnesses: hearing age, sex, vision...	Motivation	- Target language country
	Emotional state	residence

In this study, test takers are homogenous as they went through the same educational system, sat for the same English entry test before being accepted into the AEP. Bachman (1990) and Bailey and Damerow (2012) state that a high validity of a test can be achieved in population of the test takers with very similar background and experience. In other words, the high level of homogeneity of the test taker population, results of the predictive validity of a test is more reliable than the results with more diverse population because the more homogenous the population of the test takers is, the students' success is similar (Bailey & Damerow, 2012)

2.7.2 Theory-based Validity

Theory-based validity was renamed Cognitive validity (Khalifa & Weir, 2009) focuses on the

processes that test-takers use in responding to test items and tasks. This validity requires test-takers to find out if the internal mental processes that a test elicits from a candidate resemble the processes that he/she would employ in non-test conditions. Furthermore, theory-based validity includes executive resources and executive process. Executive resources consist of linguistic knowledge and content knowledge of the test-taker. The test-taker can use grammatical, discoursal, functional and sociolinguistic knowledge of the language in the test. These resources are also equivalent to Bachman's views of language components.

Weir (2005) defines language ability as comprising of two components: language knowledge and strategic competence that will provide language users with the ability to complete the tasks in the test. He also emphasizes that there is a popular method to explore the theory-based validity in which this validity can be checked through investigating test-takers' behaviours by using various types of verbal reporting (e.g., introspective, immediate retrospective, and delayed retrospective) in order to stimulate their comments on what they often do in Listening, Reading, Writing and Speaking tests (Huang, 2013; Shaw & Weir, 2007). However, this method mainly relies on subjective judgments from test-takers or it refers to the cognitive processes that the researcher cannot directly observe, thus cognitive validity seems to be unscientific (Shaw & Weir, 2007) and the understanding and use of cognitive validity to measure task demands seem to rarely use in teachers' toolbox as well as less useful for test evaluation (Huang, 2013). Bearing this in mind, cognitive validity in Weir's framework was beyond the scope of this study.

2.7.3 Content Validity

Siddick (2010) states that a test which cannot prove to be valid in content may spoil the whole

educational process because before looking at the validation process in detail, it should be noted that the practice of creating a set of detailed and exemplified test specifications is the first step in establishing evidence of the validity of the test. Context validity is concerned with content coverage, relevance and representativeness of language skills in the test. Context validity for four English components: Listening, Reading, Writing and Speaking proposed by Weir (2005) are divided into: task setting, task demands and setting administration. The task setting consists of rubrics, test purpose, response format, known criteria, weightage, order of items and time constraints. Task demands is classified into discourse mode, channel of communication, text length, nature of information, content knowledge required and other linguistic factors related to task input and output. The test administration includes physical conditions, uniformity of administration and test security. Therefore, the study focuses on investigating the content validity of the AEPET by looking at two aspects: content validity of AEPET components (Listening, Reading, Writing and Speaking) and content validity of AEPET preparation in order to find out a comprehensive picture of before, during and after the test development.

2.7.4 Scoring Validity

Weir (2005) defined scoring validity is concerned with all test aspects impacts scores' reliability. In other words, within Weir's (2005) model, reliability is subsumed under scoring validity which refers to the way how the task performance is calculated, and the test score must fit with the conceptualization of the ability being tested. This validity performs on four factors: item analysis which relates to the easiness or difficulty of the test items; internal consistency which is used in homogeneous tests in which statistics can be proved how reliable the test is; error of measurement that shows the differences between observed scores and

obtained true scores; marker reliability which has influences overall test reliability. As mentioned earlier, reliability is a part of validity because the test cannot be considered as valid unless its measurement is reliable. In other words, both reliability and validity always tend to co-exist in the valid test (Bachman, 1990; Hughes, 1989; Messick, 1996). Bearing this in my, scoring validity which concerns with all test aspects influences scores' reliability was beyond the scope of this study.

2.7.5 Consequential Validity

Consequential validity in the Weir's (2005) model is considered as a additional validation and is known as washback in the classroom. Some researchers describe that this validity as the impacts of a language test on different factors such as teaching and learning in the classroom, groups of test-takers, test takers' background (Bachman, 1990; Siddick, 2010; Shaw & Weir, 2007)). However, several researchers define consequential validity as a complex and ill-defined phenomenon in language testing. They believe that washback is a more complicated validity in which the impact of a test should be validated with the reference to the goals, values, development of the context of society and educational policy rather than simply evaluated the effect of a test on teaching and learning (Bachman & Palmer, 1996; Bailey, 1999; Iyer, 2015; Messick, 1996). Due to the ill-defined scope in language testing, few studies have been conducted to investigate the consequential validity of a language test , accordingly consequential validity was beyond the scope of this study.

2.7.6 Criterion- related Validity

Criterion-related validity refers to the relationship between test scores obtained from the same or different versions of a test administered to the same students in the same conditions

(Bachman, 1990; Hughes, 1989; Messick, 1996; Weir, 2005). There are two kinds of criterion validity: concurrent and predictive validity. As noted above, concurrent validity is the degree to which a test correlates with other valid tests testing the same thing. Predictive validity is the extent to which the test in question can be used to make predictions about future performance (Alderson et al., 1995).

2.7.7 Validity Components for Validating the AEPET

Based on the review of literature above, it is noted that Weir's framework represents unified approaches to gathering evidence for an in-house language test due to their relevance and value to test validation context. Furthermore, adopting Weir's validity frameworks guarantees more valid and reliable tests with more systematic decisions (Abidin & Bannur, 2015). Thus, the study adopts Weir's (2005) socio-cognitive framework as a guideline for validating four English sub-skills tests: Listening, Reading, Writing and Speaking. The study looks at the application of the framework for the evidence on criterion-related validity (concurrent and predictive validity) and content validity.

For criterion-related validity, validity evidence is collected based on students' AEPET, IELTS and their CGPA scores in order to find out the relationship between students' AEPET score and a standardized international English language test score (IELTS) to investigate the concurrent validity of the AEPET; look at the relationship between the students' AEPET score and academic achievement (CGPA), in comparison with the relationship between IELTS score and CGPA to examine the validity of the AEPET.

For content validity, the AEPET was investigated into two parts: content validity of AEPET

Components (Listening, Reading, Writing and Speaking) and content validity of AEPET's Preparation. The questionnaire was distributed to English language lecturers to obtain their judgments about the AEPET. While AEPET Components put an effort to cover elements in Weir's (2005) framework on weightage, time allocation, representation of skills and topics, clarity of questions and instruction, and mark allocation, AEPET's Preparation put emphasize on items which are related to test specification, course syllabus, sample AEPET papers, preparation for questions, answer keys and marking scheme, vetting process and rater training section. Figure 2.5 presents validity components for validating the AEPET adopted from Weir's (2005) Framework.

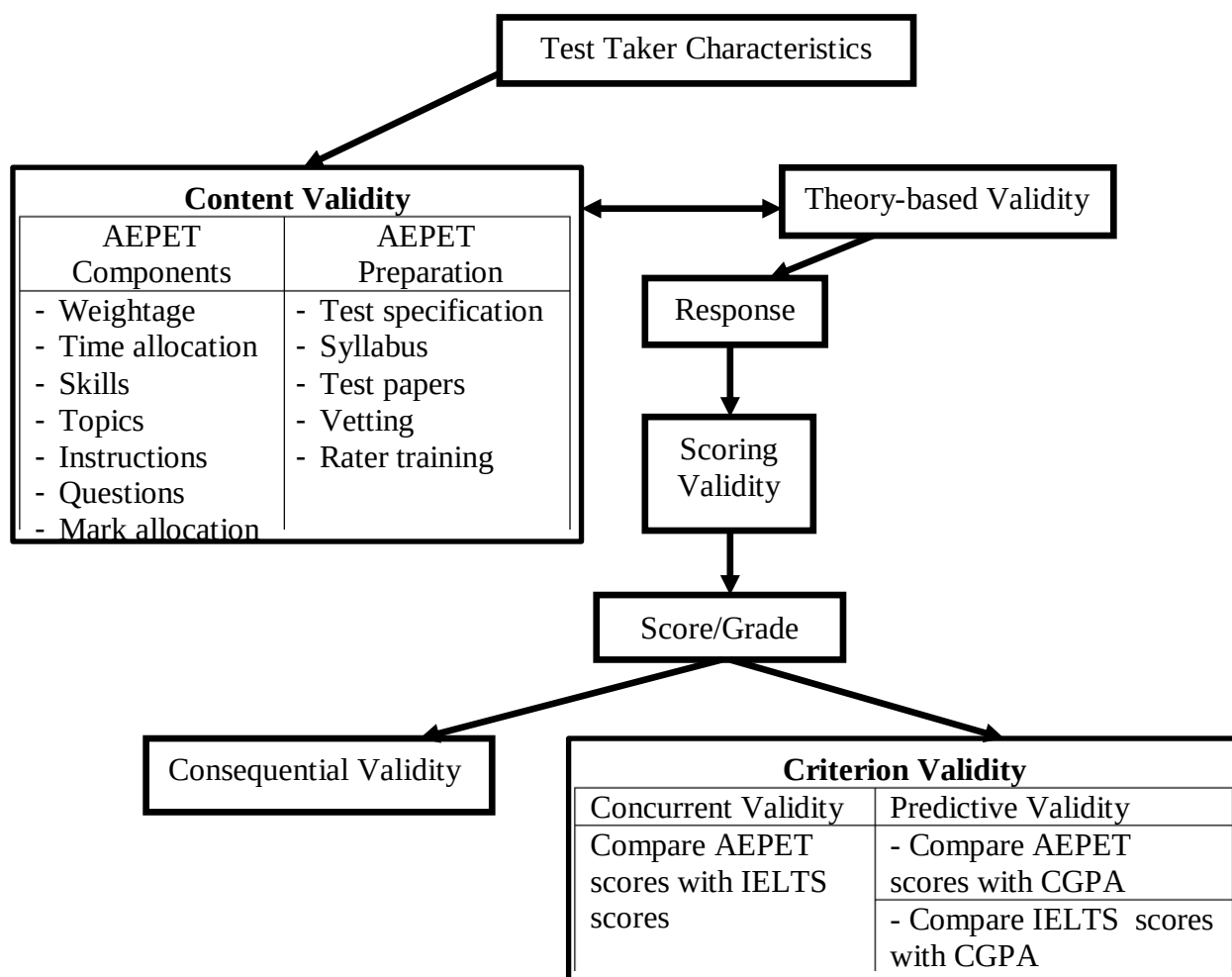


Figure 2.5: Validation components for the AEPET

In short, the socio-cognitive framework suggested by Weir (2005) is unitary concepts of validity that contains key validity elements. These elements content validity, scoring validity, consequential validity and criterion validity not only reflect the practical and nature quality of a current existing test, but also provide some possible methods for test validation. Consequently, Weir's (2005) socio-cognitive frameworks have been accepted and widely used in the public domain by the United Kingdom examination board as well as by other countries in the world. Bearing in mind, a adopted model for validating the AEPET at a public university in Vietnam has applied in respect to support validity evidence for the quality and usefulness of the university English language test. The study focuses criterion-related validity (concurrent and predictive validity). The AEPET which is modelled after IELTS and assesses students' four English skills: Listening, Reading, Writing and Speaking, thus there is a need to compare it against IELTS and CGPA in order to see to what extent the AEPET predicts students' academic achievement. It is necessary to state why it is important to look at criterion-related validity because predicting students' academic achievement is a crucial step in order to improve the efficiency of educational program, and a valid test must predict students' academic success (Elder, 1993; Grilli & Rampichini, 2013; Rethinasamy & Chuah, 2011; Rumsey, 2013). Wilson (2016) state that investigating criterion-related validity is to prevent “wastage” in educational systems and to prevent students being “set up” for failure (p. 13). More importantly, there is a need to determine whether the AEPET is fit for the main purpose, which is to provide a picture of a student's English proficiency in order to find out whether students are linguistically ready for the demands of tertiary education in which English is the medium of English. Content validity is an addition to the focus, as it would provide valuable information on the criterion validity, thus, content validity was included in the study.

2.8 Previous Studies on Test Validation

Many studies in language testing have already been conducted in an attempt to analyze the different aspects of test validation. McNamara (2000) points out that insights from such analysis provide invaluable contribution to defining the validity of language tests. Exploring how other researchers have investigated the validity of a language test can shed light on the process followed in this research study

.

2.8.1 Studies on Concurrent Validity

Muñoz et al. (2003) validated an in-house English oral test for classroom use in Spain. The purpose of this research study was to examine the concurrent validity of the Language Centre English oral test (LCEOT) with the Key English Test (KET) and Preliminary English Test (PET) administered by Cambridge Examinations. A total of 32 students from beginning and intermediate levels participated in the study. The analyses were conducted by correlating the two sets of measures through Pearson Product Moment Correlation Coefficient and Simple Linear Regression Analysis. The results showed that low range of correlations between KET and LCEOT (from $r = .228$ to $r = .255$, $p = 0.079$); strong range of correlations between PET and LCEOT (from $r = .62$ to $r = .83$) were found. The results showed that there was only concurrent validity evidence between LCEOT and PET. The researcher explained the mismatch between LCEOT and KET were that both KET and LCEOT measure similar characteristics of oral language ability but KET tended to provide students' overall performance for beginners while LCEOT just focused on the way of testing students' knowledge that they have learnt in the classroom. From the findings, the researcher suggested that the test needed to be adjust from lower levels and the training toward consistency in teacher evaluation should be improved better to design suitable test for students.

Riazi (2013) examined the concurrent and predictive validity of Pearson Test of English Academic (PTEA) which is an in-house language test at Department of Linguistics, Macquarie University, in Australia. The study involved 60 international university students who were non-native English speakers. These students had already taken IELTS Academic as a requirement for their entry into university. The researcher investigated the criterion related validity by looking at the correlation between students' scores on a criterion test (IELTS Academic) and their PTEA scores; the correlation between students' PTEA scores and their GPA. The results showed that there were strong correlations between students' overall PTEA and their overall IELTS Academic scores ($r = 0.824$), and the four communication skills of English: The results show that correlations for Speaking and Writing skills were higher than the correlations for Reading and Listening skills. This means that the Speaking and Writing tasks in the two tests had more common variance compared to the Reading and Listening tasks. The findings in this study provide useful implications for the testing community and higher education decision-makers.

2.8.2 Studies on Predictive Validity

In addition to these studies, there have been a lot of previous studies in terms of the predictive validity of in- house language tests investigated by several researchers.

Cope (2011) investigated the predictive validity of the in- house language test in three programmes: Direct Entry English Program (DEEP), Business Preparation Program (BPP) and Accounting Preparation Program (APP), which was developed on site for Non English Speaking Background international students to provide language preparation and assessment for university admissions purposes in Australia. All 138 students who had completed degree

studies via three programs: DEEP (33) , BBP (68 students) and APP (37 students) were participants in the study. Pearson correlation and Linear Regression were computed to find the relationship between the students' scores of three programs and CGPA. The results showed that the language tests for DEEP ($r=.361$, $p<0.05$) and APP ($r=.418$, $p<0.05$) students significantly correlated with their academic outcomes. The strength of correlations ranged from weak to moderate. However, there was no significant relationship between the language test for BBP students and their academic achievement ($r=.156$, $p>0.05$). The findings indicated that both language tests in DEEP and APP satisfactorily predicted academic outcomes while the predictive validity of the language test for BBP was not demonstrated.

Another study was conducted by Dogru (2013) on investigating how well various in-house assessment practices (placement test, midterms, quizzes, and readers) of the preparatory year English program in the Department of Foreign Languages predict the success of students at University of Economic and Technology (TOBB ETU). The researcher used the scores of 337 students to find out the relationship between in-house assessment practices and TOEFL ITP. All data was analysed through Pearson Product Moment Correlation and Simple Linear Regression. The predictive validity results showed that there was a significant correlation between midterms and TOEFL ITP ($r = .49$, $n= 150$, $p > .05$). For quizzes, the results indicated that the relationship between Quiz-2 and TOEFL ITP was low ($r=.30$, $n= 150$, $p> .05$); the correlation between Quiz-2 and TOEFL ITP was significant ($r=.49$, $n=150$, $p> 0.5$). For the placement test, there was a weak correlation between the placement test and TOEFL ITP ($r = .27$ $n = 150$, $p > .05$). Finally, a positive correlation between the average of reader exams and TOEFL ITP was found ($r = .44$ $n = 101$, $p > .05$). The findings show that the assessment

practices used in the Department of Foreign Languages at TOBB ETU can predict students' success. As a result, students, who get higher marks from these assessment practices, may most possibly get higher marks from the TOEFL ITP or vice versa

Fakeye (2014) examined the extent to which the English Language Proficiency Test (ELPT), which is an in-house English test for Nigerian secondary school students, predicted their overall academic achievement in Oyo and Osun States of Nigeria. A total of 400 students from eight randomly sampled secondary schools in the two states were participants in the study. The students' ELPT score was correlated with their academic achievement (CGPA) at schools by using Pearson Product Moment Correlation and Linear Regression Analysis at 0.05 level. The results showed that the ELPT has a significant positive relationship with their overall academic performance and a significant impact of English language proficiency was found on students overall academic achievement. The findings recommended that the educators should pay more attention towards making students in Nigerian more proficient in English as a way of enhancing their academic achievement

Gheghest (2014) investigated the predictive validity of in-house English Language Proficiency Test (ELPT) Preparatory Year students in three faculties: Engineering, Business (including Business Administration, Economics and Political Science) at a British University in Egypt. A total of 566 students scores were obtained from the university database and then were analysed through Pearson Correlation and Simple Linear Regression. The results found a significant but moderate positive correlation between the students' ELPT and their overall academic achievement ($r=.349$, $p=.000$). In other words, the ELPT can be a significant predictive for students' academic success. The finding indicated that the higher the students'

ELPT scores on entry to the university, the better they performed in their degree program and in their English abilities.

A similar study was conducted by Mohammadi (2009) in order to investigate the validity of Islamic Azad University's Entrance Examination (IAUEE) in Iran administered by applied English Language courses. The study focused on the predictive validity to investigate the correlation between the IAUEE and students' academic achievement regarding their Grade Point Average (GPA). The study involved 282 randomly selected senior undergraduate students who admitted to study in all language fields in Islamic Azad University. The results indicated that there was a high correlation between students' total scores in IAUEE and English proficiency subtests ($r=.74$). The total students' scores in admission exam showed a moderately significant correlation with students' total GPA $r=.43$. The findings indicated that the predictive value is much stronger among ELT students and the educators had better use IAUEE total score regarding any further planning for students

Osken (1999) focused on the validity of in-house English tests administered at the Department of Basic English (DBE) in Hacettepe University, Turkey. The study gathered validity evidence on content validity and predictive validity. A total of 510 C-level preparatory students and 34 English language lecturers participated in this study. For content validity, the questionnaires were administered to 34 English instructors to get their comments about the course content and about the content of the end of course assessment test. The results from the questionnaires revealed that the grammar and vocabulary parts of the test were represented the course objectives. For predictive validity, student's six midterm test scores were correlated with their end of course assessment test scores. The results on the comparison

of the test scores found a high correlation, ranging from $r=.72$ to $r=.90$. The findings suggested that the course content of DBE should be revised in terms of time spending on language skills and more equally fair to the students from all faculties

Rethinasamy and Chuah (2011) carried out a study which aimed at answering the predictive validity of the Malaysian University English Tests (MUET), which is an in-house English proficiency test to measure of Malaysian students' proficiency level, prerequisite for admission, as well as placement in various academic programmes. The MUET bands of 2884 undergraduates from various faculties and the grade they obtained in an English language enhancement were analysed to answer the research question. The data were collected through candidates' scores of a localised version of such tests called the MUET and then reported in a six-band scale with corresponding combined band score that ranges from 0-300. Each MUET sub-test score was scored separately and then averaged to obtain the overall band score. The Person's correlations showed a significant relationship ($r= 0.555$, $p =0.001$) between the undergraduates' MUET band scores and their grade in the Preparatory English 1 course. The results from Chi-square showed that MUET band 3 undergraduates obtained higher grades than undergraduates MUET band 2 and band 1. It means that undergraduates with higher MUET band scores can perform very well in the English language proficiency course. The study suggested to review the cut-off MUET band for placement purposes in the English language programme conducted by the university

Last but not least, Weir, Chan and Nakatsuhara(2013) investigated the criterion-related validity of the Reading and Writing components of the General English Proficiency Test which is an in-house test at a British university. A total of 171 students who were studying on a

full-time, collaborative, undergraduate program at the Business School participated in the study. Firstly, the relationships between the advanced level GEPT Reading and Writing scores and IELTS Reading and Writing scores were analysed to examine the concurrent validity. The results showed that the correlation between the participants' GEPT Writing bands and IELTS Writing bands were found low ($r=.243$, $p<.01$) due to the limited range of values for both tests. Furthermore, the correlation between GEPT Reading scores and IELTS Reading bands was significant ($r=.520$, $p<.01$). From these results, it seemed that it was more difficult for the students to pass the GEPT Advanced Writing test than Reading test. Secondly, the predictive validity of the advanced level GEPT Reading and Writing scores was correlated with students' academic performance. The results showed that the GEPT Reading and Writing total scores correlated with the participants' real-life academic performances at $.529$ ($p<.01$). Separately, GEPT Reading scores correlated with the real-life academic performances at $.499$ ($p<.01$) and GEPT Writing scores at $.294$ ($p<.01$).

2.8.3 Studies on Content validity

Abidin (2006) investigated the validity of an in-house spoken language test by using computer technology at Mara University of Technology, Malaysia. The framework for validating Speaking test proposed by Weir (2005) was adopted throughout the study to conduct validity evidence on content validity, theory-based validity and scoring validity. The study used both questionnaires and interview for gathering data for three test validation phrases on the direct Speaking test and computer-based Speaking test. The results generally showed that the overall correlation between direct and computer Speaking tests were high ($r=.70$). In particular, the data from students questionnaires indicated that students responded positively to test purpose, instructions and information for the tasks. Furthermore, students felt

pleased with the test condition and test administration in the direct test. However, the test takers had a general anxiety towards the direct test, rather than computer anxiety due to the presence of examiners. In the computer test, the interview data indicated that the computer test was more standardized than the direct test due to in the computer test students received different topics, information in tasks, linguistic variables. More importantly they did not interact with a live interlocutor, thus students' anxiety and nervousness were reduced when taking the speaking test. Furthermore, students' performance records in the computer test may be re-evaluated if needed for future rater training. In a nutshell, the researcher drew a conclusion that the computer speaking test was more valid in terms of content, theory based validity and scoring validity than the direct test.

Cumming et al. (2004) evaluated the content validity of in-house English Speaking and Writing tests to assess the domain of academic English required for studies at universities in North America. These researchers interviewed seven experienced instructors of English as a Second Language (ESL) at three universities to describe the ESL course they had been teaching, to rate their students' abilities in English and to review samples of their students' performance. At the end of interview stage, the instructors were asked to fill the questionnaire to evaluate the effectiveness of each prototype task. The results from the interviews indicated that the written and spoken responses to the readings and lectures were the best representatives for the domain of academic English required for university studies. Furthermore, 7 ESL instructors agreed that most of their students' performances on the prototype TOEFL tasks were similar to how these students usually performed on classroom tasks. Based on the results of the study, it was concluded that this study showed the relevant content validity of the prototype tasks being field-tested for a new version of the TOEFL test.

Fulcher (1997) investigated the validity of an in-house English Placement Test (EPT) at University of Surrey, in the United Kingdom. A total of 67 students at the University of Surrey were participants for this study. This study examined the validity of the EPT in terms of content validity, and concurrent validity. For content validity, 10 language teachers were interviewed to examine whether test items matched the skills taught. Furthermore, the study also correlated EPT scores with TOEFL scores to investigate the concurrent validity of the EPT. The results from interviews showed that EPT items were not very far from students' knowledge. Through Pearson correlation results, the researcher only found moderate relationship ($r=0.65$) between EPT scores and TOEFL scores. From these results above, it can be said that Fulcher's (1997) study provided some contributions to the development of test validation for in-house language test.

Kucuk (2007) examined the validity of an in-house English test administered at Zonguldak Karaelmas University Preparatory School, in Turkey. A total of 52 students and 29 English instructors participated in this study. The researcher used two questionnaires and students' test scores. The instructors and students were given questionnaires to ask for how well they judged the contents of the courses were presented on the achievement tests. The results showed that although Listening skill was not represented sufficiently on the test items, both English instructors and students still agreed that the Listening test still possessed high face and content validity.

Lee and Greene (2007) explored the validity of an English Second Language Placement Test (ESLPT) by using both qualitative and quantitative methods. A total of 100 students and 55 English lecturers at University of Illinois, in the United States participated in this study. For

content validity, the faculty evaluation questionnaire were given to 55 teaching staff to get the opinions about students' English proficiency, academic performance in the course, and the extent to which students' level of proficiency caught up their performance in the academic course. Interviews with 20 students and 10 faculty members during their office hours were conducted individually. The qualitative results showed that English skills contribute an important factor in affecting students' course performance. The quantitative results showed that there was no correlation ($r=.05$) between ESLPT score and students' GPA. Furthermore, the ESLPT did not correlate considerably with faculty members' ratings of performance in content courses ($r=.14$), but faculty members' ratings correlated moderately with first-semester GPA ($r=.53$). The findings indicated that international graduate students' English difficulties had less effect on students' academic performance than was expected, because of such other factors as sufficient background knowledge and lecture type courses during their first-semester studies.

Nofiyanti (2011) investigated the content validity of an in-house English summative test for the first year student of junior high school in Jakarta, Indonesia. The subjects included the English teacher and the first- year students. For data collection procedures, the researcher used questionnaire for students, then interviewed the English lecturers to ask about the English summative test. Then the researcher collected English course syllabus, teachers', students' profiles and students' English scores. The researcher employed both quantitative and qualitative approach methods for the descriptive analysis by getting information about the test from students and comparing the content of the English summative test with course syllabus. The results showed that 90% from 45 items was conformed to the curriculum, only 5% was considered as quite far from students; knowledge. Therefore, it can be concluded that

the summative test reached good content validity. The test developer suggested that a good English test should cover the instructions of the syllabus

Another validity investigation of an in-house English Speaking exam at Turkish University was conducted at Turkish University by Sar (2008). This research investigated the validity of the English Speaking exam at a Turkish University. A total of 70 students, English instructors and lecturers participating in this study were members of C level preparatory class students who took the December TOEFL-ITP. For this study, data were obtained through questionnaires to 70 students, interviews to English instructors, the students' English Speaking exam results, TOEFL exam results and departmental Speaking exam scores. Negatively worded items were included in the questionnaire, for example, the test content far from students' knowledge, irrelevant test instructions, and no interactions between students and teachers during the exam. The results of the interviews pointed out that the exam had the quality of content validity with a moderately high degree. More exactly, all interviewees hold the same opinions that the speaking exam met the objectives of their programme, that the students can produce simple connected texts on familiar topics such as personal interest, describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. To determine the predictive validity of the exam, students' scores of the Speaking exam and the scores of the Speaking exam given in departmental English courses were analyzed through Pearson Product Moment Correlation Coefficient and Simple Linear Regression Analysis. However, the results also showed that the Speaking exam given in preparatory year education did not seem to predict the performances of the students in the departmental Speaking exam with the p of 0.302. Based on the findings, the study concluded that English speaking exam seemed to have moderate content validity,

but there is no correlation between students' speaking test scores and TOEFL speaking scores; as a result this speaking exam could not predict students' performance at the Turkish university.

Sefisa (2012) evaluated the content validity of an in-house English test in the case of Ethiopian General Secondary Education Certificate (EGSEC-PE) in order to examine whether or not the EGSEC-PE ensures the representativeness of content coverage and sufficient sampling of the learning outcomes stated in the syllabus. Four evaluators or judges participated in this study. The study used interview and document analysis. This results show that there was a weak correlation ($p=.018$) between the syllabus and test items. That means the EGSEC- PE did not match with the syllabus contents, thus the EGSEC-PE test items did not prove to be valid for the content. The study suggested that test items should be constructed according to the syllabus and the extent of contents and a great attention should be given to exam preparation when preparing the examination

Siddiek (2010) investigated of the impact of content validity of in-house English language tests on both teachers and learners was carried out at Shaqra University, Saudi Arabia. Tests were supposed to focus on core syllabus constituents to find how much the learner achieved. In this attempt, the researcher used a critical and analytical approach by going through the examination samples that seemed to lack content validity. The findings showed that when the exam lacked this feature, most of the teaching would be focused in the techniques of examination rather than doing real teaching. It was also found that students did not refer to their textbooks as source of knowledge but exerted all efforts only to collect marks. The researcher suggested that a test without content validity would not create a high validity for

the test and easily spoil the whole educational process, so he recommended that the test writer should start with content standards to identify what students are expected to study during the academic year.

Based on these previous studies above, it can be noted that previous research studies on the field of language test validation are considered as very useful references for the researchers to have a closer look at the review of literature, methodology, findings, suggestions, or other ideas in order to complete their own studies. In addition, these previous studies may be valuable for many institutions in their endeavor to validate their in-house tests to justify the correctness of their interpretations. However, there have been some limitations in research design in terms of content and methodology. For illustrations, few studies (Advi, 2003; Al-Harj, 2013; Ayers, 1977; Dooley & Oliver, 2002; Huong, 2001; Mojtaba, 2009; Pishghadam & Khosropanah, 2011; Weir, Chan & Wu (2013) paid much attention to investigate the predictive validity of an in-house test more than other types of validity such as concurrent validity and content validity. In specific, the researchers tended to measure test scores rather than other knowledge, skills or other attributes of students. Messick (1996) emphasises that the meaning and values of test validation apply not just to interpretive and action implications derived from test scores, but also implications based on other modes of assessment. This means that in order to create high levels of validity of an in-house language test, the test should be investigated in different aspects of validity evidence both empirically and theoretically (Messick, 1996; Moore & Bounchan, 2010). Furthermore, it is generally suggested that the mixed methods (both qualitative and quantitative) in test validation helps to strengthen the degree of validity. Bachman (1990) indicates that while quantitative method helps to determine the effectiveness for predicting subsequent scores, qualitative approach

helps to understand the implementation and condition for validating in-house tests at a particular university. On the other hand, few previous studies tended to focus on one approach rather than more than the mixed methods for gathering validity evidence of their in-house English test. In the light of the review of literature above, the present study attempts to fill the gaps stated above by gathering both empirical and theoretical validity evidence and by employing both quantitative and qualitative approach methods to investigate the concurrent, predictive and content validity of the AEPET at a public university in Vietnam.

2.9 Chapter Review

This chapter reviewed the literature on important issues related to language testing. The issues included the relationship between testing with teaching and learning, definition of a test, types of test, and the qualities a good language test with an emphasis on test validity which is the most important characteristic. Then the socio-cognitive framework suggested by Weir (2005) was summarized in the chapter. Based on the importance of Weir's (2005) framework, the study adopted two validity components: content and criterion-related validity for validating the AEPET at a public university in Vietnam. Last but not least, the chapter ended with the brief summary of some previous studies that help the researchers as well as post graduate students to have closer looks at the tasks, specific methods, findings or other ideas in order to complete their final researches.

The next chapter, i.e. Chapter 3, describes the Methodology of the Study.

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

3.0 Chapter Overview

This chapter begins with the general direction of Methodology which provides appropriate methodological direction for answering the research questions presented in the previous chapters. The chapter presents detailed information of the Methodology of the study.

3.1 General Direction of the Methodology

The study aims to investigate the validity of the AEPET at a public university in Vietnam with a focus on three important validity evidence :concurrent validity, predictive validity and content validity. The study intends to answer the following research questions:

- (1) What is the relationship between the students' AEPET score and IELTS score?
(Concurrent Validity)
- (2) What is the relationship between students' AEPET score and academic achievement (CGPA), in comparison with the relationship between IELTS score and CGPA?
(Predictive Validity)
- (3) What is the Content Validity of the AEPET?

The first research question investigates the concurrent validity of the AEPET by seeking for the relationship between the students' scores in Listening, Reading, Writing, Speaking components and overall scores in the AEPET and IELTS, which is an standardized international English test. Since the concurrent validity is demonstrated when a test score

correlates with another standardized test score in which the two tests are taken at the same time and measure the same purpose. Just like IELTS, the AEPET assesses students' four English skills: Listening, Reading, Writing and Speaking and its format is similar to the format of the IELTS. This stage needs to take place immediately after the students completed the AEPET and then sit for IELTS. During this stage, the scores for each component and overall scores in the two tests would be collected from the university database through the AEP officers. The data would be keyed into SPSS and analyzed through the Pearson Product Moment Correlation Coefficient to find the correlation

The second one investigates the predictive validity of the AEPET by looking for the relationship between the students' AEPET overall and components- Listening, Reading, Writing, Speaking scores with their CGPA; the students' IELTS overall and components- Listening, Reading, Writing, Speaking scores with their CGPA. The aim of this stage is to address the question to what extent the AEPET predicts the students' academic success. This stage needs AEPET scores, IELTS scores and CGPA of the students who had graduated from university. During this stage, the scores for each component and overall in AEPET, IELTS and CGPA would be collected from the university database through the AEP officers. The data then would be keyed into SPSS and analyzed quantitatively through the Pearson Product Moment Correlation Coefficient and Simple Linear Regression to find the test of significance.

The last research question examines the content validity of the AEPET through questionnaire and semi-structured interviews for English lecturers. The aim of this stage is to determine the degree to which the AEPET content matches the AEP specification and syllabus; hence this stage needs to take place sometime after the year-end English course. The English language

lecturers would be required to assess the AEPET components: Listening, Reading, Writing, Speaking and the process of AEPET preparation

In fulfilling the requirements for carrying out this study, the research figures out the general direction of Methodology that the study would undertake in Figure 3.1 below

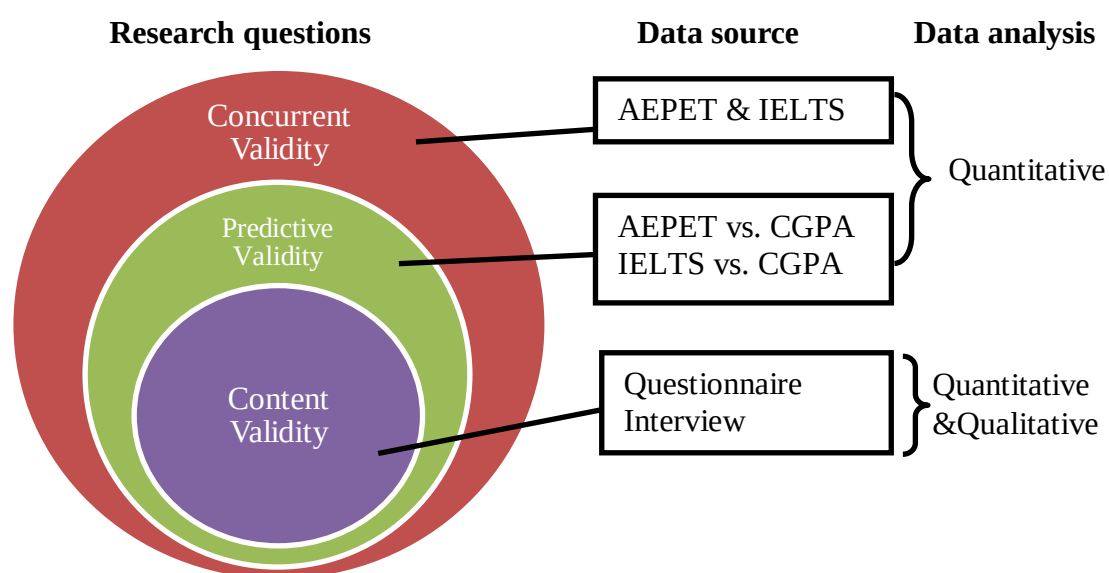


Figure 3.1: General Direction of Methodology

3.2 Research Design

In order to answer these research questions, the study adopted a two stage methodology. Stage One, which is presented in Section 3.2.1 describes the selection of materials and the design of instruments. Stage Two describes the Methodology of the Main Study which is presented in Section 3.2.2. Figure 3.2 presents a general structure of the Methodology of this study

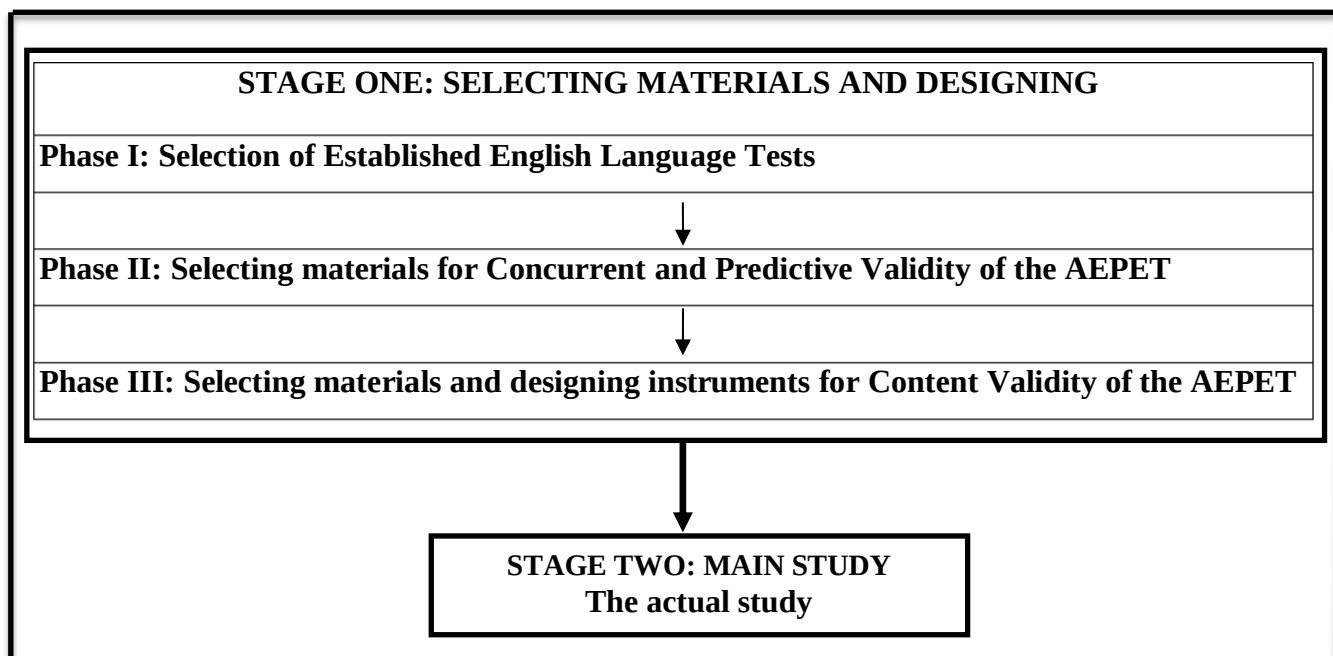


Figure 3.2: General Structure of the Methodology

3.2.1 Stage One: Selecting Materials and Designing Instruments

Stage One is the first stage in the methodology of the study. This stage consists of four phases of selecting materials and designing the instrument that were used to carry out the present study. Figure 3.3 presents the procedure for Stage One

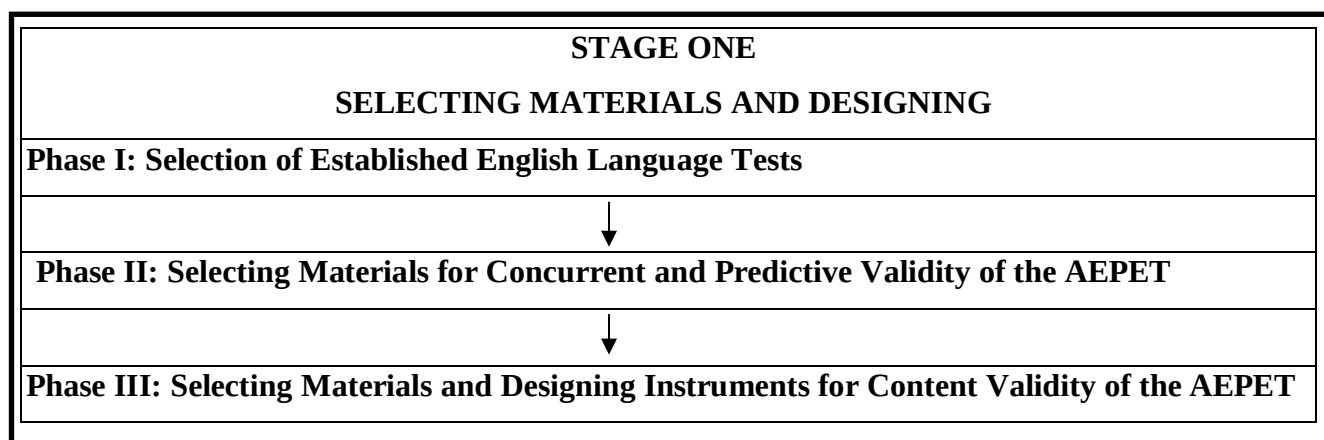


Figure 3.3: Procedure for Stage One

3.2.1.1 Phase I: Selection of Established English Language Tests

3.2.1.1.1 Selection of Standardized International English Language Test: IELTS

In recent times, there has been an increase in the number of overseas students choosing to study in universities in English-speaking countries. As a result, the issue of eligibility criteria has become an important one, with measurement of proficiency in the English language being a key aspect. At present, one of the most widely used tests for this purpose is the IELTS test developed by an international partnership between the University of Cambridge Local Examinations Syndicate (UCLES), the British Council, and the organization now known as IDP Education Australia. The IELTS scores are required to get the admission into many universities around the world and many immigration destinations require the learners to take the IELTS in order to qualify the English requirements (Dooley & Oliver, 2002; Huong, 2000; Rethinasamy, 2005). Huong (2000) emphasizes that the IELTS test serves as a useful medium to assess the language benchmark and in other cases to help the learners find good opportunities for their future career. According to Rethinasamy and Chuah (2011), selecting and using an English language test from an established examination to carry out the planned study is a crucial step which would ensure that the used test has undergone through the process of test construction and testing. In the light of the literature above, the IELTS was chosen to assess students' English proficiency in several universities in Vietnam

The IELTS assesses the candidates' English language proficiency through four English language tests: Listening, Reading, Writing and Speaking which are designed to cover the full range of ability from Non-User to Expert User through nine - band scale (Band 1.0 – Band 9.0) . There are two modules in IELTS: the General Training Module and the Academic Module. The candidates take the same Listening and Speaking components, but they choose

to take the Reading and Writing components either in the General Training Module or in the Academic Module. The General Training Module is for the candidates who plan to go to English speaking countries for secondary education, work experience or training programs. In contrast, the Academic Module is taken by the candidates, who intend to apply for academic study or higher education in an English speaking environment (IELTS, 2007). Accordingly, this study discusses on four IELTS components- Listening, Reading, Writing and Speaking in the Academic Module. The first three tests – Listening, Reading and Writing are required to complete in only one day and there is no break between them. The Speaking test may be taken in the period of seven days after the other tests in the test center. Figure 3.4 presents an overview of the IELTS

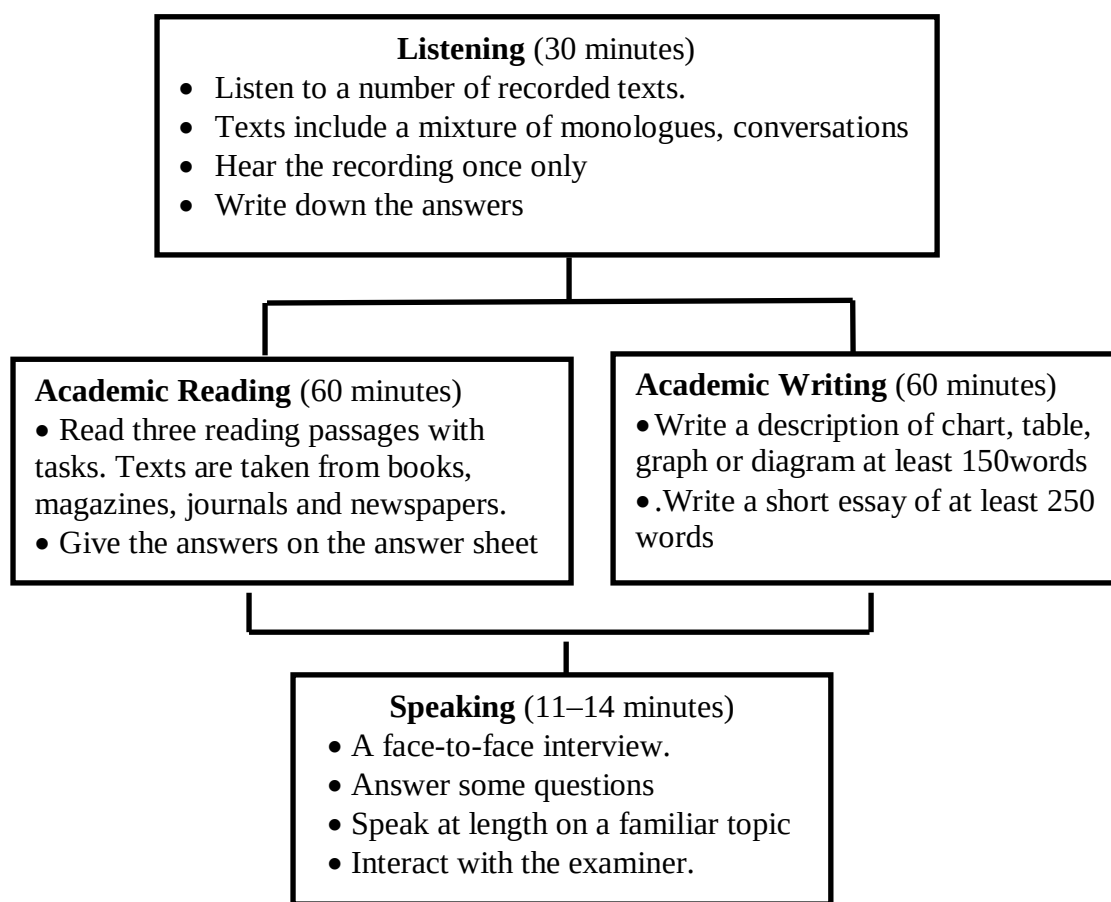


Figure 3.4: Overview of the IELTS

3.2.1.1.1a IELTS Listening Test

The Listening test consists of 40 questions in four sections with the duration of 30 minutes. The first two sections are a conversation between two speakers and then a monologue related to social needs such as travel arrangements, decisions on a night out, student services on a university campus and arrangements for meals during a conference. The final two sections are conversation between many people and a further monologue concerning with situations related more closely to educational or training contexts such as a conversation between a tutor and student about an assignment , between three students' planning a research project, and a lecture or talk of general academic interest. A variety of questions: multiple choice, short-answer questions, matching, sentence-completions, labeling a diagram and classification are used in each section. Each correct answer is awarded one mark

3.2.1.1.1b IELTS Reading Test

The Reading test consists of 40 questions based on three reading passages. The candidates must write their answers in an answer sheet during the 60-minute test. Texts are often taken from journals, newspapers, books, magazines, notices, advertisements, official documents, booklets, leaflets, and time-tables under different types of questions such as multiple choice, short-answer questions, sentence completion, matching headings for identified paragraphs/ sections of the text, identification of writer's views and identification of information in the text. One mark is awarded for each correct answer

3.2.1.1.1c IELTS Writing Test

In the Writing test, candidates are required to complete two tasks within 60 minutes. It is suggested that candidates should spend 20 minutes on task one and 40 minutes on task two.

For academic writing, candidates are asked to describe some information (graph/ table/ chart/ diagram), and to present the description in their own words in task one. In task two, candidates are required to present their points of view about an aspect of the real life. In this part, candidates have to show their ability to present and justify an opinion or present the way to give solutions for a problem. Each task is assessed independently. The assessment of Task 2 carries more weight in marking than Task one. Normally, Task one assesses on four criteria: the relevant response which fulfills the requirements set out in the task and uses the minimum of 150 words; the well- organized response which links information, ideas and language smoothly. Similarly, the Writing task two assesses on the relevant response which fulfills the requirements set out in the task with the 250 words in length; the ideas which are connected through logical sequencing and supported by evidence, examples; the wide range of vocabulary that candidates use in the specific task; the accurate use of the candidate's grammatical tenses

3.2.1.1.1d IELTS Speaking Test

The Speaking test lasts from 11 to 14 minutes and includes an oral interview between the candidate and an examiner through three parts: introduction and interview; individual long turn and two- way discussion. All Speaking tests are recorded. In part one, candidates are asked to answer general questions about themselves, their homes, families, their jobs, studies, their interests, and a range of familiar topics in the duration of five minutes. In the part two, candidates are given a verbal prompt on a card to talk about a specific topic. Before speaking at length between one and two minutes, they have one minute for preparation. Part three takes 5 minutes and candidates are required to answer further questions related to topics in the part two. Candidates' speaking performances are assessed by IELTS examiners through 3 main

criteria: the fluent and coherent speech; the wide range of vocabulary; the accurate and appropriate use of candidates' grammatical resources.

In short, IELTS has been an important internationally recognized exam that provides prove of learners' mastery of the English language (Dooley & Oliver, 2002; Huong, 2000). The IELTS with four macro-skills: Listening, Reading, Writing and Speaking, is not only closer and reflects the real life situations, but also helps to change people's lives in seeking for opportunities around the world in aspects of education, migration, and employment (IELTS, 2017). This is the reason why IELTS is strongly assessed as a high stakes test as well as are reliable measure of English language proficiency. Bearing this in mind, IELTS has been accepted as the international language standard for evaluating the quality of testing, teaching and learning English at the AEP at the public university where the study was conducted

3.2.1.1.1e Marking and Assessment for IELTS

The IELTS components – Listening, Reading, Writing and Speaking are assessed by certificated IELTS examiners who are professionally trained and approved by British Council or IDP IELTS in Australia (IELTS, 2015). Band scores are given from Band 1.0 to Band 9.0 for each test component and overall band score. The lowest is Band 1.0 and the highest is Band 9.0. The scores are rounded up to the nearest half bands or to the nearest whole band. Listening and Reading tests both consist of 40 questions, thus the results are marked on the same way. In specific, each correct answer is awarded one mark; the maximum raw score for each component is 40 and then the number of correct raw scores is converted to the IELTS nine-band scale (IELTS, 2015). Table 3.1 shows the IELTS Band Score Conversion for Academic Listening and Reading Tests taken from IELTS's (2015) handbook

Table 3.1: Band Score Conversion for IELTS Listening and Reading Tests

Band Score	9.0	8.5	7.5	7.0	6.5	6.0	5.5	4.5	4.0	3.5	3.0	2.5
No of correct answer	39-40	37-38	35-36	32-34	30-31	26-29	23-25	16-17	13-15	10-12	6-7	4-5

(IELTS, 2015)

The IELTS Writing component is marked based on the public version for IELTS Writing assessment criteria which describe detailed written performance for nine- scale IELTS bands (refer to Appendix 8 & 9). Normally, the two writing tasks are marked according to four criteria which are equally weighted (IELTS, 2015)

- Task Achievement
- Coherence and Cohesion
- Lexical Resource
- Grammatical Range and Accuracy

The IELTS Speaking component is marked based on the public version for IELTS Speaking assessment criteria which describes detailed spoken performance at each of the nine IELTS bands (refer to Appendix 13). The IELTS Speaking component is also marked according to four criteria which are equally weighted (IELTS, 2015)

- Fluency and coherence
- Lexical resource
- Grammatical range and accuracy
- Pronunciation

The overall band score is calculated by taking the average score of the four test component scores: Listening, Reading, Writing and Speaking. Then the overall band scores are rounded up to the nearest whole or half bands. For illustration, if the overall band score ends in .25, it is rounded up to the next half band, and if it ends in .75, it is rounded up to the next whole

band. Table 3.2 presents the rounding convention for IELTS components and overall band scores

Table 3.2. Rounding Convention for IELTS Components and Overall Band Scores

Listening	Reading	Writing	Speaking	Overall band score	Rounded overall band score
6.5	6.5	7	5	6.25	6.5
3.5	4.0	4.0	4.0	3.875	4.0
6.0	5.5	6.5	6.0	6.125	6.0

(IELTS, 2015)

Figure 3.5 presents the overall band descriptors and performance descriptors for the IELTS

BAND 9	EXPERT USER
Has fully operational command of the language: appropriate, accurate and fluent with complete understanding	
BAND 8	VERY GOOD USER
Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstanding may occur in unfamiliar situations. Handles complex detailed argumentation well	
BAND 7	GOOD USER
Has fully operational command of the language, touch with occasional inaccuracies, inappropriacies and misunderstanding in some situations. Generally handles complex language well and understands detailed reasoning	
BAND 6	COMPETENT USER
Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations	
BAND 5	MODEST USER
Has practical command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field	
BAND 4	LIMITED USER
Basic competence is limited to familiar situations. Has frequent problems in understanding and expression. Is not able to use complex language	
BAND 3	EXTREMELY LIMITED USER
Conveys and understands only general meaning in very familiar situations. Frequent breakdown in communication occur	
BAND 2	INTERMITTENT USER
No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty understanding spoken and written English	
BAND 1	NON USER
Essentially has no ability to use the language beyond possibly a few isolated words	
BAND 0	DID NOT ATTEMPT THE TEST
No assessable information provided	

Figure 3.5: Interpretations of IELTS Band Scores (IELTS, 2015)

3.2.1.1.2 Selection of an In-house English Language Test: AEPET

Each of a programme study's unit is followed by a test. The test is a useful tool to reflect the students' knowledge, understanding and skills; the teachers' teaching and assessment methods. Making the best selections of a test is very important for the success of the programme. Therefore, the AEPET was chosen because this is an in-house English language test which assesses the English capacities for students at the Advanced Educational Program, in Vietnam. As mentioned earlier, the AEP has been endorsed by Vietnamese Government for university students since the year 2000. Students are required to do one year English course and a three year degree program. In one year English course, students are taught Basic English subjects: Listening, Reading, Writing and Speaking. Furthermore, students have to take the midterm and semester end exams. These tests are designed to check the whole knowledge that students have learnt during the first semester. In the second semester, students are taught advanced English subjects: Reading, Writing and Speaking and take the midterm exam.

However, at the end of the second semester, students are required to take the final year English test, which is called the AEPET and then sit for IELTS one month later. The AEPET assesses students' ability in four language skills: Listening, Speaking, Writing and Reading. Basically, the test format is similar to IELTS format. Students take the Listening, Reading and Writing tests all on the same day one after the other, with no breaks in between them. The AEPET, which is given to students at the end of one-year English course, is designed to check the whole knowledge that students were taught in both semesters of the English course. Most importantly, the AEPET assesses students' ability in four language skills: Listening, Reading, Writing and Speaking. The AEPET format is similar to IELTS format. Students will take the AEPET components on the same day. Table 3.3 presents the AEPET format and weighting

Table 3.3: AEPET's format and weightage

Component	Time allocation	Part/Section	Question	Task	Scoring		Weightage (%)
					Score for a correct answer	Max score	
Listening	30 minutes including the transferring time on the answer sheet	Section 1	Q1-Q10	Complete the notes	0.2	10	20
		Section 2	Q11-Q20	Complete the notes			
		Section 3	Q21-Q32	Circle the correct answer			
		Section 4	Q33-Q41	Circle the correct answer			
		Section 5	Q41- Q50	Complete the notes			
Reading	60 minutes including the transferring time on the answer sheet	Reading passage 1	Q1-Q13	Choose the correct heading	0.2	10	20
		Reading passage 2	Q14-Q26	Choose the correct answer			
		Reading passage 3	Q27-Q40	Choose True/False/not given			
		Reading passage 4	Q42-Q50	Choose the correct answer			
		Writing task 1:	1	Report a graph/diagram/bar chart in 250 words	4.0	10	40
Writing	60 minutes	Writing task 2:	1	Write an argumentative essay in 500 words	6.0		
		Part 1: Introduction	4-5	Provide some personal information	3.0	10	20
		Part 2: Long-turn discussion	2	Retell a story/ an event with one minute	3.0		
		Part 3: Discussion	3-4	Discuss some global issues	4.0		
Speaking	11-14 minutes Part 1:4-5 mins Part 2:3-4 mins Part 3:4-5 mins						

3.2.1.1.2a AEPET Listening Test

The Listening Test has four sections which consist of 50 questions. These questions are aimed to test students' ability to understand main ideas and detailed factual information, ability to understand the opinions and attitudes of speakers, ability to understand the purpose of an utterance and the ability to follow the development of ideas. Students listen to some recordings related to everyday conversation between several people. The recordings is played only once within 30 minutes and students are asked to complete their answers into the answer sheets within the time limit. Each correct answer is awarded 2.0 marks. Section one, which starts from Q1 to Q10, is a daily conversation between some people in everyday context. Section two is a monologue or a single talk in everyday social context, starting from Q11 to Q20. Section three, which starts from Q21 to Q32, is a conversation between two people in an educational context. Section four is a number of short talks in social context, starting from Q33 to Q40. In four sections, students are required to complete the notes without more than three words for each answer or to circle the correct answer. Students are asked to listen to some tape-recordings in only one time within the duration of 30 minutes including the transferring time on the answer sheet and each correct answer is awarded 0.2 marks (refer to Appendix 4).

3.2.1.1.2b AEPET Reading Test

The Reading Test lasts 60 minutes and consists of 50 questions. A variety of question types is used in order to test a wide range of reading skills (reading for gist, reading for main ideas, reading for detail, skimming, understanding logical argument, recognizing writers' opinions, attitudes and purposes). In particular, the Reading Component consists of three reading passages with a serious of 50 questions. Reading passage one has 13 questions, starting from

Q1 to Q 13. In this part, students are asked to choose the correct heading for each paragraph from Q1 to Q5, then write True/False or Not Given in boxes from Q6 to Q10, write the correct letter in boxes from Q11 to Q13. Reading passage two also has 13 questions, starting from Q14 to Q26. Students are asked to write the correct letter in boxes from Q14 to Q26. Reading passage three has 14 questions, starts from Q27 to Q40. Students are asked to write the correct letter from Q27 to Q32, and then write True/False or Not Given from Q33 to Q40. Students have 60 minutes to complete the test and are advised to spend 20 minutes on each reading passage. Each correct answer is awarded 0.2 marks (refer to Appendix 5).

3.2.1.1.2c AEPET Writing Test

The Writing Component has two tasks in which students are advised to spend 20 minutes on task one and 40 minutes on task two. Students are asked to describe some information (graph/table/chart/diagram) in their own words in task one with the uses minimum of 250 words in Task one and then write an essay in response to a point of view, argument or problem in a formal style with the uses minimum of 500 words in the Task two. Task one is awarded 4.0 marks and 6.0 marks for Task two. This component takes 40% of the total score of the AEPET (refer to Appendix 6)

3.2.1.1.2d AEPET Speaking Test

The Speaking Test assesses students' use of spoken English, and takes between 11 and 14 minutes to complete. The Speaking paper is divided in 3 main parts. In Part one, students are asked to talk about their personal information such as family, home town, study and hobbies within 4-5 minutes. In Part two, students select one of the particular topics then talk within 3 minutes. In Part three, the examiner asks students some further questions related to topic

discussed in Part two within 3 -5 minutes. The score is given in different parts depending students' way of using language. The maximum score is 3.0 for each Part 1 and Part 2 and 4.0 for Part 3. This component takes 20% of the total score of the AEPET (refer to Appendix 7).

In short, the AEPET measures the students' English language abilities in four skills: Listening, Reading, Writing and Speaking. The AEPET format is similar to the format of the IELTS. The students sat for the AEPET and then for IELTS, therefore the AEPET is considered as a part of preparation for the IELTS.

3.2.1.1.2e Marking and Assessment for the AEPET

The AEPET provides the overall band score and band scores for each of the four components: Listening, Reading, Writing and Speaking. All results use the same band scale which ranges from Band 1 to Band 10. Scores are rounded up to the nearest half bands or whole bands

Both Listening and Reading components carry the same 20% weightage of the AEPET. Students take a series of 40 test items for each component. Each correct answer is awarded 0.2 marks, accordingly ten marks are awarded for 50 correct answers in each component. Furthermore, poor spelling and grammatical answers on the answer sheet are considered as wrong answers in the two component results

The Writing component takes 40% of the AEPET and consists of two writing tasks. Each task is marked differently, 4 marks for Task 1 and 6 marks for Task 2. Each task is assessed independently due to the fact that Writing Task 2 carries more weight in marking than Writing Task 1. Table 3.4 presents the criteria to assess the response in AEPET Writing Task 1

Table 3.4: Assessment Criteria for AEPET Writing Task 1

No	Criteria	Mark
1	Providing an overview	0.5
2	Presenting accurate factual information	0.5
3	Highlighting key features	0.5
4	Supporting details with data	0.5
5	Using accurate grammatical resource	0.5
6	Using a range of transition words to link ideas logically and clearly	0.5
7	Using appropriate vocabulary	0.5
8	Be at least 250 words in length	0.5
Total		4.0

Table 3.5 presents the criteria to assess the response in AEPET Writing Task 2

Table 3.5: Assessment Criteria for AEPET Writing Task 2

No	Criteria	Mark
1	Addressing the task	1
2	Giving relevant main points, supporting details and stating the own opinions	1
3	Using a range of transition words to link ideas logically and clearly	1
4	Using a range of grammar tenses and sentence structures	1
5	Providing a conclusion	1
6	Be at least 500 words in length	1
Total		6.0

Speaking component carries 20% weightage of the AEPET. It comprises of three main parts. Each part is marked differently. In specific, Part 1 and Part 2 are awarded the same with 3 marks while Part 3 is awarded 4 marks. Although each part is marked differently, all parts' responses in the Speaking component are assessed with the same criteria on the students' ability to

- speak fluently and clearly to fulfill the task requirements
- use the range of sentence structures
- use the range of accuracy grammatical resource

- use the logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices within and between sentences

It can be noted that each component: Listening, Reading and Speaking carry 20% while the Writing carries 40% of the AEPET. Thus, the AEPET overall band score is calculated in the following formula below:

$$\text{Overall} = \frac{((L+R+S)*20 + (W*40))}{100}$$

The overall band scores are then rounded up to the nearest half bands or whole band. Table 3.6 illustrates the formula of the AEPET Overall score

Table 3.6: Formula of AEPET Overall Score

Listening	Reading	Writing	Speaking	Overall	Rounded overall
3	3	5	6	$\frac{(3+3+6)*20 + (5*40)}{100} = 4.4$	4.5

The Band Descriptors are made in the agreements by English experts and administrators at the university. They provide ten-band level and are used as a guide for interpreting the band scores. Furthermore, the Band Descriptors scale provides a performance descriptor for each component and overall band score level, ranging from Band 1.0 to Band 10, which are classified from below average to excellent in English proficiency. Table 3.7 shows the performance descriptors and band descriptors of the AEPET band scores

Table 3.7: Band Scores and Band Descriptors for AEPET Overall Band Scores

AEPET Band	Performance Descriptors	Band Descriptors
1.0 – 1.5	Completely have no ability to use the language	Below
2.0 – 2.5	Have greatly difficulty using and understanding the language	average
3.0 – 3.5	Have limited ability to use the language with familiar situations	
4.0 – 4.5	Have basic command of language with frequent breakdowns in understanding and expression	Average
5.0 – 5.5	has practical command of the language despite many errors and can handle basic communication in own field	Moderate
6.0 – 6.5	have generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings and can use complex language in familiar situations	Competent
7.0 – 7.5	Operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations and can generally handle complex language	Good
8.0 – 8.5	Fully operational command of the language and can handle complex detailed argumentation well	Extremely good
9.0 - 10	Have fully operational command of the language; appropriate, accurate and fluent with complete understanding	Excellent

3.2.1.2 Phase II: Band scores for Criterion-related Validity of the AEPET

3.2.1.2.1 Band scores for AEPET and IELTS

Both AEPET and IELTS measure the same constructs: Listening, Reading, Writing, Speaking and Overall band scores. However, the range scores for each test are slightly different. The AEPET has 10 band scores, ranging from Band 1 to Band 10. Unlike AEPET, the IELTS has 9 band scores, ranging from Band 1 to Band 9. This means that the range of AEPET differs only in two band scores (Band 9.5 and Band 10) from the range of IELTS. Furthermore, both AEPET and IELTS are continuous ranges, which possess equal intervals between two band

scores (Treiblmaier & Filzmoser, 2009). In specific, they have the same interval band score which is equal at the length of 0.5 points. This distance remains all the same along the range in the two tests. Figure 3.6 shows the distance between band scores for AEPET and IELTS

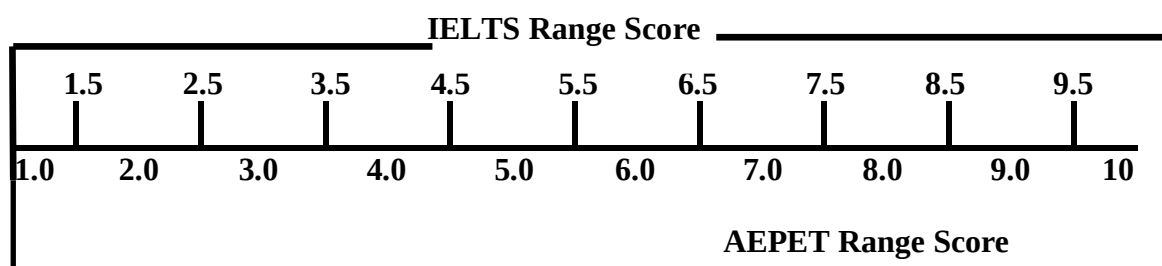


Figure 3.6: Interval Band Score for AEPET and IELTS Ranges

3.2.1.2.2 Cut-off scores for AEPET and IELTS

After completing one-year English course, the students are required to take the AEPET which comprises of four components: Listening, Reading, Writing and Speaking. The AEPET is designed to assess the whole knowledge that students have been taught during two semesters, and then one month later, they have to sit for IELTS compulsorily in order to prove their English language proficiency. In other words, both the results of AEPET and IELTS will accurately reflect students' English language level. The AEPET is a test which assesses students' English proficiency at institutional level while IELTS is a test which assesses students' English proficiency at international level. If students' AEPET and IELTS overall scores are Band 6.5 or above and their CGPA for year-year English course is 7.0 or above, they will be offered a full scholarship to continue their degree program within 3 years at a public university in the United States .Others students who do not meet the criteria, will continue their degree program at a public university in Vietnam (refer to Section 1.2.3, Chapter 1). Thus, it can be said that Band 6.5 is a common cut-off score for both AEPET and IELTS. On AEPET, the students who scored this Band seem to have effective ability in

English proficiency despite some inaccuracies, inappropriacies in language use (Refer to Chapter 3, Section 3.2.4). Similarly, on IELTS, Band 6.5 is also a common cut-off score for admission to selective universities in all over the world. The students who obtain this Band seem to have generally effective command of the language despite some misunderstandings and can understand fairly complex language, particularly in familiar situations.

3.2.1.2.3 Cumulative Grade Point Average

For universities, Vietnamese educational system has implemented a common grading system which consists of three official grades: ten-band scale, letter Grade (A, B, C, D, F) and four-band scale (Band 1.0 – Band 4.0) to assess students' Grade Point Average. Students' first scores are given under ten-band scale, and then converted into letter grade which ranks from highest to lowest: A, B, C, D and F. The grade A is excellent, D is below average, any score which is less than D, is F. Finally, their letter grades are converted to four-band-scale which ranges from highest Band 4.0 to lowest Band 1.0. The final CGPA score is not rounded to the nearest half bands or whole bands. Students' must obtain a minimum of CGPA of 5.5 according to ten-band scale (letter grade C and Band 2.0) to be considered for graduation condition at the university level. Table 3.8 presents the coding of numerical marks and letter grades for grading system in Vietnam

Table 3.8: Numerical Marks and Letter Grades for Grading System

Ten-band scale	Grade	Four-band scale	Meaning
8.5 - 10	A	4.0	Excellent
7.0 - 8.4	B	3.0	Good
5.5 - 6.9	C	2.0	Average
4.0 - 5.4	D	1.0	Below average
0.0 - 3.9	F	0.0	Fail

3.2.1.3 Phase III: Designing Instruments for Content Validity of the AEPET

3.2.1.3.1 Selecting Materials: Content Validity Questionnaire

Questionnaires are the most popular instrument in the data collection method in educational evaluation research because they help to gather information on knowledge, attitudes, opinions, behaviors and other information in a short period of time as well as at a relatively low cost (Zohrabi, 2013). Bearing this in mind, questionnaire was used to collect the lecturers' judgments about the content validity of the AEPET as well as to answer the first research question of the study. During the first phase of Stage One, some content validity questionnaires from previous studies (Cesur & Koksak, 2012; Dogru, 2013; Kucuk, 2007, Scar, 2008; Huong, 2000; Jaturapitakkul, 2013; Moore, 2015; Wang, 2006) were collected. The focus on test weightage, time allocation, representation of language skills, representation of topics, clarity of questions, clarity of instruction and mark allocations in these previous studies were listed in order to gather necessary items which are useful for examining the opinions about the validity of a language test. Next, questionnaire for content validity of the AEPET was drafted from previous studies and then refined to make sure that the adaption of the instrument would meet the requirements for investigating the lecturers' opinions about the validity of the AEPET.

3.2.1.3.2 Designing Content Validity Questionnaire

The content validity questionnaire of the AEPET is drafted for English language lecturers in this study. It consists of two main parts: Consent cover letter and Content of the questionnaire (refer to Appendix 3). The consent cover letter which is the first part of the instrument construction aimed to gain permission to conduct the data from the lecturers. It begins with a brief introductory statement about the study and the researcher. Furthermore, the promise of

confidentiality is compiled in this letter to help the participants understand that their responses will not be in any case that affects their academic study or their academic career. Finally, contact and return information that is helpful to deal with queries during the data collection procedure is also included in the letter. The content of the questionnaire content which is the main part of the instrument construction, consists of three sub-sections: Background information (Section A), AEPET Components (Section B), AEPET Preparation (Section C). An outline for the content validity questionnaire construction for both students and lecturers is presented in Figure 3.7 below

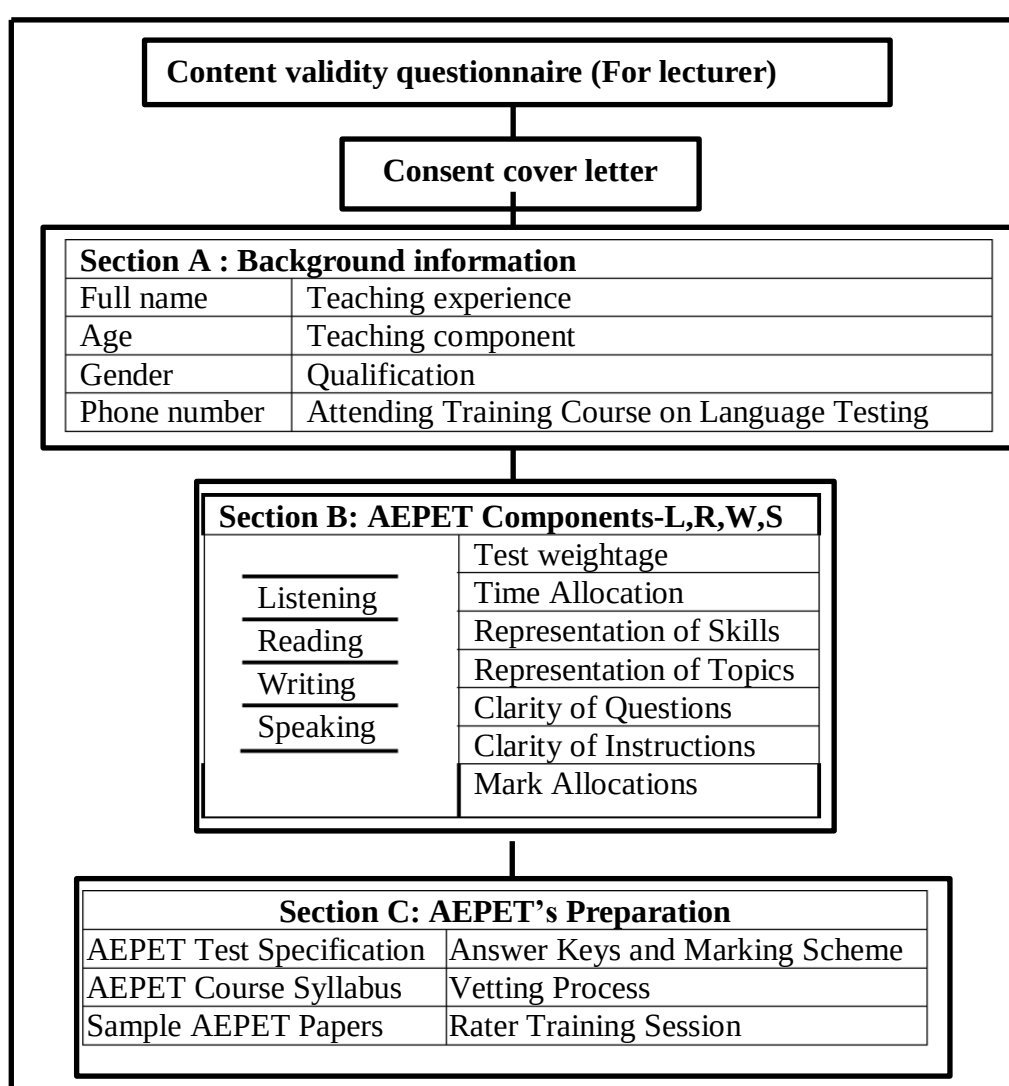


Figure 3.7: Outline of English Lecturers' Content Validity Questionnaire

3.2.1.3.2a Drafting Background information

More specifically, Section A aims to ask for the lecturers' background information. This section comprises of 10 questions (Q1-Q10) which were constructed for seeking information on the lecturers' full name gender, age, academic qualification, year of teaching English, English subjects that they teach at the university and the training course attendance on language testing

3.2.1.3.2b Drafting Content validity Questionnaire for AEPET Components

For AEPET Components, 7 items which were on the weightage, time allocation, representation of skills, representation of topics, clarity of questions, clarity of instructions and mark allocation were determined then developed the first draft to make sure that all items were well- organized and fully met the requirements of research objectives. These items were constructed repeatedly in each component: Listening, Reading, Writing and Speaking of the AEPET with the total of 28 questions. The responses for 28 questions are ranked from Strong disagreement to Strong agreement (SD-Strongly disagree; D- Disagree; N- Neutral; A- Agree; SA- Strongly agree). The modification process of the content validity questionnaire for each component is presented briefly in Figure 3.8 below

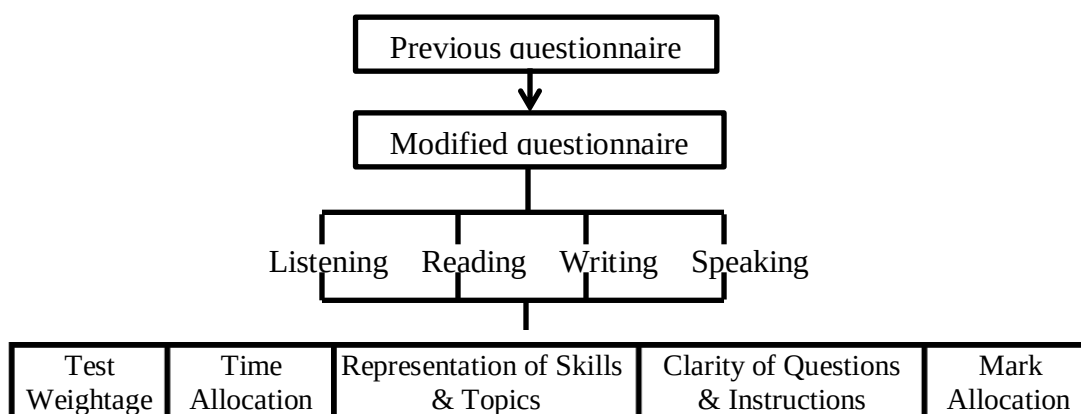


Figure 3.8: Modification Process for Content Validity Questionnaire

Test Weightage

The question item related to test weightage was modified from the original item “The weighting demonstrated on the marking scale of the test was appropriate” in Huong’s(2000) study. To be more specific, the word “weighting” in the original question was changed into “weightage” because the word “weightage” is used more frequently in the daily life, so this modification was to help the participants understand the question better and easier. Furthermore, the phrase “demonstrated on the marking scale” of the test was omitted to make the modified question simpler, and then the names of four sub- skill tests; Listening, Reading, Writing and Speaking were added to make the item more relevant. This question item aimed to ask for overall assessment of the AEPET and provide the participants in-deep understanding about the AEPET. Table 3.9 indicates the modified questions on test weightage

Table 3.9: Modified Questions on Test Weightage

Original question	Modified questions	Source
The weighting demonstrated on the marking scale of the test was appropriate	Test weightage for the listening test was appropriate Test weightage for the reading test was appropriate Test weightage for the writing test was appropriate Test weightage for the speaking test was appropriate	Huong (2000)

Time Allocation

For time allocation, the words “the midterm exams” in the original question “The time given to complete” in Liauh’s (2011) study were replaced by a new phrase “Time allowance” in order to make the sentence simpler and easier to understand. Then four names of the subtests (listening test/ reading test/ writing test/ speaking test) also was replaced the words “the midterm exams” to provide specific information. Finally, the original word “enough” was changed into “sufficient” in the modified questions in order to make the questions more

formal but still keep the similar meanings. As a result, four modified questions asking for time allocation for each English subtest were developed. Table 3.10 indicates the modified questions on time allocation

Table 3.10: Modified Questions on Time Allocation

Original question	Modified questions	Source
The time given to complete the midterm exams was enough	Time allowance for the listening test was sufficient Time allowance for the reading test was sufficient Time allowance for the writing test was sufficient Time allowance for the speaking test was sufficient	Liauh (2011)

Representation of Language Skills and Topics

Two words “language” and “exams” in the original questions “Language skills taught were sufficiently represented in the exams” were replaced by the word “Topics” and the names of four components: Listening, Reading, Writing and Speaking in order to make modified questions more specific as well as examine the presence of represented English language skills and topics in the test (Kucuk, 2007). Table 3.11 indicates the modified questions on the representation of language skills and topics

Table 3.11: Modified Questions on Representation of Language Skills and Topics

Original question	Modified questions	Source
Language skills taught were sufficiently represented in the exams	Listening skills taught were sufficiently represented in the listening test Reading skills taught were sufficiently represented in the reading test Writing skills taught were sufficiently represented in the writing test Speaking skills taught were sufficiently represented in the speaking test Topics taught were sufficiently represented in the listening test Topics taught were sufficiently represented in the reading test Topics taught were sufficiently represented in the writing test Topics taught were sufficiently represented in the speaking test	Kucuk (2007)

Clarity of Questions

Questions focused on test question items were constructed in the similar ways with the total of 4 items. These items were adapted from some researchers' questionnaire items (Cesur & Korsal, 2012; Jaturapitakkul, 2013; Kucuk, 2007; Kuntasal, 2001). Only the word “exams” in all original questions was replaced by the names of four test components consisting of Reading, Listening, Writing and Speaking while other words were kept originally. The purpose of this replacement was to make the adapted questions more relevant. The explanations of modified questionnaire items for four test components were summarized in the table below

Table 3.12: Modified Questions on Test Questions

Original question	Modified question	Source
The questions in the exams were clear	Questions in the listening test were clear	Cesur & Korsal (2012)
	Questions in the reading test were clear	
	Questions in the writing test were clear	Kucuk (2007)
	Questions in the speaking test were clear	

Clarity of Instructions

Questions about the perceptions on test instructions consisted of 4 items which were adapted from one original question in Kucuk's (2007) study. In the original question “The instructions explaining what to do in each section in the exams were explicit and clear”, the word “exams” was replaced by four English test components: Listening, Reading, Writing and Speaking. Furthermore, the word “explicit” in the original was omitted in the sentence. It was due to the fact that the original question was a double –barreled question which asked two or more issues at once. This kind of question cannot indicate the respondents how much agree or disagree of the answer, so it will cause the confusion and inaccuracies in the received answer.

Therefore, the modified question had only one choice “The instructions explaining what to do in each section in the listening/ reading/ writing/speaking test were clear”. Table 3.13 presents the modified questions on test instructions

Table 3.13: Modified Questions on Test Instructions

Original question	Modified questions	Source
The instructions explaining what to do in each section in the exams were explicit and clear	The instructions explaining what to do in each section in the listening test were clear	Kucuk &Walters (2009)
	The instructions explaining what to do in each section in the reading test were clear	
	The instructions explaining what to do in each section in the writing test were clear	
	The instructions explaining what to do in each section in the speaking test were clear	

Mark Allocations

Finally, questions about the perceptions on mark allotments consisted of 4 items which were adapted from one original question in Kucuk and Walters’ (2009) study. Some words and phrases “the; always; in the exam papers” from the original question were omitted in order to make the modified question simpler but still keep the same meaning. Furthermore, the names of four test components: Listening, Reading, Writing and Speaking were added to make the questions more relevant. Table 3.14 presents the modified questions on mark allotments

Table 3.14: Modified Questions on Mark Allocations

Original question	Modified questions	Source
The marks allotted for each section of the exam were always stated clearly in the exam papers.	Marks allotted for each section of the listening test were stated clearly	Kucuk (2007)
	Marks allotted for each section of the reading test were always stated clearly	
	Marks allotted for each section of the writing test were always stated clearly	
	Marks allotted for each section of the speaking test were stated clearly	

In short, the instruments for investigating the content validity of the AEPET were selected on the existing items in previous questionnaires designed by several researchers and then were drafted into two questionnaires which are used for collecting opinions from Students and Lecturers on weightage, time allocation, the representation of language skills, the representation of topics, the clarity of questions, the clarity of instructions and mark allocations. In order to prove the reliability of the modified instruments, the process of testing instruments would be presented the piloting stage of the section.

3.2.1.3.2c Drafting Content validity Questionnaire for AEPET Preparation

Agu, Onyekuba & Anyichie (2013) indicate that teachers use tests to check students' knowledge and skills, therefore it is necessary to examine what precautionary steps they take to ensure that the tests are valid. Thus, in order to ensure the validity of the tests used by teachers, appropriate test construction procedures should be employed by them. Furthermore, Al-Younes (2006) suggests that the administrators, educators should give more attention and care to the content validity of in house tests before and after their construction and administration. Several researchers indicate that the processes for developing and administering test items for use which are normally based on

- 1) referring to carefully constructed test specification (Popham, 2002; Siddiek (2011)
- 2) referring to course objectives as provided for in the curriculum for each subject (Al-Shara, 2011; Al-Younes, 2006; Siddiek (2011)
- 3) assuring the clarity of the questions submitted by item writers to prevent the misunderstanding the questions between students (Wang, 2006)
- 4) sampling of the topics, situations, types, function, number of test items, structures and tasks that were covered in the class (Karr, 2011)

- 5) assuring of acceptable properties of the difficulty level between test items (Daniel & Deber, 1998; Popham, 2002)
- 6) evaluating and assessing open-answer questions (Agu, Onyekuba & Anyichie, 2013; Al-Shara, 2011)
- 7) reviewing and updating test items to meet current requirements of the test (Al-Shara, 2011; Popham, 2002)
- 8) practicing the procedures of analyzing test results (Al-Shara, 2011; Al-Younes, 2006)
- 9) discussing of the test results with other experts (Al-Shara, 2011)

These standard criteria not only ensure high content validity for the examinations, but also a guarantee that what is expected to be tested at a particular level of schooling is what is actually being tested. The tests which possess content validity of before, during and after test development give a comprehensive coverage of the specific language elements to be included, provide a clear picture of what students have or have not achieved, help teachers to diagnose the weaknesses and strengths of their students and to follow up their students' learning progress and then provide a comprehensive educational process.

From the outcome of the literature review above, questionnaire items on Section C which is on content validity of AEPET Preparation was designed by the researcher to further evaluate the presence of the procedures for the development of test items as well as to investigate whether teachers use appropriate construction procedures which confer on tests the requisite validity. This questionnaire focused on four main areas which are on reference to test specification, course syllabus, test sample papers when constructing the test items; preparation for questions, answer keys for the objective questions and marking scheme during the item

construction; vetting session for questions, answer keys for the objective questions and marking scheme; and rater training session before assessing the answer scripts assigned to the English language lecturers. This section consists of 25 questions (Q39-Q63). The responses for 25 questions are ranked from Strong disagreement to Strong agreement (refer to Appendix 3).

3.2.1.3.2d Semi-structured Interview

According to Blandford (2013) and Szombatová (2016), a semi-structured interview is a meeting in which the interviewer does not strictly follow a formalized list of questions. The questions' orders can be modified based upon the interviewer's perception of what seems most appropriate. Question wording can be changed and explanations given; inappropriate questions for a particular interviewee can be omitted, or additional ones included. The interviewer asks more open-ended questions, allowing for a discussion with the interviewee to gain extra information. Bearing this in mind, the semi-structured interview for investigating the content validity of the AEPET was conducted after the questionnaire survey was administrated. The interviewer asked follow-up questions to gain a more in-depth understanding about the process of AEPET preparation. Notes taken down during the interview and sample excerpts from the English language lecturers are presented in the results chapter.

3.2.1.3.3 Pilot Study

In this section, the content validity questionnaire was piloted in order to test the applicability of the first version of the instruments for investigating the content validity of the AEPET, make necessary modifications, refine the items in the questionnaire for the purpose of the

main study provide hands on experience to carry out the main study. Firstly, the appointments with the 3 English language lecturers were made at the program meeting room at the university. Secondly, the instructions of the questionnaire were read out to make sure that the lecturers feel free to participate in the survey and ensure that the lecturers would not have any difficulty with the items. Secondly, the lecturers were asked to fill out their background information questionnaire (Q1-Q10). All of lecturers had at least five three years of experience in teaching English and designing tests. All of them had Master degrees or Ph.D. degrees in teaching English as a secondary language. Two of the lecturers teach four English skills while one lecturer teaches Writing skill. Three of them did not attend any training course con language testing and assessment. After fulfilling their personal information, the lecturers were provided sample of the AEPET components: Listening, Reading, Writing and Speaking. The lecturers were asked to circle only one answer from Question 11 to Question 63. The time for the lecturers to complete the questionnaire was noted down. Finally, the responses were checked and administered.

The results showed that the instructions of the questionnaire were clear, thus the lecturers easily understood and followed the instructions which were read to them from the instruction sheet and they felt free to fill in the questionnaire. Furthermore, all lecturers did not have any problem in filling out the background information questionnaire. It took 2 minutes to fill out 10 questions related to their personal information. Then the second part lasted 25 minutes for the lecturers to answer 53 questions. The lecturers held the same view that items in the questionnaire reflected the content validity of AEPET components and AEPET preparations were easy to understand. In other words, it can be concluded that pilot participants did not have any problems in the following instructions, completing the items with the time allocated

and filling out the data collection instruments designed for questionnaires. Thus, no modifications were made to the instruments and time allocated for completing the items in the questionnaire. From these findings, it can be concluded that the instruments using in Pilot study can be completely used in the Main Study.

3.2.1.4 Section Review

This section is on Stage One of the study presented the selection of materials and construction of instruments for the purpose of the Main Study. This section also presented the modification process on the first draft of the instruments for content validity of the AEPET and the pilot study which was conducted to test the reliability of the instrument as well as make the instruments better fit for the purposes of the Main Study. The next stage of Chapter, i.e. Stage Two describes the Main Study.

3.2.2. Stage Two: Main Study

The methodology of the research consists of 2 stages. Stage One, which was presented in Section 3.2.1 described the selection of materials for the validity of the AEPET and the design, testing and refining the content validity questionnaire. This section- Stage Two describes the Methodology of the Main Study.

3.2.2.1 Participants

For concurrent validity investigation, data from three batches of students were included. The data of 103 AEP students who completed their English and then sat for IELTS were obtained from the university database.

For predictive validity investigation, participants of this study comprised of 143 graduates who had taken the AEPET and then sit for IELTS.

For content validity investigation, 20 English language lecturers participated in the survey and interview. These lecturers teach English at the AEP and their ages range from 30 to 50 years. Moreover, they have been experience in teaching, designing the English test as well as assessing the students' language abilities.

3.2.2.2 Purpose

The purposes of the Main Study are

- to determine to what extent to results of AEPET components scores correspond to those of the IELTS (Concurrent Validity).
- to address the question to what extent the AEPET determines academic success for the AEP students (Predictive validity).
- to access lecturers' judgments about the AEPET whether the AEPET content reflects the knowledge/ skills required to do in the AEP Course syllabus as well as to find out to what extent the process of AEPET preparation is present and applicable before the examination is administrated (Content Validity).

3.2.2.3 Materials and Instruments

For Concurrent Validity:

- 103 Vietnamese undergraduate students' AEPET and IELTS scores

For Predictive Validity:

- 143 graduates' AEPET, IELTS scores and CGPA

For Content Validity

- CVQ for AEP English language lecturers (refer to Appendix 3)
- Semi-structured interview for AEP English language lecturers

3.2.2.4 Data Collection Procedures

For Concurrent Validity: A formal letter requesting for AEPET scores and IELTS scores for the research was sent to the rector of the university. Thus, when the permission was granted, 103 students' AEPET scores and their IELTS scores for investigating the concurrent validity of the AEPET were collected from the university database given by AEP officers.

For Predictive Validity: For determining the predictive validity of AEPET, a formal letter requesting for the AEPET, IELTS and the CGPA scores of students, who have completed their study, was sent to the rector of the university. Once the permission was granted, 143 students' scores for validating the predictive validity of the AEPET were retrieved from the university database and were given by AEP officers.

For Content Validity: In order to answer the third research question which relates to the content validity of the AEPET, the questionnaire survey was conducted 1 month after the English language lecturers finished their teaching English course. 20 questionnaires for English language lecturers were distributed and 20 completed questionnaires were returned. The semi-structured interviews were carried out after administrating all the responses in order to gain additional information and also to clarify some interesting and unclear responses in the

questionnaire. During the interview, notes were taken down and then sample excerpts from the lecturers' would be presented in the third part of the results section in the thesis.

3.2.2.5 Data Analysis Procedures

For Concurrent Validity: Firstly, 103 students' AEPET and IELTS scores were keyed into Statistical Software for Social Science (SPSS) version 23 and then descriptive statistics were computed in order to see how the students have performed on each component and overall score in the two tests. Lastly, Pearson Correlation was used to determine the correlation between AEPET and IELTS scores.

For Predictive Validity: Firstly, 143 students' AEPET, IELTS and CGPA scores were coded and processed in SPSS version 23. Secondly, Pearson Correlation was used to determine the correlation between AEPET scores and CGPA; between IELTS scores and CGPA. Finally, Simple Linear Regression was calculated to determine the impact of the AEPET and IELTS on students' CGPA.

For Content Validity: The data obtained from 20 English lecturers were coded in to SPSS, version 23. The data then was analyzed separately in three steps. Firstly, the demographic information was computed to see the status of the respondents' background. Secondly, the Cronbach's alpha values were computed to check the reliability the questionnaire. Finally, the responses for each item in each part in the questionnaire were analyzed quantitatively and qualitatively with descriptive statistics in order to establish the content validity of the AEPET. Table 3.15 presents the summary of data collection and analysis.

Table 3.15: Summary of Data Analysis Procedures

Research questions	Instrument	Data analysis
Concurrent validity	Students' (103) AEPET scores and IELTS scores	Pearson Product Moment Correlation Coefficient
Predictive validity	Graduates' (143) AEPET, IELTS scores and their CGPA	- Pearson Product Moment Correlation Coefficient - Simple Linear
Content validity	- Questionnaire – (20) English lecturers - Semi-structured interview –(20) English lecturers	The responses was analyzed quantitatively (mean scores) and qualitatively (thematic)

In conclusion, in order to be able to investigate the validity of the in house language test, different kinds of instruments: students' AEPET, IELTS and CGPA scores; questionnaire and semi-structured interviews for English language lecturers were used to collect data. Bachman (1990) also indicated that the mixed method approaches in test validation can create effectiveness in combination with one another. He explained that while qualitative approach helps to understand the implementation and condition for validating in-house language tests at a particular university, quantitative method helps to determine the effectiveness for predicting subsequent scores. By employing these methods, the aspects of the validity such as concurrent, predictive and content validity create much more valid for the tests (Messick, 1995) . Furthermore, using a combination of qualitative and quantitative data can improve a better evaluation in which the limitations of one type of data will be balanced by the strengths of another (Bernstein, Moore & Bounchan, 2010). For these reasons, this study employs both qualitative and quantitative approach methods in order to investigate the validity of AEPET.

3.2.2.5.1 Pearson's Correlation

According to Garrett and Woodworth (1958), “correlation method is used to examine the relationship of one variable to another” (p. 122) and that ‘correlation is simply a measure of

mutual association between two variables’ (p. 180). Similarly, Glass and Hopkins (1984) assert that correlation is a measure to describe the degree of relationship between two variables, providing an example to illustrate the meaning of correlation in which “students’ scores might be paired with their English scores to determine whether students who earn high scores in one subject also earn them in another” (p. 79). Furthermore, several researchers used Pearson Correlation for investigating the concurrent and predictive validity of their language proficiency tests (Fulcher, 1997a; Al Hajr, 2014; Mauritoshi, 2002; Nakamura, 2006; Riazi, 2013). The researchers examined the relationship between their in- house English language proficiency test with a standardized international English test such as IELTS, TOEIC and TOEF or with student’s academic achievement measured by their CGPA. As all these references indicated above, Pearson Correlation was used in this study to answer the research question which seeks possible relationship between the students’ AEPET scores with IELTS scores and CGPA. In Pearson’s correlation, the strength of the relationship is indicated by the correlation coefficient **r** value. The larger **r** value is, the stronger the relationship between two variables is found. The significance the relationship is expressed by **p**- value. The smaller p value is, the more significant the relationship between two variables is found. Furthermore, correlation is an effect size and so we can verbally describe the strength of the correlation using the guide Evans (1996) suggest for the absolute value of Pearson’ r. Table 3.16 displays the strength of correlation coefficients.

Table 3.16: Strength of Correlation Coefficients

R- value	Strength of correlations
.00-.19	Very weak
.20-.39	Weak
.40-.59	Moderate
.60-.79	Strong
.80-1.0	Very strong

(Evans, 1996)

Alderson, Wall and Clapham (1996) suggest that a correlation coefficient of $r=.30$ which signifies a positive relationship between variables would be satisfactory as a standard sufficient for test validation studies. Therefore, several researchers have used the correlation coefficient ($r=.030$) as the standard strength of relationship between two variables (Criper & Davies, 1988; Davies, 1999; D’Este, 2012; Graham, 1987; Hill, Storch & Lynch, 1999; Huong, 2001; Moritoshi, 2002; Willson, 2016).

P- value gives evidence of statistical significance as well as indicates the strength of the relationship between two variables. Normally P- value ranges from 0- 1, a small p- value ($p<0.05$) indicates strong evidence against the null hypothesis. A large p- value ($p>0.05$) indicates weak evidence against the null hypothesis. Table 3.17 presents the strength of significant level (P – value).

Table 3.17: Strength of Significance

P- value	Strength of significance
$P > 0.10$	No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
$0.05 < P < 0.10$	Weak evidence against the null hypothesis in favor of the alternative.
$0.01 < P < 0.05$	Moderate evidence against the null hypothesis in favor of the alternative.
$0.001 < P < 0.01$	Strong evidence against the null hypothesis in favor of the alternative.
$P < 0.001$	Very strong evidence against the null hypothesis in favor of the alternative

3.2.2.5.2 Linear Regression

In statistics, regression analysis refers to “a broad class of statistical techniques that are designed to study the relationship between a criterion (or dependent variable), Y, and one or more predictors (or independent variables), $X_1, X_2, \dots, X$ ” (Tatsuoka, 1997, p. 648). In the same vein, regression analysis is considered as one of the most commonly used methods in prediction studies of academic success (McKenzie & Schweitzer, 2001; Othman & Nordin,

2013; Sar, 2008). Indeed, many researchers used regression analysis to conduct the relationship of their in-house English language proficiency tests with academic performance measured by CGPA (Fakeye, 2014; Feast, 2001; Karakaya & Tavsancil, 2008; Kucuk, 2001; Othman & Nordin, 2013; Sar, 2008). In the light of the references above, in this confirmatory stage, the key independent variables were Listening, Reading, Writing , Speaking and Overall scores of the AEPET while students' academic performance measured by CGPA was used as a dependent variable in order to determine the impact of AEPET scores on CGPA.

3.3 Chapter Review

This chapter presented some related areas where the data for this study were collected. Then the chapter also explained on two stages of Methodology in this study. Stage One involved the collection of materials, construction and procedure of testing instrument. Stage Two presented the Methodology of the Main Study.

The next chapter, i.e., Chapter 4 presents the results chapter for the Main Study, and reports the quantitative results on Concurrent and Predictive Validity and both quantitative and qualitative results on Content Validity of the AEPET.

CHAPTER 4

RESULTS AND FINDINGS

4.0 Chapter Overview

This chapter which is the results chapter for the Main Study consists of three parts. Result Part I reports the quantitative results on the Concurrent Validity of the AEPET. Result Part II reports the quantitative results on the Predictive Validity of the AEPET, in comparison with the Predictive Validity of the IELTS. Result Part III reports both quantitative and qualitative results on the Content Validity of the AEPET. The chapter ends with a summary of the findings.

4.1 Result Part I: Concurrent Validity of the AEPET

4.1.1. Data for Concurrent Validity

The data for concurrent validity was obtained from all 103 students who were doing the AEP. The students score on the AEPET and their IELTS scores were obtained from the University database by writing officially and requesting for the data. Table 4.1 presents the demographics of the student participants.

Table 4.1: Demographics of Student Participants (N=103)

Group	Frequency	Percent (%)
Male	40	38.8
Female	63	61.2
Total	103	100

As shown in Table 4.1, 61.2% were females and 38.8% of them were males, aged between 18 and 22. This gender imbalance is a common phenomenon in higher institutions not only in Vietnam but also in India, Malaysia, Thailand, China, Canada, Australia, the United States and many other countries in all over the world (Khan, 2007; Loan, 2002; Zacharia, 2014).

4.1.2 Band Score Distribution

Both AEPET and IELTS measure the same constructs: Listening, Reading, Writing, Speaking and Overall band scores. However, the range scores for each test are slightly different. The AEPET has 10 band scores, ranging from Band 1 to Band 10. Unlike AEPET, the IELTS has 9 band scores, ranging from Band 1 to Band 9. This means that the range of AEPET differs only in two band scores (Band 9.5 and Band 10) from the range of IELTS. Table 4.2 presents the students' band score distributions for AEPET and IELTS.

Table 4.2: Band Score Distributions for AEPET and IELTS (N=103)

	Listening		Reading		Writing		Speaking		Overall	
Band score	AEPET (%)	IELTS (%)	AEPET (%)	IELTS (%)	AEPET (%)	IELTS (%)	AEPET (%)	IELTS (%)	AEPET (%)	IELTS (%)
1.0	0	0	0	0	0	0	0	0	0	0
1.5	0	0	0	0	0	0	0	0	0	0
2.0	1	0	0	0	0	0	0	1.9	0	0
2.5	0	1	0	0	0	0	0	0	0	0
3.0	10.7	3.9	1.9	3.9	3.9	1.9	2.9	5.8	0	0
3.5	2.8	1	0	1	2.9	1	1	6.8	0	1
4.0	4.9	10.7	2.9	5.7	6.8	10.7	10.7	4.9	1.9	1
4.5	0	6.8	0	8.7	1.9	9.7	1	3.9	3.9	3.9
5.0	14.6	19.4	6.8	17.5	7.8	23.3	16.4	4.9	9.7	23.3
5.5	8.7	14.6	9.7	9.7	6.8	17.5	10.7	9.7	16.5	29.1
6.0	12.6	19.4	14.6	28.2	16.5	17.5	20.4	13.5	20.4	26.2
6.5	17.5	14.6	19.4	10.7	18.4	10.7	4.9	28.2	27.1	14.5
7.0	10.7	8.6	24.3	10.7	20.4	6.8	14.6	12.6	11.7	1
7.5	3.9	0	6.8	3.9	7.8	1	3.9	6.8	6.8	0
8.0	11.6	0	9.7	0	5.8	0	10.7	1	1	0
8.5	1	0	2.9	0	1	0	1.9	0	1	0
9.0	0	0	1	0	0	0	1	0	0	0
9.5	0	-	0	-	0	-	0	-	0	-
10.0	0	-	0	-	0	-	0	-	0	-

4.1.2.1 AEPET Listening Component Score

The band score distribution for Listening Component of the two tests is presented graphically in Figure 4.1

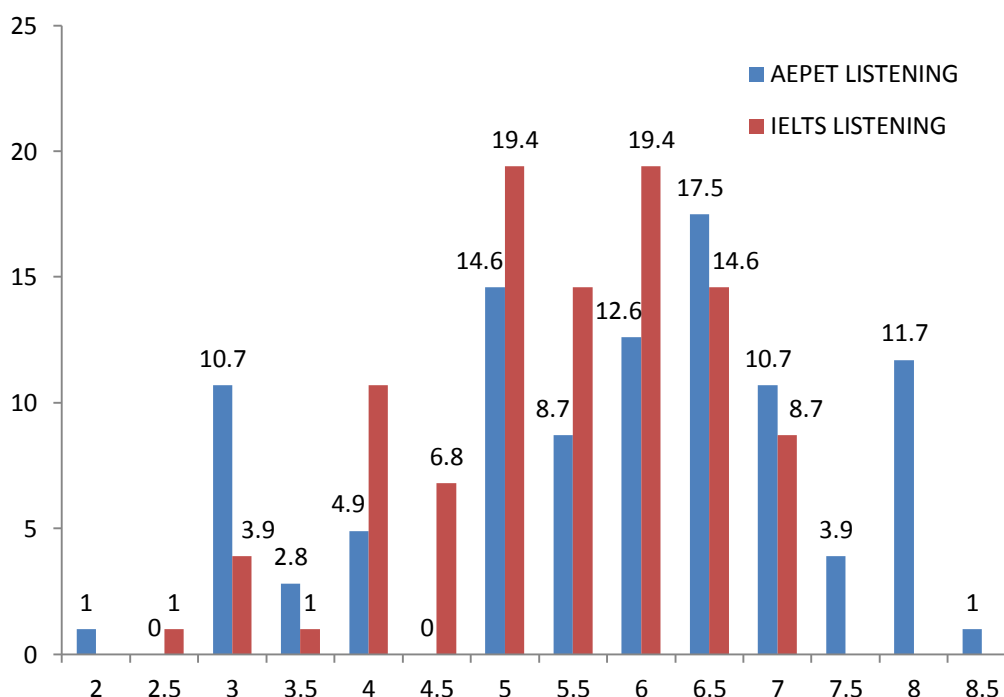


Figure 4.1: Distribution for Listening Component Score

Figure 4.1 shows that the scores of the AEPET Listening Component range from Band 2.0 to Band 8.5, which are classified from below average to very good ability in English proficiency. More specifically, the highest frequency for the bands is 17.5% for Band 6.5, followed by 14.6% for Band 5 and 12.6% for Band 6. The percentages for the rest of AEPET Listening band scores range from 1% to 11.7%.

The scores of IELTS Listening component range from Band 2.5 to Band 7.5, which are classified from intermittent to good ability in English proficiency. 19.4% of the students scored for Band 5.0 and Band 6.0, followed by Band 5.5 and Band 6.5 with 14.6%. The percentages for the rest of IELTS Listening band scores range from 1% to 10.6%.

4.1.2.2 AEPET Reading Component Score

The band score distribution for Reading Component of the two tests is presented graphically in Figure 4.2

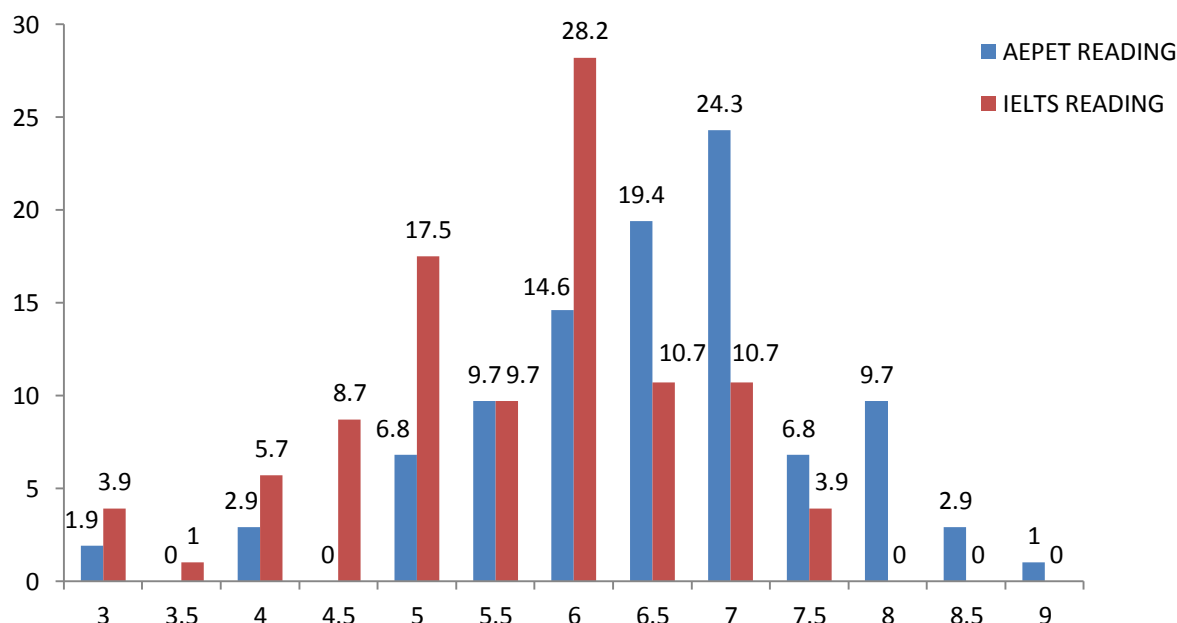


Figure 4.2: Distribution for Reading Component Score

Figure 4.2 shows that the scores of the AEPET Reading Component range from Band 3.0 to Band 9.0, which are classified from below average to very good ability in English proficiency. In specific, the highest percentage for the bands is for Band 7.0 with 24.3%, and then it is followed by Band 6.5 (19.4%) and Band 6.0 (14.6%). The percentages for the rest of AEPET Reading band scores range from 1% to 10.7%.

The scores of IELTS Reading Component range from Band 3.0 to Band 7.5, which are classified from very limited to good ability in English ability. Particularly, the highest percentage for the bands is for Band 6.0 with 28.2%, followed by Band 5.0 (17.5%), Band 6.5 and Band 7.0 (10.7%). The percentages for the rest of IELTS Reading band scores range from 1% to 9.7%.

4.1.2.3 AEPET Writing Component Score

The band score distribution for Writing Component of the two tests is presented graphically in Figure

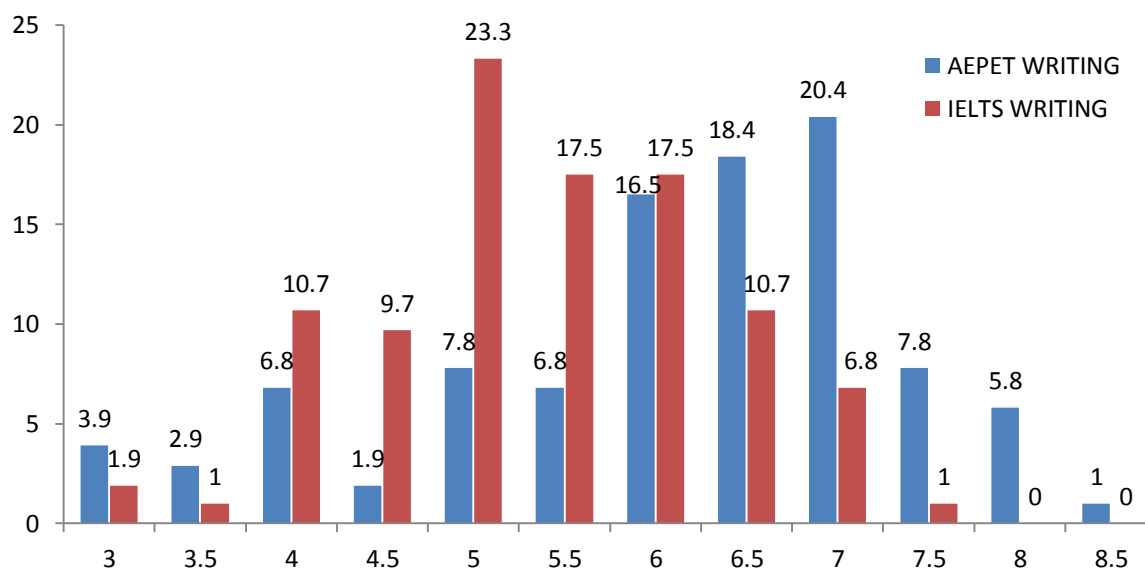


Figure 4.3: Distribution for Writing Component Score

Figure 4.3 shows that the scores of the AEPET Writing Component range from Band 3.0 to Band 8.5, which are classified from below average to very good ability in English proficiency. More precisely, the highest percentage for the bands is for Band 7.0 with 20.4%, followed closely by Band 6.5 with 18.4% and 16.5% for Band 6.0. The percentages for the rest of AEPET Writing band scores range from 1% to 7.8%.

The scores of IELTS Writing Component range from Band 3.0 to Band 7.5, which are classified from very limited to good ability in English proficiency. In particular, the highest percentage for the bands is for Band 5.0 with 23.3%, and followed by Band 5.5 and Band 6.0 with the same percentage of 17.5%. The percentages for the rest of IELTS Writing band scores range from 1% to 10.7%.

4.1.2.4 AEPET Speaking Component Score

The band score distribution for Speaking Component of the two tests is presented graphically in Figure 4.4

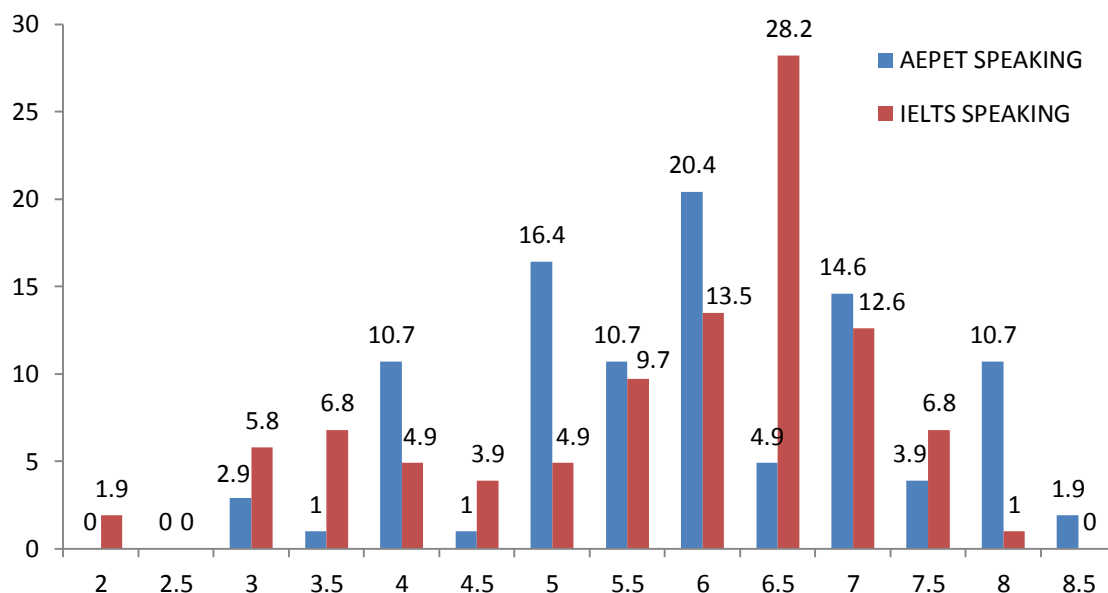


Figure 4.4: Distribution for Speaking Component Score

Figure 4.4 shows that the scores of the AEPET Speaking Component range from Band 3.0 to Band 9.0, which are classified from below average to excellent ability in English proficiency. Particularly, the students obtained the highest percentage for Band 6.0 with 20.6%, followed by Band 5.0 (16.4%) and Band 7.0 (14.6%). The percentages for the rest of AEPET Speaking band scores range from 1 % to 10.7%.

The scores of IELTS Speaking Component range from Band 2.0 to Band 8.0, which are classified from intermittent to very good ability in English proficiency. Particularly, most of the students obtained the highest percentage for Band 6.5 with 28.2%, followed by Band 6.0 with 13.5% and 12.6% for Band 7.0. The percentages for the rest of IELTS Speaking band scores range from 1 % to 9.7%.

4.1.2.5 AEPET Overall Band Score

The band score distribution for Overall band of the two tests is presented graphically in Figure 4.4

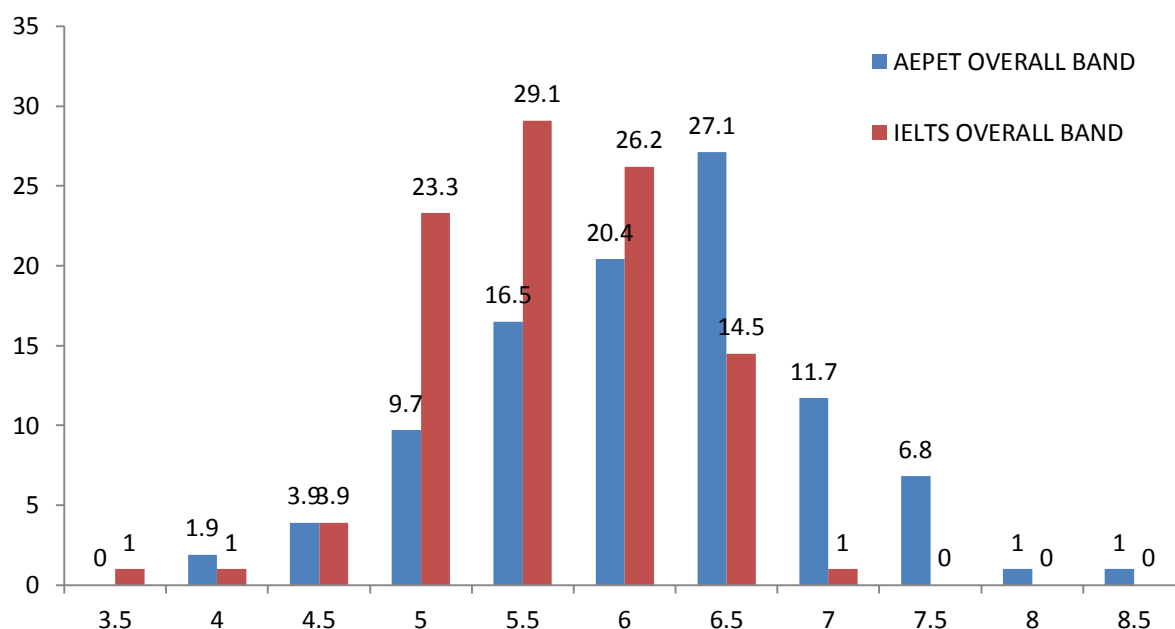


Figure 4.5: Distribution for Overall Band Score

Figure 4.5 shows that the scores of the AEPET Overall band range from Band 4.0 to Band 8.5, which are classified from average to very good ability in English language proficiency. More precisely, the students obtained the highest percentage for Band 6.5 with 27.2% , and then followed by Band 6.0 (20.4%) and Band 5.5 (16.5%). The percentages for the rest of AEPET overall band scores range from 1% to 11.7%.

The scores of IELTS Overall band range from Band 3.5 to Band 7.0, which are classified from limited to good ability in English language proficiency. Specifically, most of the students scored the highest percentage for Band 5.5 with 29.1%, followed by Band 6.0 (26.2%) and Band 5.0 (23.3%). The percentages for the rest of IELTS Overall band scores range from 1% to 14.5%.

4.1.2.6 Band Score Distribution according to Cut - Off Score

Based on the common cut-off score under the AEP programme for the selection of the students for further study in the United States, which is Band 6.5, for AEPET and IELTS (Refer to Chapter 3, Section 3.2.1.1), the range scores are classified into two groups Below Band 6.5 and Band 6.5 and Above . Table 4.3 shows the distribution for between two ranges of the two tests below Band 6.5 and Band 6.5 and above.

Table 4.3: Distribution for AEPET and IELTS Cut - Off Band Scores

Component	Below Band 6.5		Band 6.5 and Above	
	AEPET (%)	IELTS (%)	AEPET (%)	IELTS (%)
Listening	55.3	76.8	44.7	23.2
Reading	35.9	74.7	64.1	25.3
Writing	46.6	81.6	53.4	18.4
Speaking	62.1	51.4	37.9	48.6
Overall	52.4	84.5	47.6	15.5

Table 4.3 is presented graphically in Figure 4.6

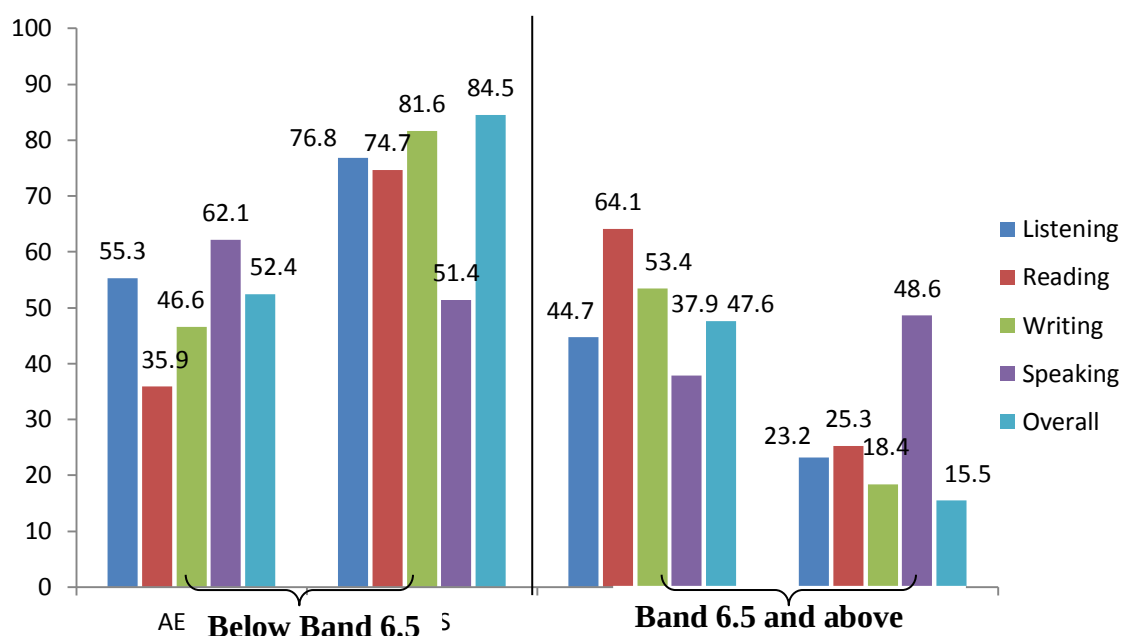


Figure 4.6: Distributions for AEPET and IELTS Cut- off Band Scores

Figure 4.6 shows that the students scored Band 6.5 and above on AEPET components and overall band are higher than those who obtained on the same sections of IELTS. On AEPET, most of the students scored Band 6.5 and above which ranges from 37.9% to 64.1%. In contrast, on IELTS, a lower percentage of the students scored Band 6.5 and above which ranges from 15.5% to 48.6%.

For AEPET, the highest percentage for score Band 6.5 is for Reading with 64.1%, followed by Writing (53.4%), Overall (47.6%), Listening (44.7%) and Speaking (37.9%). For IELTS, the highest percentage is for Speaking with 48.6%, followed by Reading (25.3%), Listening (23.3%), Writing (18.4%) and Overall (15.5%).

From the results for concurrent validity, it can be observed that the students perform better on four components: Listening, Reading, Writing and Overall scores in the AEPET than in the IELTS. However, the students perform better on Speaking component in the IELTS than the AEPET.

4.1.3 Correlation between AEPET and IELTS Score

In attempting to establish the concurrent validity for AEPET, the Pearson's correlation was run between the four components: Listening, Reading, Writing, Speaking and Overall scores on AEPET and IELTS.

Although the range of AEPET is two band score (Band 9.5 and Band 10) larger than the range of IELTS, the correlations between two test scores can be analyzed using Pearson Correlation and Linear Regression because both has the same interval which is .05. It is due to the fact that Pearson's correlation coefficient was designed to determine the strength and the significance of the association between two variables which are measured on an interval scale,

however, both variables do not need to be measured on the same scale or same units (Pagano, 2004; Treiblmaier & Filzmoser, 2009).

In Pearson's correlation, the strength of the relationship is indicated by the correlation coefficient **r** value. The significance of the relationship is expressed by **p**- value. The details of the strength and the significance of the relationship between two variables are explained in Chapter 3, Section 3.2.3.5.1. Table 4.4 shows the relationship between AEPET and IELTS components: Listening, Reading, Writing, Speaking and Overall scores.

Table 4.4: Correlations Results between AEPET and IELTS Scores

	IELTS Listening	IELTS Reading	IELTS Writing	IELTS Speaking	IELTS Overall
AEPE Listening	.261** (p= .008)				
AEPE Reading		.351** (p= .000)			
AEPE Writing			.307** (p= .002)		
AEPE Speaking				.517** (p= .000)	
AEPE Overall					.398** (p= .000)

****Correlation is significant at the 0.01 level (2-tailed)**

The results presented in Table 4.4 show that the relationship between AEPET and IELTS is significant. The p- value is less than 0.01 level, showing very high significant correlation between each component: Listening, Reading, Writing, Speaking and Overall band scores between the two tests. In addition, the r value which ranges from .261 to .517 is positive, indicating from weak to moderate correlations. More specifically, the highest correlation is found for Speaking Component (r=.517) which is a moderate correlation, followed by Overall (r= 0.398), Reading component (r=0.351), Writing (r= 0.307) and Listening (r=.0.261) which are weaker correlations.

4.1.4 Summary of Results on the Concurrent Validity of the AEPET

Table 4.5 summarizes the p values, r values between the two tests: AEPET and IELTS

Table 4.5: Summary of Concurrent Validity of the AEPET

Component	AEPET vs. IELTS	
	P - value	R - value
Speaking	.000	Moderate ($r=.517$)
Overall	.000	Weak ($r=.398$)
Reading	.000	Weak ($r=.351$)
Writing	.002	Weak ($r=.307$)
Listening	.008	Weak ($r=.261$)

Notes

	:Significant at 0.01 level		: Very Strong correlation ($r=.80-1.0$)
			: Strong correlation ($r=.60-.79$)
			: Moderate correlation ($r=.40 - .59$)
			: Weak correlation ($r=.20-.39$)
			: Very weak correlation ($r=.00-.19$)

Table 4.5 is presented graphically in Figure 4.7 which presents path loadings which are derived from the relationship between AEPET and IELTS

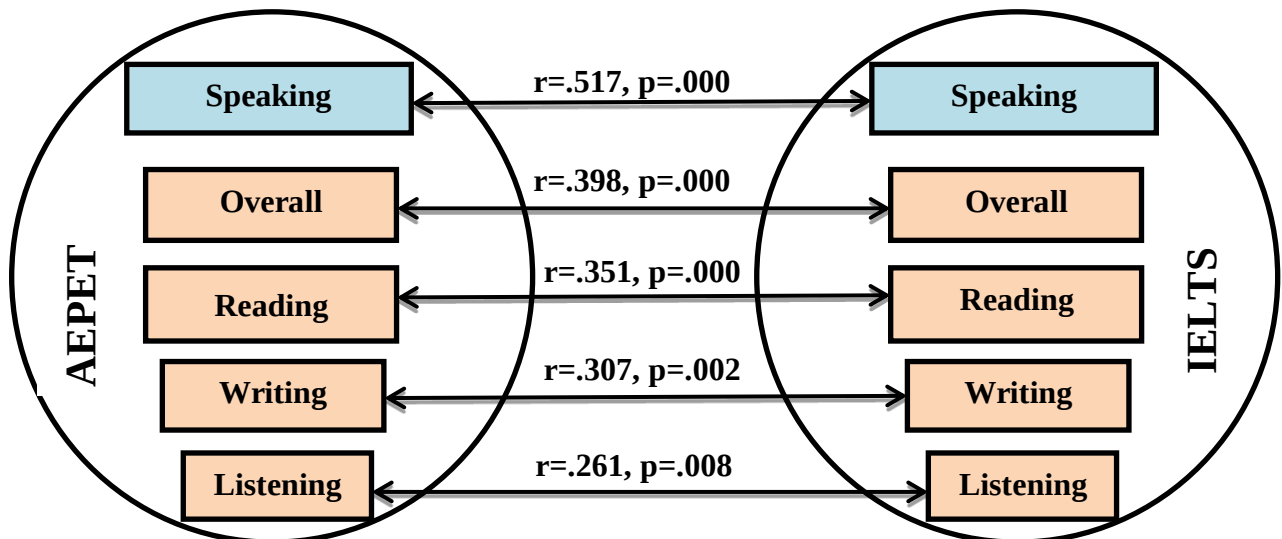


Figure 4.7: Path Loadings for Correlation Results between AEPET and IELTS

Figure 4.7 indicates that two main sources of evidence support the concurrent validity of AEPET. The first source is positive and significant r - values between each component and overall band score of the AEPET and IELTS. The second source is the significant p -values between each component and overall band score of the two tests. In other words, there are significant correlations between AEPET Listening, Reading, Writing, Speaking scores and IELTS Listening, Reading, Writing, Speaking scores. Similarly, the overall band score for AEPET and IELTS also significantly correlates. From the results, it can be concluded that the AEPET score significantly correlates with the IELTS score. The strength of correlations ranges from weak to moderate. Therefore, it is suggested that students who achieve higher scores on Listening, Reading, Writing, Speaking components and overall band scores of the AEPET have the tendency to gain similar scores on Listening, Reading, Writing, Speaking sections and overall band scores on IELTS.

4.2 Result Part II: Predictive Validity of the AEPET

4.2.1 Data for Predictive Validity

Just like the data for concurrent validity, the data for predictive validity were obtained from the University's database. The scores of a total of 143 students who have taken the AEPET, IELTS and completed their study from the public university were given to the researcher. Of the 143 students, 60.1% were female and 39.9% were male. The students' AEPET scores, IELTS scores and their CGPA were keyed into SPSS for analysis. Table 4.6 shows the composition of the participants.

Table 4.6: Composition of the Participants

	Frequency	Percent (%)
Female	86	60.1
Male	57	39.9
Total	143	100

4.2.2 Correlation Results

In attempt to establish the predictive validity of the AEPET and IELTS (N=143), data collected were analyzed through two stages. This stage used Pearson Correlation analysis to examine the degree of relationship between AEPET scores and CGPA, IELTS scores and CGPA. The correlation coefficient r value and significant p value between AEPET scores and CGPA, IELTS scores and CGPA were calculated. Table 4.7 presents the correlations between the AEPET and CGPA; IELTS and CGPA (N=143).

Table 4.7: Correlations Results between AEPET scores and CGPA

AEPET vs CGPA (N=143)			IELTS vs CGPA (N=143)		
Component	P	R	Component	P	R
Overall	.000	.613	Overall	.000	.614
Speaking	.000	.549	Speaking	.000	.535
Reading	.000	.451	Listening	.000	.471
Listening	.000	.414	Reading	.000	.428
Writing	.000	.366	Writing	.000	.314

:Significant at 0.01 level

	: Very Strong correlation (.80-1.0)
	: Strong correlation ($r=.60-.79$)
	: Moderate correlation ($r=.40-.59$)
	: Weak correlation ($r=.20-.39$)
	: Very weak correlation ($r=.00-.19$)

Table 4.7 shows that, the relationship between AEPET scores and CGPA is significant. The p value is less than 0.01, showing very high significant level between each component, overall scores of the AEPET and CGPA. Furthermore, the r values which range from .366 to .614 are all positive, indicating from weak to strong correlations. Particularly, the highest r value is observed for Overall AEPET score and CGPA ($r=.613$, $p=.000$), showing strong correlation. Then, it is followed by the relationship between AEPET Speaking scores and CGPA ($r=.549$, $p=.000$), Reading ($r=.451$, $p=.000$), Listening ($r=.414$, $p=.000$), showing moderate strength

of correlations. However, lower r value is found for AEPET Writing scores and CGPA ($r=.366$, $p=.000$), corresponding to the weak strength of correlation.

Just like the relationship between AEPET and CGPA, the relationship between IELTS scores and CGPA is also significant. The p value is less than 0.01, showing very high significance level between each component, Overall band scores of the IELTS and CGPA. Additionally, the r values, which range from .314 to .614, are all positive, corresponding from weak to strong. More precisely, the highest r value is for IELTS Overall scores and CGPA ($r=.546$, $p=.000$), show the strong strength of correlation. Then, it is followed by the relationship between IELTS Speaking scores and CGPA ($r=.535$, $p=.000$), IELTS Listening scores ($r=.471$, $p=.000$) and IELTS Reading scores ($r=.428$, $p=.000$), indicating the moderation strength of correlations. In contrast, the lower r - values is observed for IELTS Writing scores and CGPA ($r=.314$, $p=.000$), signifying weak strength of correlation.

4.2.3 Regression Results

Linear Regression analysis was used to determine the impact of the students' AEPET and IELTS scores on their academic CGPA as well as to determine which of AEPET, IELTS components and Overall band score prove to be the best predictor of academic achievement. The key independent variables were AEPET, IELTS components: Listening, Reading, Writing, Speaking and Overall band scores while students' CGPA scores were used as a dependent variable. Table 4.8 shows the regression results between AEPET vs. CGPA and IELTS vs. CGPA ($N=143$).

Table 4.8: Regression Results between AEPET vs. CGPA and IELTS vs. CGPA.

Component	AEPET vs. CGPA					IELTS vs. CGPA				
	R ²	β	t	F	P	R ²	β	t	F	P
Overall	.376	.613	9.212	84.868	.000**	.376	.614	9.228	85.110	.000**
Speaking	.301	.549	7.795	60.770	.000**	.286	.535	7.513	56.446	.000**
Reading	.171	.414	5.402	29.178	.000**	.183	.428	5.622	31.602	.000**
Listening	.204	.451	6.006	36.073	.000**	.225	.474	6.390	40.828	.000**
Writing	.134	.366	4.674	21.857	.000**	.099	.314	3.932	15.461	.000**

Notes

** : Significant at the 0.01 level (2-tailed)

Predictors : AEPET Listening, Reading, Writing, Speaking and Overall scores

Dependent variable : CGPA

Table 4.8 shows that there is significant correlation between the AEPET scores and CGPA because the p value is less than 0.01, showing a strong regression line between each component, overall band scores of the AEPET scores and CGPA. Furthermore, the r squared values are all positive, ranging from .134 to .376. The highest r-squared is observed for the agreement between AEPET Overall scores and CGPA ($R^2 = .376$, $\beta = .613$, $t(143) = 7.372$, $F(1,141) = 84.868$, $p = .000$), and then followed by AEPET Speaking scores ($R^2 = .301$, $\beta = .549$, $t(143) = 7.759$, $F(1,141) = 60.770$, $p = .000$), indicating 37.6% and 30.1% respectively of the variance of success in CGPA. By contrast, lower r squared values are found for the relationships between CGPA and AEPET Listening scores ($R^2 = .204$, $\beta = .451$, $t(143) = 6.006$, $F(1,141) = 36.073$, $p = .000$); AEPET Reading scores ($R^2 = .171$, $\beta = .414$, $t(143) = 5.402$, $F(1,141) = 29.178$, $p = .000$); AEPET Writing scores ($R^2 = .134$, $\beta = .366$, $t(143) = 4.674$, $F(1,141) = 21.857$, $p = .000$), corresponding to 20.4%, 17.1% and 13.4% respectively of the variance of academic success.

Just like the relationship between AEPET and CGPA, the IELTS scores significantly correlates with CGPA due the p value being less than 0.01, showing a strong regression line

between each component, overall band scores of the IELTS and CGPA. Furthermore, the r-squared values are all positive, but small, ranging from .099 to 37.6. The highest r-squared value is accounted for the agreement between CGPA and IELTS Overall scores ($R^2=.376$, $\beta =.614$, $t(143) = 9.228$, $F(1,141) = 85.110$, $p= .000$), and then followed by Speaking ($R^2=.286$, $\beta =.535$, $t(143) = 7.513$, $F(1,141) = 56.446$, $p= .000$), predicting 37.8% and 28.6% respectively of the variance of success in CGPA. In contrast, the lower r-squared are observed for the associations between CGPA and IELTS Listening scores ($R^2=.225$, $\beta =.474$, $t(143) = 6.390$, $F(1,141) = 40.828$, $p= .000$); IELTS Reading scores ($R^2=.183$, $\beta =.428$, $t(143) = 5.622$, $F(1,141) = 31.602$, $p=.000$); IELTS Writing scores ($R^2=.099$, $\beta =.314$, $t(143) = 3.932$, $F(1,141) = 15.461$, $p= .000$), signifying 22.5%, 18.3% and 9.9% respectively of the variance of success in CGPA.

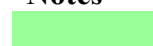
4.2.4 Summary of Predictive Validity results for the AEPET and IELTS

Table 4.9 summarizes the predictive validity results for AEPET and IELTS.

Table 4.9: Summary of Predictive Validity of the AEPET and IELTS

AEPET vs. CGPA			IELTS vs. CGPA		
Component	Significance	Correlation	Component	Significance	Correlation
Overall	Sig	S ($r=.613$)	Overall	Sig	S ($r=.614$)
Speaking	Sig	M ($r=.549$)	Speaking	Sig	M ($r=.535$)
Reading	Sig	M ($r=.451$)	Listening	Sig	M ($r=.471$)
Listening	Sig	M ($r=.414$)	Reading	Sig	M ($r=.428$)
Writing	Sig	W ($r=.366$)	Writing	Sig	W ($r=.314$)

Notes

 :Significant at 0.01 level

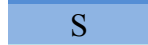
 VS : Very Strong correlation (.80-1.0)
 S : Strong correlation ($r=.60-.79$)
 M : Moderate correlation ($r=.40 - .59$)
 W : Weak correlation ($r=.20-.39$)
 VW : Very weak correlation ($r=.00-.19$)

Table 4.9 shows that both AEPET and IELTS components significantly correlate with CGPA. The p value is less than 0.01, showing a very high significant level between each component: Overall, Speaking, Reading, Listening and Writing scores in the two tests and CGPA. In addition, the r- values in the two tests are all positive, indicating from weak to strong correlations. Across the two tests, only strong correlation is found for the relationship between AEPET Overall and CGPA; IELTS Overall scores and CGPA, thus overall scores of the two tests emerge as the most significant predictors for academic success. Likewise, the moderate correlations are observed for the association between Speaking, Reading and Listening scores with CGPA in the two tests. However, weak correlation is accounted for the relationship between Writing and CGPA.

In a nut shell, just like IELTS, the AEPET is able to predict students' academic achievement. The more students get higher AEPET scores, especially on AEPET Overall, Speaking, Reading and Listening scores, the better CGPA scores they can achieve.

4.3 Result Part III: Content Validity of the AEPET

4.3.1 Participants (N=20)

The questionnaire was administrated with 20 lecturers who were teaching English in the AEP. Background information about the English lecturers were collected using questionnaire (refer to Appendix 3, section A). Table 4.10 below presents the demographics of the lecturer participants

Table 4.10: Demographics of English Lecturer Participants (N=20)

Item	Group	Frequency	Percent
Gender	Male	4	20
	Female	16	80
Academic qualification	Master Degree	9	45
	Ph.D.	11	55
Year of teaching English	Less than 5 years	1	5
	5-10 years	8	40
	11 to 15 years	5	25
	More than 15 years	6	30
Teaching English subjects	Listening	2	10
	Reading	3	15
	Writing	4	20
	Speaking	4	20
	Listening and Reading	3	15
	Writing and Speaking	3	15
	Reading and Writing	1	5
Attended training in language testing	Yes	7	35
	No	13	65
Training course on English components	All four components	7	35
	Did not attend	13	65

As shown in Table 4.10, the participants were 20 English lecturers consisting of 4 (20%) males and 16 (80%) females. Their age ranges from 25 to 50. 55% of English language lecturers hold Ph.D. degrees and 45% possess Master degrees.

In relation to teaching experience, 40% of the lecturers have been teaching English at the university for 5 to 10 years, 30% have been teaching more than 15 years and 25% from 11 to 15 years. By contrast, only 5% have less than 5 years teaching experience.

In relation to English subjects that the lecturers teach at the university, some lecturers teach only one English component such as Speaking (20%), Writing (20%), Reading (15%) and Listening (10%). However, some of them teach two English components: Listening and Reading (20%); Writing and Speaking (20%); Reading and Writing (5%).

Referring to training in language testing, 35% of the lecturers have attended the training course for all English components: Listening, Reading, Writing and Speaking. In contrast, 65% said they did not attend any training course on language testing.

4.3.2 Reliability of Content Validity Questionnaire

A total of 53 items relating to content validity of the AEPET were included in the questionnaire (refer to Section B and C in Appendix 3). Of the 53 items, Section B (Q11-Q38) was on test components and Section C was (Q39-Q63) on test preparations. Details of the questionnaire are provided in Chapter 3, Section 3.2.1.3). After administering the Content Validity Questionnaire, the data obtained were coded and run through SPSS to check the reliability of the items in the questionnaire. The most popular statistical index that is used to measure the internal consistency for the questionnaire items is the alpha value which was developed by Cronbach (1951) (cited in Wells & Wollack, 2003).

Several researchers (George & Mallery, 2003; Gliem & Gliem, 2003;Tavakol & Dennick, 2011) recommend that the benchmark Cronbach's alpha value should be at least 0.7. Cronbach's alpha value satisfying 0.7 shows acceptable reliability. If the Cronbach's alpha is below 0.7, the evidence proves that the items do not have much in common and are unable to measure the same construct. Table 4.11 presents the rules of alpha values suggested by George and Mallery (2003), Gliem and Gliem (2003) and Tavakol and Dennick (2011).

Table 4.11: Rules of Alpha Values

Values	Levels
$\alpha \geq .90$	Excellent
$\alpha \geq .80$	Good
$\alpha \geq .70$	Acceptable
$\alpha \geq .60$	Questionable
$\alpha \geq .50$	Poor
$\alpha \leq .50$	Unacceptable

(George & Mallery, 2003; Gliem & Gliem, 2003;Tavakol & Dennick, 2011)

For Content Validity Questionnaire (refer to Appendix 3.), Cronbach's Alpha values at Likert- scale levels were computed to investigate the reliability of the questionnaire. Table 4.12 shows alpha value of the items in each section in the questionnaire.

Table 4.12: Cronbach's Alpha Values for Lecturers' Questionnaire

	Section	No of Items	Question	Cronbach's Alpha	Levels
Test Component	Listening Test	7	Q11-Q17	.881	Good
	Reading Test	7	Q18-Q24	.762	Acceptable
	Writing Test	7	Q25-Q31	.717	Acceptable
	Speaking Test	7	Q32-Q38	.840	Good
Test Preparation		25	Q39-Q63	.894	Good
Overall		53	Q1-Q63	.854	Good

Table 4.12 shows that the alpha values for all sections in the questionnaire are above 0.70. In particular, for the four English components, alpha values reach the good level for Listening Test ($\alpha = 0.881$), and Speaking Test ($\alpha = 0.840$). In contrast, the values correspond to the acceptable degree for Reading Test ($\alpha = 0.762$) and Writing Test ($\alpha = 0.717$). For test preparation, the alpha value is good level ($\alpha = 0.894$). Similarly, the overall alpha value of the Content Validity Questionnaire is good ($\alpha = 0.854$). Thus, it can be inferred that the items in content validity questionnaire are sufficiently reliable.

4.3.3 Content Validity Results Interpretation

Content validity plays the most important role in test validation because this validity component reflects the match between test items and the objectives and content knowledge in the course syllabus (Siddiek, 2010). Therefore, in order to establish the content validity results obtained from the lecturers' questionnaire, the analysis was done by computing the mean scores for each item in each section of the questionnaire. Table 4.13 presents the

interpretation of the mean scores recommended by Abidin (2006), Kucuk (2007) and Rumsey (2013).

Table 4.13: Interpretation of the Mean Scores

Mean	Option	Interpretation
4.5 - 5.0	Strong Agreement	Very high
3.5 - 4.4	Agreement	High
2.5 - 3.4	Neutral	Moderate
1.5 - 2.4	Disagreement	Low
1.0 - 1.4	Strong Disagreement	Very Low

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

(Abidin, 2006 ; Kucuk, 2007; Rumsey, 2013)

4.3.3.1 Content Validity Results for AEPET Components

This section presents the results interpretation for the content validity of the four AEPET components which consist of four components: Listening, Reading, Writing and Speaking from the questionnaires which were administered with the lecturers.

4.3.3.1.1AEPET Listening Component

The first component of AEPET is Listening Test in which students are required to answer 50 questions within 30 minutes. The Listening component carries 20% of the AEPET. Table 4.14 presents the mean score results fromQ11 to Q17.

Table 4.14: Mean Scores for AEPET Listening Component

No	Items	Mean	SD	Degree
11	Weightage	2.80	.95145	M
12	Time allocation	2.95	.88704	M
13	Representation of listening skills	3.50	.94591	H
14	Representation of topics	3.20	.95145	M
15	Clarity of questions	3.40	1.14248	M
16	Clarity of instructions	3.55	.99868	H
17	Mark allocations	3.70	1.08094	H
Overall		3.30	.70362	M

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.14 shows that the mean scores for all of items range from 2.80 to 3.70. In specific, among the items, the highest mean score is for sufficiency of mark allocations (Q17, M=3.70), followed by clarity of instructions (Q16, M=3.55) and representation of listening skills (Q13, M=3.50), indicating the group of high level of agreement. This indicates that the lecturers highly agree on AEPET Listening Component's mark allocation, clarity of instructions and representation of listening skills. However, the lower mean scores are observed for weightage (Q11, M=2.80), time allocation (Q12, M=2.95) and clarity of questions (Q15, M=3.40), falling in the category of moderate level of agreement. This implies that the lecturers only moderately agree on AEPET Listening Component's appropriacy of the weightage, sufficiency of time allocation, representation of topics, and clarity of questions.

In summary, the overall mean score which is **3.30 implies** that AEPET Listening Component on the whole seems to obtain moderate content validity from the lecturers' judgments.

4.3.3.1.2 AEPET Reading Component

The second component of AEPET is Reading Test in which students are required to complete 50 questions within 60 minutes. The reading component carries 20% in the AEPET. Table 4.15 shows the descriptive statistic results for the reading component from Q18 to Q24.

Table 4.15: Mean scores for AEPET Reading Component

No	Items	Mean	SD	Degree
18	Weightage	3.00	.97333	M
19	Time allocation	3.25	1.01955	M
20	Representation of reading skills	3.50	.88852	H
21	Representation of reading topics	3.15	.98809	M
22	Clarity of questions	3.75	.71636	H
23	Clarity of instructions	4.00	.56195	H
24	Mark allocations	3.95	.60481	H
Overall		3.51	.82265	H

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

As shown in Table 4.15, the mean scores for all of items range from 3.00 to 4.00. More precisely, among the items, the highest mean score is for clarity of instructions(Q23, M=4.00), followed by mark allocations (Q24, M=3.95), clarity of questions (Q22, M=3.75) and representation of reading skills (Q20, M=3.50), showing the category of high agreement degree. This means that the lecturers highly agree on AEPET Reading Component's clarity of instructions, mark allocation, clarity of question and representation of reading skills.

On the other hand, the lower mean scores are for time allocation (Q19, M=3.25), representation of reading topics (Q21, M=3.15) and weightage (Q18, M=3.00), falling in the group of moderate agreement level. This proves that the lecturers adequately agree on AEPET Reading Component's sufficiency of time allocation, representation of topics and appropriacy of weightage.

The overall mean score for all items in this component is 3.51. This shows that AEPET Reading Component seems to possess high content validity from the lecturers' judgments.

4.3.3.1.3 AEPET Writing Component

The third component of AEPET is Writing Test in which students are required to complete two given tasks within 60 minutes. The Writing Component carries 40% in the AEPET. Table 4.16 presents the mean score results for Writing component from Q25 to Q31.

Table 4.16: Mean Scores for AEPET Writing Component

No	Items	Mean	SD	Degree
25	Weightage	3.20	.95145	M
26	Time allocation	3.10	.91191	M
27	Representation of writing skills	3.65	.58714	H
28	Representation of writing topics	3.50	.68825	H
29	Clarity of questions	3.95	.22361	H
30	Clarity of instructions	4.10	.30779	H
31	Mark allocations	3.40	.94032	M
Overall mean		3.05	.66335	M

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.16 shows that the mean scores for all of items in the Writing component are above 3.0, ranging from 3.10 to 4.10. In particular, the highest mean score is for clarity of instructions (Q30, M=4.10), followed by clarity of questions (Q29, M=3.95), representation of writing skills (Q27, M=3.65) and representation of writing topics (Q28, M=3.50), falling in the group of high agreement degree. It means that the lecturers highly agree on AEPET Writing Component's clarity of instructions and questions, representation of writing skills and topics.

In contrast, the lower mean scores are accounted for mark allocation (Q31, M=3.40), weightage (Q25, M=3.20) and time allocation (Q26, M=3.10), sharing the group of moderate level of agreement. This suggests that the lecturers moderately agree on AEPET Writing Component's mark allocations appropriacy of weightage and sufficiency of time allocation.

4.3.3.1.4 AEPET Speaking Component

The last important AEPET component is Speaking Test in which students are required to answer 3 parts within 11-14 minutes. This component takes 20% in the AEPET. Table 4.17 presents the mean scores for Speaking Component from Q32 to Q38.

Table 4.17: Mean Scores for AEPET Speaking Component

No	Items	Mean	SD	Degree
32	Weightage	3.30	.923	M
33	Time allocation	3.05	.998	M
34	Representation of speaking skills	3.75	.850	H
35	Representation of speaking topics	3.85	.833	H
36	Clarity of questions	3.95	.604	H
37	Clarity of instructions	3.95	.759	H
38	Mark allocations	3.05	1.050	M
Overall mean		3.56	.860	H

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

As shown in Table 4.17, it can be found that the mean scores for all of items range from 3.05 to 3.95. Particularly, clarity of questions (Q36) and instructions (Q37) rank first with the same

highest mean score (M=3.95), followed by representation of speaking topics (Q35, M=3.85), and skills (Q34, M=3.75), sharing in the group of high agreement level. These results state that the lecturers highly agree on AEPET Speaking Component's clarity of questions and instructions, representation of speaking topics and skills.

On the other hand, the lower mean scores are found for appropriacy of weightage (Q32, M=3.30), sufficiency of time allocation (Q33, M=3.05) and appropriacy of mark allocation (Q38, M=3.05), having the category of moderate agreement level. This shows that the lecturers moderately agree on AEPET Speaking Component's appropriacy of weightage, sufficiency of time allocation and appropriacy of mark allocations.

Overall, the averaged mean score (M = 3.56) for all items is 3.56. This shows AEPET Speaking Component on the whole seems to achieve high content validity from the lecturers' judgments.

4.3.3.1.5 Overall Mean Scores for AEPET Components

Table 4.18 summarizes the content validity results between AEPET Components: Listening, Reading, Writing and Speaking.

Table 4.18: Overall Mean Scores for AEPET Components

Component	Overall mean	SD	Degree
Speaking Test	3.56	.860	H
Reading Test	3.51	.822	H
Listening Test	3.30	.703	M
Writing Test	3.05	.663	M
Overall mean	3.35	.762	M

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.18 shows that the content validity of the AEPET is not in parallel between components. Both Speaking and Reading Tests have high content validity while Listening and Writing Tests have moderate content validity. In other words, the course content is highly

represented in Speaking and Reading Tests meanwhile it is moderately represented in Listening and Writing Tests. The overall mean scores of four English components is 3.35 and this shows that the AEPET on the whole has moderate content validity (M=3.35).

4.3.3.2 Content Validity Results for AEPET Test Preparation

In order to better understand the results for AEPET Components, all the 25 questions which are in relation to the process of AEPET preparation were collected and then analyzed quantitatively and qualitatively. To determine the validity for this process, the mean scores for each item and overall mean scores were computed.

4.3.3.2.1 AEPET Test Specification

Table 4.19 presents the mean scores for the AEPET's Test Specification (Q39– Q45)

Table 4.19: Mean scores for AEPET Test Specification

No	Items	Mean	SD	Degree
39	I am given information about AEPET (test specification) before constructing test items.	1.75	.550	L
40	When constructing the test items, I refer to the test specification .	1.75	.550	L
41	When constructing the test items, I refer to the test specification to determine the topics to be covered.	1.75	.550	L
42	When constructing the test items, I refer to the test specification to determine the time allocation .	1.75	.550	L
43	When constructing the test items, I refer to the test specification to determine the number of test items .	1.75	.550	L
44	When constructing the test items, I refer to the test specification to determine the difficulty level .	1.75	.550	L
45	When constructing the test items, I refer to the test specification to determine the marks allocation (weightage) .	1.75	.550	L

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.19 shows that of the 7 items on reference to test specification for the construction of AEPET (Q39- Q45), the mean score for each item is the same (M=1.75), falling in the

category of low agreement level. This seems to indicate that the lecturers did not refer to the test specification to determine the topics to be covered, time allocation, number of test items, difficulty level, and marks allocation when constructing AEPET items.

Referring to test specification is a standard construction guideline for a test development which helps the lecturers map the objectives, instructions and assessment. However, during the interview, the lecturers explained that they were not provided with the AEPET's test specification; hence they did not refer to it. Below are excerpts that illustrate their point.

Lecturer 4 : Because I wasn't given any information on the test specification, so if you ask me, I don't have much to say.

Lecturer 5 : Nobody provided me any information on AEPET Test specification, so I did not refer to test specification when constructing the test.

4.3.3.2.2 English Course Syllabus

Table 4.20 presents the mean scores for English course syllabus

Table 4.20: Mean Scores for English Course Syllabus

No	Items	Mean	SD	Degree
46	I refer to English course syllabus when constructing the test items	2.30	.923	M
47	When constructing the test items, I refer to English course syllabus to determine the topics to be covered.	2.30	.923	M
48	When constructing the test items, I refer to English course syllabus to determine the time allocation .	2.30	.923	M
49	When constructing the test items, I refer to English course syllabus to determine the number of test items .	2.30	.923	M
50	When constructing the test items, I refer to English course syllabus to determine the difficulty level .	2.30	.923	M
51	When constructing the test items, I refer to English course syllabus to determine the marks allocation (weightage) .	2.30	.923	M

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.20 shows that of the 6 items on English course syllabus, (Q46– Q51), the mean score for each item is the same (M=2.30), indicating the moderate level of agreement. This shows that the lecturers are uncertain in referring to the English course syllabus to determine the topics to be covered, time allocation, number of test items, difficulty level, and marks allocation (weightage) when constructing AEPET items.

The interview data on whether the lecturers referred to the English course syllabus when designing the AEPET, show mix responses among them. The lecturers who referred to the syllabus agreed that the course syllabus helps them to be sure of the minimal educational goals for their teaching and testing. Below is an example that illustrates this view.

Lecturer 3 : Referring to English course syllabus makes me feel that I'm on the right track of teaching and designing test for students

Furthermore, some lecturers mentioned that referring to the syllabus provides them an important basis for developing the test such as clear descriptions to determine test purpose, relevant types of language use, students' knowledge levels, weightage, time allocation, difficulty level of test item, number of test items and so on. The excerpt below is the evidence which illustrates for their response

Lecturer 2 : To make sure no mistakes, not too far from students' knowledge in my test, I tend to use the syllabus when I write the test. Of course, when I design the test, I always pay attention the difficulty level, time allocation and other elements such as topics to be covered in the test, number of test items and mark allocations

On the other hand, the lecturers who did not refer to English course syllabus, expressed their lack of understanding about the syllabus. The following excerpt exemplifies this finding

Lecturer 19 : I'm a new lecturer here, so I don't seem to understand it much

Another lecturer lamented that the full coverage of the syllabus' requirements in the test was not practical and accordingly this made students get stressed due to over studying. This may affect students' performance. The following excerpt illustrates her statement.

Lecturer 12 : If you try to cover the syllabus in the test, you will make students more pressured while they are taking the test because there are a lot of aspects in the syllabus, so there is no need to refer to the syllabus when designing a test

4.3.3.2.3 Sample AEPET Papers

Table 4.21 presents the mean scores for referring to sample AEPET papers (Q52– Q57)

Table 4.21: Mean Scores for Sample AEPET Papers

No	Items	Mean	SD	Degree
52	I am given sample AEPET papers before constructing test items	3.50	.945	H
53	When constructing the test items, I refer to the sample test paper to determine the topics to be covered.	3.50	.945	H
54	When constructing the test items, I refer to the sample test paper to determine the time allocation .	3.50	.945	H
55	When constructing the test items, I refer to the sample test paper to determine the number of test items .	3.50	.945	H
56	When constructing the test items, I refer to the sample test paper to determine the difficulty level .	3.50	.945	H
57	When constructing the test items, I refer to the sample test paper to determine the marks allocation (weightage) .	3.50	.945	H

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.21 shows that of the 6 items on sample AEPET papers (Q52– Q57), the mean score for each item, is the same (M=3.50). This indicates the high degree of agreement. The results

shows that the lecturers refer to test sample to decide on the topics to be covered, time allocation, number of test items, difficulty level, and marks allocation (weightage) when constructing AEPET items.

The frequent revision with sample papers for test development is one of quickest ways that helps the test developers to work out with the test content and format. Thus, when the lecturers were asked whether they referred to sample AEPET papers, they explained that referring to sample AEPET papers helped them to determine topics covered, time allocation, numbers of test items, difficulty level of test items and mark allocation (weightage). The excerpts below are some examples supporting what they claimed.

Lecturer 1 : Admittedly, I collected some AEPET papers from AEP administrators and based on the format, grammatical level, number of items, topics, difficulty level when designing the test items

Lecturer 5 : If you don't want to redesign or correct the mistakes again and again in your test, using the old test sample is the easiest way to save your time and energy

4.3.3.2.4 Answer Keys and Marking Scheme (Rubrics)

Table 4.22 presents the mean scores for answer keys and marking scheme (rubrics) (Q58-Q59)

Table 4.22: Mean Scores for Answer Keys and Marking Scheme (rubrics)

No	Items	Mean	SD	Degree
58	The answer keys for the objective questions are prepared during the item construction.	3.55	.887	H
59	The marking scheme (rubrics) is prepared before the examination is administered.	3.50	.888	H

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.22 shows that both Q58 (M=3.55) and Q59 (M=3.50) fall in the category of high degree of agreement degree. In other words, the lecturers highly agree that the answer keys for the objective questions are prepared during the item construction and the marking scheme (rubrics) is prepared before the examination is administered.

In response to the follow-up interview, most of the lecturers claimed that the answer keys for objective questions and marking scheme (rubrics) were prepared and it is a requirement before the examination is administered. The following segments from the lecturers' transcripts support this point.

Lecturer 8 : When I prepare the test, I have to prepare the answer keys and marking schemes for the test banking system

Lecturer 10 : Before sending the test to the administrators, all answers keys for the objective questions must be prepared completely during the item instructions and the same with marking schemes. The administrators will not accept a test without answer keys or marking schemes

4.3.3.2.5 Vetting for Questions, Answer Keys and Marking Scheme (Rubrics)

Table 4.23 presents the mean scores of the vetting process for questions, answer keys and marking scheme (Q60-Q62).

Table 4.23: Mean Scores for Vetting Session

No	Items	Mean	SD	Degree
60	The questions for the test are vetted in a meeting.	1.95	.510	L
61	The answer keys for the objective questions are vetted in a meeting.	1.95	.510	L
62	The marking scheme (rubrics) is vetted in a meeting.	2.10	.852	L

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.23 shows that Q60 (M=1.95), Q61 (M=1.95) and Q62 (M=2.10) fall in the category of low of agreement. The vetting of assessment tools is a crucial process which helps to

review and evaluate question items, answer keys, marking scheme (rubrics) based on the specific criteria with the aims to detect errors and accordingly to improve the quality. However, these results suggest that the vetting was not practiced.

During the interview, the lecturers revealed that there was no vetting for questions, answer keys and marking scheme for AEPET. One lecturer explained that each lecture seemed to be entirely responsible for the process of constructing each AEPET component, accordingly there was no test -vetting session with other colleagues. The following excerpt illustrates for her statement.

Lecturer 10 : Because each lecturer will do everything by themselves, so there is no vetting for questions, answer keys and marking schemes in the meeting

4.3.3.2.6 Rater Training Session

Table 4.24 presents the mean scores for rater training session (Q63)

Table 4.24: Mean Scores for AEPET Rater Training Session

No	Items	Mean	SD	Degree
Q63	I attend rater training session (moderation) before assessing the answer scripts assigned to me.	2.25	.786	L

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.24 shows that lower mean score is for Q63 (M=2.25) which is item on attending rater training (moderation) session before assessing the answer scripts. Q63 also falls in to the group of low level of agreement. Rater training session plays an important role which helps to make sure that the judges have similar understanding in marking scheme, thus when they do the marking; they can give their judgments which are based on the same thing. Lacking of rater training session will lead to poor assessment for students' performance. However, this

result shows that rater training sessions for the AEPET were not carried out before the lecturers assess the answer scripts assigned to them.

During the interview, one lecturer explained that she only referred to the provided answer keys and marking scheme (rubrics) and then did the marking by herself. The excerpt below is an example that exemplifies for her statement.

Lecturer 19 : I have never attended any vetting or rater training session because I check and do the marking by myself

4.3.3.2.7 Overall Mean Scores for AEPET Test Preparation

Table 4.25 presents the overall mean scores for all items in the AEPET preparation

Table 4.25: Overall mean scores for AEPET's preparation process

Items	Overall Mean	SD	Degree
Test Preparation (Q39-Q63)	2.48	.780	L

Note: *VH =Very High, H=High, M=Moderate, L = Low, VL = Very Low

Table 4.25 shows that the overall mean score for content validity related to test preparation is **2.48**, and this shows that the content validity related to test preparation is rather low.

4.3.4 Summary of Content Validity Results for the AEPET

Based on all results on AEPET's components and preparation, it can be summarized that the content validity in relation to AEPET components is moderate (M=3.35). This result can be explained by the fact that the content validity related to AEPET's preparation is rather low (M=2.48). In other words, the process for constructing the AEPET is not well-prepared; consequently this leads to the moderate content validity for AEPET components. The interview responses in relation to the process of AEPET preparation support for this finding. Weir (1993) states that in order to establish the high content validity of a test, test

specification, test sample and course syllabus should go hand in hand in the process of test construction. Although, the lecturers do refer to the sample AEPET papers as the quickest way to save time and energy, avoid mistakes and more importantly this was the easiest way to get the form and format right, there seem to be a lack of standard procedure for test construction. However, AEPET's test specification it is not available for reference. Similarly, they were uncertain in referring to the course syllabus as crucial guide in preparing the test items and ensure coverage of topics.

Although the questions, answer keys and marking scheme were prepared before the examination, they were not vetted before the test is administered. Thus, they were not put through any quality assurance to remove possible mistakes or ambiguity. This may also lead to confusion or wrong understanding of the questions by the students which would be reflected in their answers as well.

Moreover, rater training session is crucial for ensuring the consistency in interpreting the marking scheme and judging the performance assessments, especially for Writing and Speaking, it was not carried out. This may result in inconsistency in interpreting the marking scheme and eventually lead to the failure of the quality assessment for students' performance.

In a nut shell, it can be concluded that the low content validity for test preparation seems to be the reason which causes the moderate content validity for AEPET components.

4.4 Summary of Results on Concurrent, Predictive and Content validity of the AEPET

This section aims to summarize the results on concurrent, predictive and content validity in order to further set up an approach to assess multiple aspects of test validation research. Table 4.26 presents the overall results on concurrent, predictive and content validity of the AEPET Components: Listening, Reading, Writing and Speaking.

Table 4.26: Overall Results on the Validity of AEPET Components

Component	Concurrent Validity (AEPET vs. IELTS)	Predictive Validity (AEPET vs. CGPA)	Content Validity
	Strength of Correlation		Degree of Agreement
Speaking	Significant (Moderate, $r=.517$)	Significant (Moderate, $r=.549$)	High ($M=3.56$)
Reading	Significant (Weak, $r=.351$)	Significant (Moderate, $r=.451$)	High ($M=3.51$)
Listening	Significant (Weak, $r=.261$)	Significant (Moderate, $r=.414$)	Moderate ($M=.3.30$)
Writing	Significant (Weak, $r=.307$)	Significant (Weak, $r=.266$)	Moderate ($M=.3.01$)
Overall	Significant (Weak, $r=.398$)	Significant (Strong, $r=.613$)	Moderate ($M=3.35$)

Notes

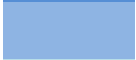
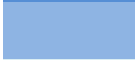
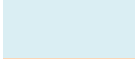
Strength of Correlation		Degree of Agreement	
	: Very Strong correlation ($r=.80-1.0$)		: Very high (4.5-5.0)
	: Strong correlation ($r=.60-.79$)		: High (3.5-4.4)
	: Moderate correlation ($r=.40 - .59$)		: Moderate (2.5 – 3.4)
	: Weak correlation ($r=.20-.39$)		: Weak (1.5- 2.4)
	: Very weak correlation ($r=.00-.19$)		: Very weak (1.0-1.4)

Table 4.26 shows each component of the AEPET: Speaking, Reading, Listening and Writing has concurrent, predictive and content validity. For concurrent validity, all AEPET components scores significantly correlate with IELTS scores. The strength of correlation ranges from weak to moderate. Across the two tests, moderate correlation is for Speaking, while weak correlations are for Reading, Listening, Writing and Overall. In other words, on the whole AEPET seems to possess concurrent validity with the IELTS. This suggests that students who achieve high scores on AEPET components tend to obtain similar scores on those of the IELTS. Similarly, for predictive validity, all AEPET components significantly correlate with CGPA. The strength of correlation ranges from weak to strong. Among these components, strong correlation is for the relationship between CGPA and AEPET Overall

score, followed by AEPET Speaking, AEPET Reading and AEPET Listening which are moderate correlations. However, weak correlation is for the relationship between AEPET Writing and CGPA. Overall, the AEPET is able to predict in a way similar to IELTS, thus the AEPET can be considered as a significant predictor for students' academic success. This suggests that those students, who have high scores in the AEPET, are likely to perform academically better in CGPA scores.

For content validity results, both AEPET Speaking and Reading possess high content validity and then it is followed by AEPET Listening and Writing which achieve moderate content validity from the lecturers' judgments. Overall, the AEPET has moderate content validity. The contribution for the moderate content validity of the AEPET comes from the weaknesses in AEPET Preparation. There is a lack of uniformity in referring to relevant documents for constructing the AEPET. Furthermore, the crucial vetting session for questions, answer keys and marking scheme is not practiced. Last but not least, rater training session for ensuring the consistency in interpreting the marking scheme and judging the performance assessments, especially for AEPET Writing and Speaking components is not carried out.

In a nut shell, it can be concluded that the AEPET can be considered as quite a good test due to the consistent results on three validity evidence: concurrent, predictive and content validity. However, the process of preparation for the AEPET needs to be improved in order to increase the strength of relationship, as well as make the test more valid.

4.5 Chapter Review

This chapter, which is part of the results chapter for the Main Study, presented quantitative results on Concurrent and Predictive validity of the AEPET and both quantitative and qualitative results on content validity of the AEPET.

The next chapter, i.e. Chapter 5, presents the overall discussion of the findings, pedagogical implications, some limitations of the study and recommendations for further research and concluding remark.

CHAPTER 5

OVERALL DISCUSSION AND CONCLUSION

5.0 Chapter Overview

This final chapter discusses the findings and answers the research questions. Subsequently, the chapter discusses the implication of the findings on language test organizations in general and test validation in specific. The chapter then concludes with discussions on the limitations of the study and suggestions for future research.

5.1 Overall Discussion

This section brings together all aspects of results from the study, answers the research questions and discusses the findings. The discussion is thus guided by the research questions

5.1.1 RQ1: What is the relationship between the AEPET and IELTS?

The descriptive statistics results in section 4.1, Chapter 4 provided evidence for concurrent validity of the AEPET. The results show that there is a significant correlation between AEPET and IELTS. The AEPET Overall band, Listening, Reading, Writing, Speaking scores significantly correlate with the same components of the IELTS. The strength of correlations ranges from weak to moderate. Across the two tests, a moderate correlation was accounted for Speaking ($r = 0.517$) while the correlations for Reading ($r = .351$), Writing ($r = .307$) and Listening ($r = .261$) were found to be weak. Among the four AEPET components, AEPET Listening is the component which appears to have the lowest concurrent validity value ($r = .261$) than other components. Brindley and Slatyer (2002), who investigated the validity of

the Listening test at a public university suggested that during the test students were allowed to listen to the text only once, and this may cause stressed to the students or restricted by the time when taking the test (Chang & Read, 2006). In addition, Zarin (2013) stated that the speaker's accent is one of challenges that affect students' listening comprehension. She mentioned that during the listening test in which its tape scripts were recorded by native English speakers, most students had discomfort, anxiousness and difficulty because they had to struggle with unfamiliar accent, thus students felt demotivated to do the listening test, and as a result, they failed to acquire the important aspect of second language learning. Last but not least, other factors related to the condition of audio-tape, the nature of text (dialect, speech rate, length) (Chao, 2013) are similar reasons which make students to have more difficulties in their listening comprehension. In the present study, AEP students also have similar problems in the listening skill, therefore they take the wrong way when performing the test and this leads to their poor results. Similar these factors could have contributed to the weak concurrent validity of listening scores between the AEPET and IELTS.

On the other hand, the AEPET Overall band scores ($r=.398$) show that on the whole AEPET seems to possess concurrent validity with the IELTS. This suggests that students who achieve high scores on AEPET, especially on AEPET Speaking, Reading, Writing and Overall band scores tend to obtain similar scores on those of the IELTS. The finding matches similar conclusion made in previous studies. For example, Glass and Hopkins (1984) declared that "correlation is a measure to describe the degree of relationship between two variables, providing an example to illustrate the meaning of correlation in which students' scores might be paired with their English scores to determine whether students who earn high scores in one subject also earn them in another" (p. 79). Similarly, Riazi (2013) investigated the concurrent

validity between the Pearson Test of English Academic (PTEA) and IELTS at Macquarie University in Australia. The results showed that there were strong correlations between students' overall PTEA and their overall IELTS Academic scores ($r = 0.824$), and the four communication skills of: Listening ($r = 0.661$), Reading ($r = 0.677$), Speaking ($r = 0.723$), and Writing ($r = 0.686$). The researcher drew a conclusion that those who achieve higher scores on PTEA will achieve higher scores on IELTS and vice versa.

Similarly, Weir, Chan and Nakatsuhara (2013) investigated the concurrent validity of the Reading and Writing components of the Advanced Level General English Proficiency Test (GEPT) at a British University. The results showed that the correlation between the participants' GEPT Writing scores and IELTS Writing scores were found to be moderate ($r=.432$, $p<.01$) due to the limited range of values for both tests, in which GEPT Writing has five band scale (Band 1- Band 5) while IELTS Writing has nine band scale (Band 1- Band 9). Furthermore, the correlation between GEPT Reading scores and IELTS Reading scores was significant ($r=. 520$, $p<.01$). The findings suggested that both GEPT Reading and Writing scores had concurrent validity with IELTS Reading and Writing scores and students who achieved high scores on GEPT Reading and Writing scores tend to gain similar scores on the same sections of the IETLS. However, it seemed to face more difficulty for the students to pass the GEPT Advanced Writing Test than Reading Test. Similar to the case of the this study, AEP students seemed more difficult to pass AEPET Writing than AEPET Reading. Neena and Dash (2007) believe that learning reading skill is not so difficult as writing skill because students can improve their reading skill by reading a lot, however, they have to pay more time in studying and enhancing their writing skill. Writing in English requires different stages such as planning, collecting data, drafting, revising, rewriting and editing. Thus, in

order to improve this skill, students have to spend more time to develop their grammar, vocabulary, writing strategies, cultural knowledge and other social contexts. Additionally, Lap and Truc (2014) highlighted that writing has been the most problematic to Vietnamese students because they are not aware of the importance of writing skill in their learning. Thus, they often get low marks when doing the tests on writing skill and it affects their results (Huy, 2015; Nguyen, 2009). These factors could have contributed to the fact that AEP students tend to achieve better score on reading skill rather than writing skill, and explains the reason why AEPET Writing had less predictive value than AEPET Reading.

In a nut shell, it can be concluded that the scores of AEPET components: Listening, Reading, Writing, Speaking and Overall band significantly correlate with the scores of the same sections of the IELTS. Hence, it suggests that those students who achieve higher scores on the AEPET have the tendency to gain similar scores on the IELTS.

5.1.2 RQ2: What is the relationship between students' AEPET scores and academic achievement (CGPA), in comparison with the relationship between IELTS and CGPA?

The descriptive statistics results in section 4.2, Chapter 4 showed that there is a significant and positive correlation between students' AEPET and CGPA. The strength of correlations ranges from weak to strong. This indicates that student's English ability is directly reflected in the AEPET influences students' academic achievement. Since English language is the medium of instruction for the degree programme, AEP students are required to speak, read, listen and write in English at an appropriate level in their studies. Thus, the positive relationship between AEPET and academic performance demonstrates that English proficiency is a determinant of academic success for students. This finding is in accordance

with other research findings (Maleki & Zangani, 2007; Sivaraman et al., 2014; Sadeghi et al., 2013; Roche & Harrington, 2013; Ghenghesh, 2015). For illustration, Alawdah and Mapurisa (2015) investigated the predictive validity of an English language proficiency test among microeconomics students in Saudi Arabia. The microeconomics courses were taught in English. This study found a positive significant correlation ($r=.327$, $p<.005$) between microeconomics students' English proficiency test and their overall academic performance. The results concluded that the English language proficiency test was a good predictor of students' academic achievement in microeconomics. The study also recommended paying special attention at students' proficiency in English because higher language proficiency test would obtain higher academic achievement scores. Similarly, Fakeye (2014) compared English Language Proficiency Test (ELPT) and students' academic achievement at Nigerian secondary school. The results showed that the ELPT had a significant positive correlation with overall academic achievement measured by CGPA ($r=.499$, $p<.05$). The study suggested that the more proficient the participants were in the English language test, the better was in their academic performance. Thus, they were recommended that more effective teaching and learning English strategies should be applied in order to improve the participants' English proficiency.

In this study, the results show that AEPET Overall band, Listening, Reading, Writing, Speaking scores significantly correlate with CGPA. AEPET Overall ($r=.613$) strongly correlates with CGPA, showing 37.6% of the CGPA variance. Furthermore, AEPET Speaking ($r=.549$), Reading ($r=.451$), and Listening moderately correlate with CGPA. In contrast, AEPET Writing ($r=.266$) weakly correlated with CGPA. Among the four AEPET components, AEPET Writing has less predictive validity value than other components due to

its lower correlation coefficient ($r=.266$). As mentioned earlier, AEPET writing seemed to be a challenge for AEP students, thus they tend to make a variety of errors or mistakes when writing. Huy (2015) explained that the weakness in the writing skill is a common phenomenon among Vietnamese students due to the fact that students have limitations related to grammar, vocabulary, material sources, time and motivation to improve their writing skill (Lap & Truc, 2014; Nguyen, 2009). Furthermore, he showed that the topics and exercises of writing lesson in textbooks are inappropriately designed to students' levels; many topics have a lot of new words and its grammar points which make students feel complicated and more importantly, many lecturers are busy with their teaching, so they do not have time to correct students' written works. Similarly, Kansopon (2012) investigated the validity of writing test on the undergraduates of Institute of International Studies, Ramkhamhaeng University (IISRU) where English is a medium of instruction. She states that writing ability is very important to students' academic achievement at IISRU. However, the results showed that at the end of the first year, many students had difficulties in achieving the academic writing skills required and were unable to meet the benchmarks of writing. As a result, some students failed the course while others wanted to change their programme. The findings found that the weakness of writing skills among the students came from the lack of critical and analytical thinking skills because students were unable to compose their thoughts and ideas clearly and logically. Furthermore, Al-Khasawneh (2010) emphasizes that in order to succeed in writing skills students need to surround themselves in a language learning environment. However, students from non-native English backgrounds have little opportunities to use English language in their society, therefore, the weaknesses in writing skills occur when students study at a university where English is the medium of instruction. Similar to the case of this study, AEP students also have little opportunities to learn English writing skills through

natural interaction with native English speakers. Furthermore, they have to struggle with a lot of assignments, subject requirements and examinations during their one year English course, thus they have no time to develop writing skill. More importantly, during the writing test, students are given under the time limits of one hour responding to two writing tasks, so they easily get stressed while performing the test. From these problems above, AEP students often have difficulties in learning writing skill, and this could contribute to the lower writing performance of the students as well as the effect of depressing the predictive validity of coefficient in AEPET Writing component.

The AEPET Overall band score ($r=.614$) strongly correlates with CGPA. Thus, the results show that on the whole AEPET is considered as a significant predictor for students' academic performance. This suggests that those students, who have high scores in the AEPET, are likely to perform academically better in CGPA scores. These findings are in line with some researchers (Ajibade 1993; Elder, 1993; Fakeye & Ogunsiji, 2009, Al Hajr, 2013; Maleki & Zangani, 2007; Othman & Nordin, 2013; Sahragard, Baharloo & Soozandehfar, 2011) who found a significant and positive relationship between in-house English language tests at university and academic performance, as measured by CGPA. In these previous studies, the medium of instruction is English. The results showed that English language tests can predict the students' academic performance. These researchers also suggested that as the English test score increases, students' academic success also increases. Thus, in order to have better CGPA scores, students in the present study are suggested to achieve good competence in Speaking, Reading and Listening scores because a combination three component scores could make a contribution to the large variance of academic success. This suggestion concurs with

the previous suggestion of Al Haijr (2013), Heitner, Hoekje and Braciszewski (2014) and Yinxin and Daller (2015).

Regarding the predictive validity of IELTS, the results show that the IELTS also significantly correlates with CGPA. Among the four components, IELTS Speaking ($r=.535$), Listening ($r=.471$) and Reading scores ($r=.428$) demonstrate moderate correlations with CGPA while IELTS Writing scores ($r=.314$) weakly correlate with CGPA. On average, IELTS Overall band scores ($r=.614$) strongly correlate with CGPA. The existence of significant correlations amongst other components of IELTS is aligned with a number of previous studies (Elder, 1993; Feast, 2000; Huong, 2001; Kerstjens & Nery, 2000; Schoepp & Garinger, 2016; Woodrow, 2006; Yen & Kuzma, 2009) in which the IELTS is regarded as a significant predictor for students' study success. The comparison between the predictive validity AEPET and IELTS showed that AEPET scores generally seemed to have similar predictive validity as IELTS scores. The results highlight a random consistency in correlations and regression results which are found weak to strong. In particular, both AEPET and IELTS Overall strongly correlate with CGPA while Speaking, Reading and Listening scores moderately correlate with CGPA. However, Writing scores in the two tests weakly correlate with CGPA. This indicates that just like IELTS, the AEPET can predict students' academic success. Additionally, AEPET scores also significantly correlates with IELTS scores (refer to section 4.1, Chapter 4) and the AEPET format is similar to the format of the IELTS in which it measures the students' English language ability in four skills: Listening, Reading, Writing and Speaking. The students sat for the AEPET and then for IELTS, therefore the AEPET is considered as a part of preparation for the IELTS (refer to section 3.2.1.1.2, Chapter 3). Thus, based on these consistent results in the predictive validity of the two tests, this suggests that

the construct of the two tests is similar; accordingly the students could draw somewhat similar scores on the same component of both tests (Hoa, 2008).

In recent years, several commercially produced proficiency exams such as TOEFL, IELTS, TOEIC, PET, KET and so on are available on the market. Nevertheless, these exams are costly, only offered at limited times and may not be appropriate for the needs of some language programmes (Sim, 2015). After taking the AEPET, students have to sit for the IELTS because they need a particular IELTS score to apply for a visa and university admission in the United States. However, paying about USD250 for the IELTS per each student now could affect Vietnamese public universities' educational budgets. Thus, just like IELTS, the AEPET's ability to predict students' study success could be the potential for the university to use the AEPET to check students' knowledge accomplishment as well as an admission test for overseas study. More importantly, this could save time and budgets for students and the public university in which the AEP is offered.

To sum up, the AEPET is able to predict students' academic success in a similar way to IELTS. The AEPET can be considered as a valid measure of students' language ability. The higher students' AEPET scores, especially on AEPET Speaking and Reading scores, the better CGPA scores they can achieve. The cross-test comparability for predictive validity results supports the evidence that both AEPET and IELTS can predict students' academic achievement. Therefore, the higher scores on both AEPET and IELTS students obtain the better CGPA scores they can achieve. However, the findings suggest the focus on using the AEPET as a university English language test to check students' English skills and

competencies and as admission test could help the public university save more time and educational budgets.

5.1.3 RQ3: What is the Content Validity of the AEPET?

The AEPET is an end-year English achievement test which is designed to measure the knowledge that students have been taught during their one-year English course. This test consists of four English components: Listening, Reading, Writing and Speaking. Furthermore, the test content and knowledge has been used since the AEP was established. The validity of a test is one of the most important factors that reflect the success of an English language instruction programme and an analysis of the content of a test is the most essential part of validity as well (Bachman, 1990). If the test content is less valid, it cannot reach the desired objectives of teaching and cannot measure students' English skills appropriately, then how can teaching method and students' true English abilities be improved? Will students feel waste their time and money studying the course?. As a result, a less valid test will spoil the whole process of educational system (Lee & Greene, 2007; Messick, 1996; Siddiek, 2010). Bearing this in mind, it has always been useful and necessary to acquire insights into the content validity of the AEPET. The content validity of the AEPET was investigated into two parts: content validity of AEPET components: Listening, Reading, Writing and Speaking and content validity of the process of AEPET's preparation. The answers on content validity of AEPET components and AEPET preparation are presented individually as follows.

5.1.3.1 Content Validity of AEPET Components

In order to investigate the content validity of the AEPET Components: Listening, Reading, Writing and Speaking, the English language lecturers were asked to give comments on 7

items which are related to weightage, time allocation, representation of skills, representation of topics, clarity of questions, clarity of instructions and mark allocation. The findings on each component are presented individually.

For AEPET Listening, the mark allocation, clarity of instructions and representation of listening skills are highly represented in the test. Nonetheless, the coverage for appropriacy of the weightage, sufficiency of time allocation, representation of topics and clarity of questions are at moderate level. On the whole, AEPET Listening seems to obtain moderate content validity from the lecturers' judgments.

For AEPET Reading, the clarity of instructions, mark allocations and clarity of questions are highly represented in the test. Meanwhile, the coverage for sufficiency of time allocation, representation of reading topics and appropriacy of weightage are at moderate level. Unlike AEPET Listening, AEPET Reading Component seems to possess high content validity from the lecturers' judgments.

For AEPET Writing, the clarity of instructions and questions, representation of writing skills and topics are highly represented in the test. In contrast, mark allocations, the coverage for appropriacy of weightage and sufficiency of time allocation are at moderate level. Thus, just like AEPET Listening, AEPET Writing component appears to have moderate content validity from the lecturers' judgments.

For AEPET Speaking, the clarity of questions and instructions, representation of speaking skills and topics are highly represented in the test. However, the coverage for appropriacy of

weightage, sufficiency of time allocation and mark allocations are at moderate level. On the whole, AEPET Speaking component seems to achieve high content validity from the lecturers' judgments.

Overall, the results for content validity show that both Speaking and Reading Tests have high content validity while Listening and Writing Tests have moderate content validity. Thus in order to know more the reason leading to this non- parallelism, there was a need to investigate to what extent AEPET preparation adhered to crucial test preparation procedures before the examination was administrated.

5.1.3.2 Content Validity of AEPET's Preparation

The content validity of the AEPET Preparation was investigated in terms of six items: AEPET's test specification, English course syllabus, sample AEPET papers, preparation for answer keys and marking scheme, test vetting and marking scheme and rater training sessions. The results on content validity of each component are presented individually.

On the whole, the descriptive statistics results on AEPET Test Specification showed that the lecturers did not refer to the test specification to determine the topics to be covered, time allocation, number of test items, difficulty level, and marks allocation when constructing the items. The results from the interview showed that AEPET's test specification was not provided to the lecturers. Referring to test specification for test construction not only helps the teachers have a clear picture about who the test is aimed at, what its purpose is, what the content should to be covered, how long the test takes and so on, but also helps the administrators whether the test is valid for specific academic purposes (Sadie, 2010). Thus,

this suggests that the absence of test specification could have made the lecturers unable to design test items effectively. Accordingly, this may have led to poor content validity of the AEPET. This finding is in line with previous findings which pertain that using the test specification helps the teachers map the objectives, instructions and assessment, thus teachers who do not use test specifications during the test construction are likely to fall short of assessing students achievement well and likely to have poor content validity (Ali & Aube, 1988; Fives & Donato-Barnes, 2013; Kahsay, Devardhi & Nelson, 2014; Ing, Musah, Al-Hudawi, Tahir, & Kamil, 2015; Kindeya, 2002; Kinyua & Okunya, 2014; Marso, 1989; Notar et al., 2004). Supporting this finding, Siddiek (2010) also indicated that lack of using the test specification before test development will increase the limited awareness about the perception of content validity and test application. Thus, he suggested that it should be better if the test-developers set up test specification in details before writing test items.

In relation to English course syllabus, the results showed that the lecturers seemed to be uncertain in referring to the English course syllabus when constructing AEPET items. The interview showed mix responses among the lectures that could be explained for this uncertainty. The lecturers who referred to English course syllabus during test construction supported the view that referring to the syllabus provides them an important basis for developing the test such as clear descriptions to determine test purpose, relevant types of language use, students' knowledge levels, weightage, time allocation, difficulty level of test item, number of test items and so on. However, other lecturers who did not refer to the course syllabus expressed their lack of understanding about the syllabus. This suggests that the lack of referring to course syllabus during test construction has likely to be lower the validity of the test content. This finding concurs with the finding of Nwagu (2009), Kindeya (2002),

Wang (2006), Siddick (2010), Kahsay and Devardhi (2014) and Ing et al. (2015) who indicated that without referring to the course syllabus for test development will make the test have no guarantee in educational goals, thus this may cause poor content validity. Siddick (2010) also emphasized that when the items do not cover the constituents of the syllabus, no content validity is provided for the test and as a result, “the whole process of teaching will be for obtaining nothingness” (p.141), Consequently, Kindeya (2002) emphasized that during the process of test preparation and construction the test must be according to the content and weight of the given syllabus. Thus, referring to the course syllabus is a crucial step in test preparation which was found to be rather lacking among the AEPET lecturers.

The lecturers referred to sample AEPET papers when constructing AEPET items. They revealed that referring to sample papers was one of quickest ways that helped them to work out with the test content and format in order to decide on the topics to be covered, time allocation, number of test items, difficulty level, and marks allocation (weightage) for AEPET items. Weir (1990) stated that “sample test items are a key issue in language testing” (p. 11). The frequent revision with sample papers for test development is one of quickest ways that helps the test developers to work out with the test content and format. Furthermore, Hughes (1998) and Kindeya (2002) addressed that only referring to proper sample of very relevant test; the test would have content validity. However, the lecturers purely referred to the former sample AEPET papers and more importantly, no empirical test validation studies have been published by the university where the study was conducted, thus there seems to be lack of evidence which helps to prove the past AEPET sample paper to be proper ones. Weir (1993) suggests the best way to establish the high content validity of a test is that the lecturers should refer to test specification, test sample and course syllabus in the process of test construction.

Preparing answer keys and marking scheme before the examination is a requirement of the administrators. Hence all the questions, answer keys and marking scheme for the AEPET were prepared before the examination. However, they were not put through any quality assurance to remove possible mistakes or ambiguity in the questions. Test vetting is a very important process to sustain the validity of questions, answer keys and marking scheme because it helps to review and evaluate the questions, answer keys and marking scheme with the aims of removing errors and then make appropriate modifications to improve their qualities (Abedi, 2006; Educational Testing Service, 2007; Kopriva, 2008; Wadi, 2012). The test vetting session should be reviewed carefully by a vetting committee of an examination unit. Therefore, this suggests that the lack of test vetting process may lead to students' confusion or wrong understanding of the questions which would be reflected in their performance as well. This also may lead to the wrong understanding of the answer keys and marking scheme by the lecturers which would be reflected in their assessment (Bailey, 2007; Educational Testing Service, 2006; Wadi, 2012). Consequently, the findings suggest that the absence of test vetting process may be one of the factors that cause the rather low content validity for the test.

In addition, the results show that rater training sessions for the AEPET were not carried out before the lecturers assessed the answer scripts assigned to them. During the interviews, the lecturers explained that they had tendency to do the marking by themselves. Language testing experts have emphasized that rater training session is crucial for ensuring the consistency in interpreting the marking scheme and judging the performance assessments, especially for Writing and Speaking (Kindeya, 2002; Kopriva, 2000; Nwagu, 2009). However, when rater training is not conducted, this will prompt the lecturers to make their own judgments in

assessing the answer transcripts (Nwagu, 2009; Wang, 2006). This suggests that the absence of rater training session will lead to subjectivity in judgments, and this will result in poor assessment of students' performance. This interpretation is in accordance with Kindeya (2002), Montee, Bach, Donovan and Thompson (2013) and Nwagu (2009) who also found teachers' shortcoming in test construction due to lacking training course in assessment. Similarly, Al-Mahrooqi et al. (2017) states that when the teachers have not attended any training course on language testing and assessment, or have not received any guidelines to mark students' test papers, they logically do not have sufficient assessment literacy. Hence, they have tendency to carry on the test assessment based on their own teaching experience; as a result, the scores which they assigned to students might be misrepresenting students' true abilities. In the light of the literature review above, it can be reasonably assumed that English language lecturers at the public university where the AEP is offered seem to lack assessment literacy, and that is also one of important factors which could have led to rather low content validity of the designed test.

In short, the results show that AEPET components achieve moderate content validity while the process of AEPET preparation possesses rather low content validity due to lack of basic construction guidance for test development, lack of test vetting process and rater training session. Overall, the AEPET has moderate content validity. The absence and lack of crucial steps in the AEPET's preparation is a reasonable explanation for the moderate content validity results on AEPET components. Thus, in order to improve the concurrent, predictive and content validity values for the AEPET, it would be better to start from improving the process of AEPET's preparation. The following section discusses the overall discussion for concurrent validity, predictive validity and content validity of the AEPET.

5.1.4 Overall Discussion for the Validity of AEPET

Overall, the results show that all the AEPET components: Listening, Reading, Writing and Speaking has concurrent, predictive and content validity. Thus, they can be considered as a valid English language test.

More specifically, for concurrent validity, there is a significant correlation between AEPET scores and IELTS scores. The strength of correlations ranges from weak to moderate. Thus, it is suggested that those students who achieve higher scores on the AEPET on Listening, Reading, Writing and Speaking have the tendency to gain similar scores on the same sections of the IELTS. For predictive validity, the AEPET significantly correlate with CGPA. The strength of relationship ranges from weak to strong. Thus, it can be concluded that the AEPET is considered as good predictor for students' academic performance. The finding suggests that those students, who have high scores on the AEPET, are likely to achieve better CGPA scores. For content validity, the content validity of the AEPET was investigated in to two aspects: AEPET's components and preparation. In terms of AEPET components, both Speaking and Reading Tests have high content validity while Listening and Writing Tests have moderate content validity. Overall, the AEPET has moderate content validity. However, the process of AEPET preparation seems to be rather deficient in content validity due to lack of uniformity in referring to relevant documents for constructing the AEPET. Furthermore, the crucial vetting session for questions, answer keys and marking scheme is not practiced. Similarly, rater training session for ensuring the consistency in interpreting the marking scheme and judging the performance assessments, especially for AEPET Writing and Speaking components is not carried out. These could have been contributed for the moderate content validity of the AEPET, especially the low content validity of AEPET Writing

component. Thus, in order to increase the content validity of the AEPET, there is a need to look at the standard procedure for test development in which the process of test preparation must be followed the basic test construction guidelines. Cunningham, Callahan and Field (2013) state that the combination of multi-material (specification, textbooks, syllabus) and multi-step internal review helps to design a good test. More importantly, Weir(1993) suggests that in order for test construction to be effective and useful for educators, the test writer should bring all the contents that students have already learned in class in the test. This will provide a good test which produces beneficial wash back effect on both teaching and learning. The findings from this study provide valuable information for the evidence of construct validity of the AEPET. As mentioned earlier, construct validity is the process of gathering evidence to determine the meaning of scores or score interpretation from the test to assure that the scores mean what is expected to mean (Abedi, 2006; Alderson et al., 1995; Bachman & Palmer, 1996; Ebel & Frisble, 1991). Construct validation is a unitary validity which requires a series of validity evidence in which content and criterion-related evidence are contributory factors (Bachman, 1990; Bachman & Palmer, 1996; Loevinger, 1957; Messick, 1989; Weir, Chan, & Nakatsuhara, 2013). In the light of the literature review, it is evidentially linked to the study in which the findings on concurrent, predictive and content validity the contributory factors to support evidence for construct validity of the AEPET.

To conclude, the AEPET can be considered as a quite good test due to the consistent results on three validity evidence: concurrent, predictive and content validity. Furthermore, The findings on concurrent, predictive and content validity of the study provides valuable information for the evidence of construct validity of the AEPET. However, the AEPET is room for improvement. The evidence for a room for improvement comes from the

weaknesses in AEPET's Preparation. Thus, there is a need to provide the lectures with test specification, a requirement for the lectures to refer to course syllabus, to do the vetting and rater training sessions. By improving these suggestions, the strength of correlation for concurrent and predictive validity will increase will be and as results the AEPET will be more valid.

5.2 Implications of the Findings

Teaching and learning English language plays a very important role in educational systems and in the daily life situations because each person is required to use English fluently to catch up with the development of the social sciences and technologies all over the world. To enhance the quality of teaching and learning English in educational settings, it is suggested that validating in-house English tests should always be cared and examined because the valid tests help to improve the quality of education, pinpoint strengths and weaknesses in the curriculum, programme appropriations, students' true language abilities (Bachman & Palmer, 1996; Cumming & Berwick, 1998).

Furthermore, English language test validation will show how much the educators and administrators know how much knowledge their student achieved during a course or which subjects the student are not good at, and then make immediate changes in syllabus. However, the strengths and weaknesses in test validation would not be revealed if the researchers had looked at only one validity evidence, i.e. content validity or concurrent validity or only predictive validity. Thus, this study supports the idea that the combination multiple validity evidence suggests whole picture of the validity of a language test, provides information about the status of teaching, learning and testing in order to give the appropriate decisions on

educational administration, teaching methods as well as students' English language abilities (Lofgren, 2005; Maki, 2002; Wang, 2006) and then can create a comprehensive procedure of before, during and after the test construction and administration (Zubairi, 2001; Zucker, 2003). Additionally, the study also supports for the ideas that a good in-house language test should be validated from different aspects both empirically and theoretically (Bernstein, Moore & Bounchan, 2010) and using both quantitative and qualitative approach methods may create high levels of validity of gathered data (Bachman,1990).

Due to the importance of certain key aspects related to test validation process, several researchers (Abidin, 2006; Saville, 2009; Weir, Chan & Nakatsuhara, 2013; Wu, 2011) have highly recommended Weir's (2005) socio-cognitive validation framework which appears as the most comprehensive and integrated framework. Bearing this in mind, this study adapted the test validation procedures proposed in the Weir's (2005) socio-cognitive validation framework to gather validity evidence for the AEPET. The study offers support to the idea that the frameworks for validating four English sub-skills tests: Listening, Reading, Writing and Speaking provided by Weir (2005) are useful in term of their purposes and applicability to the study.

Weir's (2005) framework suggests concurrent and predictive validity as a form of external evidence in which concurrent validity involves comparing scores of the AEPET and IELTS taken at the same time whereas the predictive validation involved comparing scores of the AEPET and CGPA of the same students some time later. The weak to strong correlations from the predictive evidence support findings from earlier studies (Alawdah & Mapurisa, 2015; Fakeye, 2014; Ghenghesh, 2015; Maleki & Zangani, 2007; Roche & Harrington, 2013;

Shadeghi et al., 2013; Sivaraman et al., 2014) which suggest English proficiency is a determinant of academic success for students. When English is a medium of instruction, the higher scores for English language, the better the students perform their academic performance. The findings highlight that the one-year English language course at university facilitates the students' English proficiency and contributes their academic success. Therefore, it suggests that more attention should also be given to effective teaching and learning English strategies in order to improve the students' English proficiency during the one year English programme.

Furthermore, Weir's (2005) framework for content validity component addresses the concern for determining test format, purpose, content knowledge, uniformity of test construction which is very helpful for the process of test construction. In the present study, the findings on content validity revealed that the lecturers seemed to be uncertain in referring to the English course syllabus and did not refer to the test specification to determine the topics to be covered, time allocation, number of test items, difficulty level and marks allocation when constructing the items. The absence of referring to test specification and course syllabus during test construction supports the findings in Ali and Akubue (1987), Fives and Didonato-Barnes (2013), Ing, Musah, Al-Hudawi, Tahir and Kamil (2015), Kahsay, Devardhi and Nelson (2014), Kindeya (2002), Kinyua and Okunya (2014) and Siddick's (2010) studies. They indicated that referring to the test specification and course syllabus helps the teachers map the objectives, instructions and assessment, thus test writers who do not use such references are likely to fall short of assessing students' achievement well and likely to have poor content validity. Therefore, it should be better if the test-developers set up test specification as well as course syllabus in detail before writing test items (Siddick, 2010). The study also found that

test vetting which is a very important process to sustain the validity of questions, answer keys and marking scheme was not practiced. Moreover, rater training session is an important step for ensuring the consistency in interpreting the marking scheme and judging the performance assessments, especially for AEPET Writing and Speaking, was not carried out. The absence of test vetting and rater training sessions support Kindeya (2002), Montee, Bach, Donovan and Thompson (2013), Wadi (2012) and Weir (1993)'s arguments for retraining to ensure the quality of answer keys and marking schemes, increase assessment literacy among the teachers as well as relook at the standard procedure for test development.

Additionally, the lower concurrent, predictive and content validity value than other components proved that AEPET Writing seemed to be a challenge for AEP students. Huy (2015) highlighted that writing is a crucial skill which the students need in order to successfully go through their academic years. When the students possess good writing skills, they tend to automatically do better at their assignments, essays and exams. Therefore, it should be better if the teachers can facilitate their students' writing skill by providing opportunities for students to work together because group work discussion will help them to think critically, generate ideas and proofread each other's work. Furthermore, the teachers should select suitable materials that are not far from their knowledge and then ask them to do more reading and do more writing practice. The connection between reading and writing is very important for developing students' writing skills. Reading will help them to improve their grammar, vocabulary, styles, forms and genres of writing in order to understand of how the language works. Therefore, the more they read and write, the better their writing skill they will achieve. More importantly, the teachers should give students an effective feedback which contains clear specific tips and corrections for each student. A quality feedback will offer

students clear and specific strategies for activating and improving their writing skills. It is hope that these brief suggestions in teaching writing skill could be potential to help students enhance their writing skills.

Last but not least, the findings of the study are beneficial to future studies in other scientific and technological departments and faculties at different advanced educational programme in other universities, where English skills are being taught and tested. Also, the study contributes useful insights to English language teaching and learning, especially in-house English test validation and prevent the mismatch between learners' true performance and their test scores.

In brief, the study has addressed some theoretical and pedagogical implications related to in-house test language test validation. In the light of these implications above, it is hoped that other institutions may take this study as a guideline to examine the quality of their locally-designed assessment tools, revisit the test construction and renew course syllabus as well as to justify the correctness of their interpretations. The study also highly recommends conducting future validation research in different levels to validate different language tests by using the four frameworks proposed by Weir (2005).

5.3 Limitations and Recommendations for Future Research

Like many research studies on language testing, the present study also has some limitations. Therefore, this section discusses the possible weaknesses of the study and makes necessary recommendation for future research

First and foremost, the present study is limited in that it refers to concurrent and predictive validity in the context of on students from AEP in one public university. The AEP is a special

programme, thus there was a need to focus on one particular setting as the in- house test is unique to that university. It is hoped that future studies investigate the validity of other in-house tests carried out in similar programmes at other universities in Vietnam or at other similar contexts in institutions around the world.

Furthermore, the present study used score data as the source for investigating the concurrent and predictive validity. While test scores are considered crucial for validity studies, future studies could extend the search to include other factors such as gender, learning background, motivation and self-assessment which could help to provide more comprehensive picture of how in- house English language test relate to students' academic performance

Secondly, the content validity data were collected from questionnaire and interview. While these are important sources of data for content validity, it is suggested that future research that focus specifically on content validity add wider range of approaches such as observations, think-aloud protocols and stimulated recall to elicit more extensive data from the expert judgments. Additionally, it would be more beneficial for future studies to place a focus on the judgments from the administrators who importantly function as the intermediates that help classroom teachers put the curriculum policies into practice (Wang, 2006) in order to improve the content validity of the designed language tests as well as better understand how administrators and lecturers work together in designing practical and useful language tests at institutional levels.

Thirdly, the present study applied Weir's (2005) socio- cognitive validation framework, and focused on three crucial components (concurrent, predictive and content validity) of validity

were addressed. Other aspects of validity such as theory- based validity; scoring validity and consequential validity were beyond the scope of the present study. Therefore, future studies could consider including these aspects of validity.

Last but not least, it would be a good idea for future studies to investigate the perceptions of the English language tests from the students who are the receivers of learning outcomes in order to better understand whether the designed tests are successfully implemented or not. Alderson (1986) and Thiel (1995), who appealed for consideration of a student's perspective on test validity, suggested that students should be encouraged to contribute to future developments in language testing by reflecting on their responses to test items, and giving comments on any inconsistency between their true test performance and their own opinions, or their teachers' views, about their true language proficiency. Thus, it is hoped that the future research could investigate how students respond to the in-house English language test after they complete their English course. This suggestion may fulfill a picture of how in-house English test reflects teaching methods, improve their English abilities and promote their learning outcomes from the students' perceptions.

In short, the study has addressed the aspects of constraints in test evaluation research. Some of these limitations are inevitably unavoidable because of the nature of the study that involve extensive data collection, analyses and contribution from many participants. In order to avoid these limitations, the study has offered guidelines and suggestions in the area of in-house testing evaluation for future research.

5.4 Concluding Remark

The present study has investigated three validity components: concurrent validity, predictive validity and content validity. Three validity components are then combined to seek for the effects on the validity of the AEPET. In fact these aspects have not been investigated in combination in any test validation studies before, especially using the mix methods: quantitative and qualitative. The study demonstrates that judgments from the lecturers are very useful for test development and the improvement of students' English abilities can contribute in prediction of their academic success.

Like other research studies on language testing and assessment, the present study also has some limitations. However, despite the limitations, the findings from the study have shed light on important factors which relate to the effect of the process of test preparation and test construction. Thus it is further hoped that this study will be a contribution to the ongoing efforts to provide more validity evidence for in-house English language tests.

REFERENCES

- Abella, R., Urutia, J., & Shneyderman, A. (2005). An examination of the validity of English-language achievement test scores in an English language learner population. *Bilingual Research Journal*, 29(1), 127-144.
- Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377-398). Mahwah, New Jersey: Erlbaum.
- Abidin, Z. A. Z. (2006). Context validity of speaking test. *Journal of Language Studies*, 2(7), 125-142.
- Alderson, J. C. (1988). New procedures for validating proficiency tests of ESP?. Theory and practice. *Language Testing*, 5(2), 220-232.
- Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8(1), 41-66.
- Advi, E. (2003). IELTS as a predictor of academic achievement in a Master's Program, *English Australia Journal*, 26(2), 42-49.
- Agu, N. N., Onyekuba, C., & Anyichie, A. C. (2013). Measuring teachers' competence in construction classroom-based tests in Nigerian secondary schools: Need for a test construction inventory. *Educational Research and Reviews*, 8(8), 431-439.

- Ajibade, A. Y. (1993). *Proficiency in English and affective factors as predictors of senior secondary school students' achievement in French* (Unpublished doctoral thesis). University of Ibadan, Nigeria.
- Aksan, M. M. (2001). *Instructors' perceptions of the content validity of the English language exams at Niğde University* (Unpublished master's thesis). Niğde University, Turkey.
- Alavi, T. (2012). The predictive validity of final English exams as a measure of success in Iranian National University Entrance English Exam. *Journal of Language Teaching and Research*, 3(1), 224–228.
- Alawdah, S. A., & Mapurisa, N. (2015). A study on English language proficiency as a predictor of academic achievement among freshmen students in Saudi Arabia. *International Journal of Arts and Commerce*. 4(9), 50-59.
- Alderson, J. C. (2000). *Technology in testing: the present and the future*. Cambridge: Cambridge University Press.
- Alderson, J. C, Clapham, C., & Wall, D. (1995). *Language test construction and evaluation* Cambridge: Cambridge University Press.
- Al-Khasawneh, S. M. (2010). Writing for academic purposes: Problems faced by Arab postgraduate students of the College of Business, UUM. *ESP World*, 2(9), 1-23.
- Bachman, L. F. (1988). Problems in examining the validity of the oral proficiency interview. *Studies in Second Language Acquisition*, 10, 149-164.

- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*, Oxford, UK: Oxford University Press.
- Bannur, F. M & Abidin, Z. A. Z. (2015). A validation a process of ESP test using Weir's (2005) socio-cognitive framework. *Procedia - Social and Behavioral Sciences*, 202(2015), 199-208.
- Blandford, A. (2013). Semi-structured qualitative studies. In: Soegaard, Mads and Dam, Rikke Friis (Eds.). *The Encyclopedia of Human-Computer Interaction* (2nd ed). Aarhus, Denmark: The Interaction Design Foundation.
- Al Hajri, F. (2014). The predictive validity of language assessment in a pre-university programme in two Colleges in Oman: A mixed method analysis. *International Multilingual Journal of Contemporary Research*, 2(2), 121-147.
- Ali, A., & Akubue, A. (1988). Nigerian primary schools' compliance with Nigeria national policy on education. An evaluation of continuous assessment practices. *Evaluation Review*, 15, 183-191.
- Al-Mahrooqi, R., Coombe, C., & Al-Maamari, F. (2017). *Revisiting EFL Assessment*, Germany :Springer International Publishing AG.
- Al-Younes, Y. A. (2006). The extent of the University of Jordan faculty members' practices of test preparation, correction, analysis, interpretation and discussion of results with students. *Assyout University Journal of the Faculty of Education*, 22(1), 153-193.

- Ayers, J. B., & Peters, R. M. (1977). Predictive validity of the test of English as a foreign language for Asian graduate students in Engineering, Chemistry and Mathematics. *Educational and Psychological Measurement*, 37(2), 461–463.
- Bachman, L. F. (1990). *Fundamental considerations in Language Testing*, Oxford, UK: Oxford University Press.
- Bachman, L. F. (1981). Languages for specific purposes: Program design and evaluation. *Formative evaluation in ESP program development*. Rowley, Massachusetts: Newbury House.
- Bachman, L. F., & Palmer, A. (1996). The construct validation of self-ratings of communicative language ability. *Language Testing*, 6(2), 13–20.
- Bachman., L. F., Davidson, F. & Milanovic, M. (1996). The use of test methods in the content analysis and design of EFL proficiency tests. *Language Testing*, 13(2), 125-150.
- Bailey, K.M. (1999). *Washback in Language Testing*. Princeton, New Jersey: Educational Testing Service.
- Bailey, A. L. (2007). *The language demands of school: Putting academic English to the test*. New Haven, U.S: Yale University Press.
- Bailey, K. M & Damerow, R. M. (2012). *Teaching and learning English in the Arabic-speaking world*. New York: Routledge.
- Bayliss, A., & Ingram, D. E. (2006). IELTS as a predictor of academic language performance, *Australian International Education Conference*, 1-12.

- Bekuretsion, S. (1997). *Communicative language testing: Its validity and reliabiltiy at Addia Ababa University* (Unpublised master thesis). Addia Ababa University, U.S.
- Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3), 355–377.
- Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10, 277–301.
- Brown, H. D. (2000). *Principles of language learning and teaching*. New York, US: Addison Wesley Longman.
- Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains, New York : Pearson Education.
- Buglear, J. (2013). *Practical Statistics: A handbook for Business Project*. London, UK: Koga Page.
- Campbell, K. (1996). *The world rushes to speak and write 'American' English*. London, UK: Oxford University Press.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, Illinois: Rand McNally.
- Canh, L. V., & Barnard, R. (2009). Curricular innovation behind closed classroom doors. *Prospect*, 24(2), 20–33.

- Canh, L. V. (2003). Local mind, global practice: ELT in the context of globalization. *Teachers Edition*, 13, 36–42.
- Caroll, J. B. (1968). *The psychology of language testing*. London, UK: Oxford University Press.
- Chan, S. H. C. (2013). *Establishing the construct validity of the reading-into-writing test tasks the UK academic context* (Unpublished doctoral thesis). University of Bedfordshire, England.
- Chang, C. S., & Read, J. (2006). The effect of listening support on the listening performance of EFL learners. *TESOL Quarterly*, 40, 375-397.
- Chao, J. Y. G. (2013). Factors affecting college English Foreign Language learners' listening comprehension and listening problems. *NCUE Journal of Humanities*, 8, 71-84.
- Cheng, L. (2008). *Washback, impact and consequences*. New York: Springer.
- Criper, C., & Davies, A. (1988). *The ELTS Validation Project Report*. London, U.K: British Council and University of Cambridge Local Examinations Syndicate
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition)*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Coley, M. (1999). The English language entry requirements of Australian universities for students of non-English speaking background. *Higher Education Research and Development*, 18(1), 7–17.

- Cope, N. (2011). Evaluating locally-developed language testing: A predictive study of ‘Direct Entry’ language programs at an Australian University. *Australian Review of Applied Linguistics*, 34(1), 40-59
- Cuieford, J. P. (1965). *Fundamental Statistics in Psychology and Education*. New York, U.S: McGraw-Hill.
- Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, 21(2), 107–145.
- Cumming, A., & Alister. H., Berwick, R. (1988). *Validating in language testing*. Clevedon, England: Multilingual Matters.
- Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Philadelphia, U. S: Harcourt Brace Jovanovich College Publishers.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrical*, 16(3), 297-334.
- Cronbach, L. J & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
- Crystal, D. (2000). *English as global language*. Cambridge, UK: Cambridge University Press.
- Daniel, L. G., & Debra, A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *Journal of Educational Research*, 91(6), 331-344.

- Davies, A. (1965.). *Proficiency in English as a second language*. Birmingham,UK: University of Birmingham.
- Davies, A. (1978). Language testing: survey articles 1 and 2. *Language Teaching and Linguistic Abstracts*, 11, 145–59
- Davies, A. (1984). Validating three tests of English language proficiency. *Language Testing*, 1(1), 50–69.
- Dooley, P., & Oliver, T. R. (2002). An investigation into the predictive validity of the IELTS as an indicator of future academic success. *Prospect*, 17(1), 36–54.
- Educational Testing Service. (2006). *ETS guidelines for constructed-response and other performance assessments*. Princeton, New Jersey: Author.
- Educational Testing Service. (2007). *ETS international principles for fairness review of assessments*. Princeton, New Jersey: Author.
- Ebel, R. L., Frisbie, D. (1991). *Essentials of Educational Measurement (5th ed.)*. Englewood Cliffs, New Jersey: Prentice Hall.
- Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing*, 2, 68-85.
- Evans, J. S. B. T, & Over, D. E (1996). *Rationality and reasoning*. London, United Kingdom: Psychology Press.

- Fayeke, D. O., & Ogunsiji, Y. (2009). English language proficiency as a predictor of academic Achievement among EFL students in Nigeria. *European Journal of Scientific Research*. 37(3), 490-495.
- Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal*, 3(4), 70- 85.
- Ferrara (2007). Our Field Needs a framework to guide development of validity research Agendas and identification of validity research questions and threats to validity. *Measurement: Interdisciplinary Research and Perspectives*, 5(3), 156-164.
- Fernandes, H. J. X. (1984). *Testing and Measurement*. Jakarta, Indonesia: National Education Planning.
- Ferguson, G., & White, E. (1998). A small scale study in predictive validity. *Melbourne Papers in Language Testing*, 7, 15-63.
- Fives, H., & Didonato-Barnes, N. (2013). Using a table of specification to improve teacher – constructed traditional tests: an experimental design. *Assessment in Education: Principles, Policy and Practice*, 21(1), 90-108.
- Fulcher, G. (1997a). An English language placement test: issues in reliability and validity. *Language Testing*, 14(2), 113–139.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Boston, Massachusetts, U.S: Allyn & Bacon.

- Giri, R. A. (2011). Languages and language politics: How invisible language politics produces visible results in Nepal. *Language Problems and Language Planning*, 35(3), 197-221
- Ghenghesh, P. (2015). The relationship between English language proficiency and academic performance of university students. *International Journal of Applied Linguistics and English Literature*, 4(2), 91 - 97.
- Glass, G. V & Hopkins, K. D. (1984). *Test Bank for Statistical Methods in Education & Psychology, Second Edition*. Englewood Cliffs, New Jersey: Prentice-Hall.
- Government of the Socialist Republic of Vietnam. (2008). *National Foreign Language 2020 Project*. Article 1.1, Decision 1400.
- Government of the Socialist Republic of Vietnam. (2001). *Decree No14/2001/TC-TT on the Renovation of the Curriculum of Vietnam's General Education*.
- Graham, J. G. (1987). English language proficiency and prediction of academic success. *TESOL Quarterly*, 21(3), 505-521.
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications (2nd ed.)*. New York, U.S: Routledge
- Haladyna, T.M. (2006). Roles and importance of validity studies. In Downing, S. M., & Haladyna, T. M. (Eds.), *Handbook of Test Development* (pp. 739-755). Mahwah, New Jersey: LEA Press.

- Hancock, C. R. (2006). Language teaching and language testing: a way and ways to make appropriate connections. *Action Pedagogical*, 15, 6-13
- Harley, T. A. (2001) *The pschological of language*. New York: Psychology Press.
- Harrison, A. (1983). *Communicative testing: Jam tomorrow?* London, UK : Academic Press.
- Heaton, J. (1988). *Writing English language tests*. London, UK: Longman.
- Heaton, J. (1990). *Classroom Testing*. London, UK: Longman.
- Heitner, R. M., Hoekje, B. J., & Braciszewski, P. L. (2014). Tracking English language proficiency and IELTS test scores in an international undergraduate conditional admission program in the United States. In C. J. Linton & L. V. Amoroso (Eds.), *Measured language: Quantitative studies of acquisition, assessment and variation*. Washington, DC: Georgetown University Press.
- Henning, G. (1984). Current Developments in Language Testing. *Language Testing in Asia*, 1, 237–241.
- Henning, G. (1987). A guide to language testing: Development, evaluation, research. Cambridge, UK: Newbury House.
- Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *International English Language Testing System (IELTS) Research Report*, 2. 62-73.
- Hughes, A. (1989). *Testing for Language Teachers* (2²ⁿ ed). UK: Cambridge University Press.

- Hung, N.N (2013, March) *National Foreign Language 2020 Project: Challenges, Opportunities, and Solutions* . Paper presented at the Vietnam Engineering Education , Hanoi National University. <http://bruneiusprogramme.org/wp-content/uploads/2013-Forum-Publication-Complete.63-65.pdf>
- Huong, T. T. T. (2000). *The predictive validity of IELTS among Vietnamese students* (Unpublished doctoral thesis). La Trobe University, Australia.
- Huong, T. T. T. (2001). The predictive validity of International English Language Testing System (IELTS) Test. *Post-Script*, 2(1), 66-96.
- Huong, N. N. T., & Pham, H. D. (2010). Vietnam's teachers' and students' perceptions of global English. *Language Education in Asia*, 1(1), 48-61.
- Huy, N. T. (2015). Problems affecting learning writing skills of Grade 11 at Thong Linh high school. *Asian Journal of Educational Research*, 3(2), 1-17.
- Ing, L. M., Musah, M. B., Al-Hudawi, S. H. V., Tahir, L. M., & Kamil, N. M. (2015). Validity of teacher-made assessment: A table of specification approach. *Asian Social Science*, 11(5), 193-200.
- Iyer, S. M. (2015). Impact of wash back in English as Second Language Classrooms : An investigation in the University of Jaffna – Sri Lanka. *ELT Voice- International Journal for Teachers of English*, 5(3), 88-91.
- Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing*, 27(4), 555–584.

- Johnson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability and validity and educational consequences. *Educational Research Review*, 2(2007), 130-144.
- Kahsay, M. Y., Devardhi, J., & Nelson, D. (2014). Assessment of Content Validity in Ethiopian General Secondary Education . *Global Scholars Journal of Linguistics and Communication Studies*, 1(1), 1-13.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Eds.), *Educational measurement*, 4, 17-64.
- Kansopon, V. (2012). An investigation of the Writing test used at the Institute of International Studies Ramkhamhaeng University (II-RU), Thailand. *Language Testing in Asia*, 2(4), 1-15.
- Kappe, F. R., Boekholt, L., Den Rooyen, C., & Van der Flier, H. (2009). A predictive validity study of the learning style questionnaire (LSQ) using multiple, specific learning criteria. *Learning and Individual Differences*, 19(4), 464–467.
- Karr, C. (2011). *The attitude of teachers towards teaching reading in the content areas* (Unpublished master thesis). The University of the West Indies, U.S.
- Kerstijens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports*, 3, 85-108.

- Khalifa, H., & Weir, C. (2009). Examining Reading: Research and practice in assessing second language reading. *Journal of Studies in Language Testing*, 29, 276-280.
- Kindeya, N. T. (2002). *The content validity of the Ethiopian general secondary education certificate English examination* (Unpublished master's thesis). Addis Ababa University, Ethiopia.
- Kinyua, K., & Okunya, L. O. (2014). Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu District of Kenya. *African Educational Research Journal*, 2(2), 61-71.
- Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System*, 38(1), 63–74.
- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Kopriva, R. J. (2008). *Improving testing for English language learners*. New York: Routledge.
- Kuntasal, I. (2001). *Perceptions of teachers and testers of achievement tests prepared by testers in the Department of Basic English at the Middle East Technical University* (Unpublished master's thesis), Bilkent University, Turkey.
- Kluitmann, S. (2008). *Testing English as a foreign language: Two EFL tests used in Germany* (Unpublished master's thesis), University of Albert-Ludwig, Germany.

- Lado, R. (1961). *Language testing : The construction and use of foreign language tests : A teacher's book*. London: Longman.
- Lap, T. Q., & Truc, N, T. (2014). Enhancing Vietnamese learners' ability in writing argumentative essays. *The Journal of Asia TELF*, 11(2), 63-91.
- Lee, Y. J., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research*, 1(4), 366–389.
- Lewkowicz, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing*, 17(1), 43–64.
- Liao, Y. (2002). Issues of validity and reliability in second language performance assessment. *Language Testing*, 4(2), 2–5.
- Liauh, Y. H. (2011). *A study of the perceptions of English faculty and students of exit English examinations at Taiwan's Technological and Vocational Higher Education Institutions* (Unpublished doctoral thesis). The University of Montana, Missoula, U.S.
- Littlewood, W., Liu, N. F., & Yu, C. (1996). Hong Kong tertiary students' attitudes and proficiency in spoken English. *RELC Journal*, 27(1), 70–88.
- Loan, D. T. N. (2002). *Gender Issues in Education in Vietnam* (Unpublished doctoral thesis). National Institute for Education and Development, Vietnam.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635—694.

- Lofgren, R. E. (2005). *Validity evidence of multiple choice test and a performance test in employment setting* (Unpublished doctoral thesis). University of Pittsburgh, Pennsylvania, U.S.
- Maki, P. (2002). Using multiple assessment methods to explore student learning and development inside and outside of the classroom. *Net Results*. Retrieved from <http://www.naspa.org/>.
- Maleki, A., & Zangani, E. (2007). A survey on the relationship between English language proficiency and the academic achievement of Iranian EFL students. *Asian EFL Journal*, 9(1), 86-96.
- McNamara, T. (2000). *Language Testing*. Oxford, UK: Oxford University Press.
- McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3(1), 31-51.
- McKenzie, K., & Schweitzer, R. (2001). Who Succeeds at University?. Factors Predicting Academic Performance in First Year Australian University Students. *Higher Education Research and Development*, 20, 21-33.
- Messick, S. (1989). *Educational measurement*. New York, U.S: Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance and assessment. *Educational Researcher*, 23(2), 13-25

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Minh, N. T. T. (2008). Textbooks evaluation: The case of English textbooks currently used in Vietnam's upper secondary school (Unpublished doctoral thesis). National University Hanoi, Vietnam.
- Mohammadi, M. (2009). The predictive validity of Islamic Azad University's entrance examination: Does access mean success. *The Journal of Modern Thoughts in Education*, 4, 59-72.
- Montee, M., Bach, A., Donovan, A., & Thompson, L. (2013). Teachers' assessment knowledge and practices: An exploratory study. *Journal of the National Council of Less Commonly Taught Languages*, 13, 1-31.
- Moore, K. (1992). *Classroom testing skills*. New York, U.S: McGraw Hill, Inc
- Moritoshi, T. P. (2002). *Validation of the test of English conversation proficiency* (Unpublished doctoral thesis). University of Birmingham, UK.
- Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the criterion of graduate success examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement*, 55, 309-316.
- Nakamura, Y. (1997). Establishing construct validity of an English speaking test. *Journal of Communication*, 6, 13-30.

- Neena, D., & Dash, M. (2007). *Teaching English as an additional language*. New Delhi, India: Atlantic Publishers & Distributors(p) Ltd.
- Nha, T. T. Vu., & Burns, A. (2014). English as a medium of instruction: Challenges for Vietnamese tertiary lecturers. *The Journal of Asia TEFL*, 11(3), 1-31.
- Nguyen, B., & Crabbe, D. (2007). The design and use of English language textbooks in Vietnamese secondary schools. *Proceedings of the Fourth International Conference on Language and Development* (pp. 13-15). Hanoi, Vietnam.
- Nguyen, T. H. H. (2009). Teaching EFL writing in Vietnam: Problems and solutions – a discussion from the outlook of applied linguistics. *VNU Journal of Science*, 25(1), 61-66
- Notar, C. E., Zuelke, D. C., Wilson, J. D. & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology*, 31, 115-129.
- Nwagu, E. K. N. (2009). Content validity of teacher-made geography tests used in secondary schools in Anambra and Enugu States of Nigeria. *Social Science Review*, 2(1). 109-115.
- Oller, J. W. J. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in ESL. *Modern Language Journal*, 56, 151–158.
- Oller, J. W. J. (1997). *Language test at school*. London, UK: Longman.

- Omo-Egbekuse, J., Afemikhe, O.A., & Imobekhai, S. Y. (2012). Nigerian teachers' ratings of their competence in standards for educational assessment of students. *Benin Journal of Educational Studies*, 21(1), 1- 11.
- O'Sullivan, B., Weir, C. & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, 19(1), 33-56.
- O'Sullivan, B. (2006). *Issues in Testing Business English: The BEC Revision Project, Studies in Language Testing 17*. Cambridge, UK: Cambridge University Press.
- O'Sullivan, B. (2011). *Language Testing*. Oxford, UK: Routledge.
- O'Sullivan, B., & Weir, C. (2011). *Language testing and validation*. Oxford, UK: Palgrave.
- Othman, J., & Nordin, N. (2013). MUET as a predictor of academic achievement in ESL teacher education. *GEMA Online Journal of Language Studies*, 13(1), 99-111.
- Pagano, R. R. (2001). *Understanding Statistics in the Behavioral Sciences*. California, U.S: Wadsworth.
- Pallant, J. (2013). *SPSS survival manual*. UK: McGraw-Hill Education.
- Parhizgar, S. (2012). Testing and technology: Past, Present and Future. *Theory and Practices in Language Studies*, 2(1), 174- 178.
- Pishghadam, R., & Khosropanah, F. (2011). Predictive validity of the English language teacher competency test. *International Journal of Education*, 3(1), 1-19.
- Popham, W. J. (2001). Teaching to the test. *Educational Leadership*, 58(6), 16-20.

- Popham, W.J. (2002). *Classroom Assessment, What Teachers need to know*. Boston: Allyn & Bacon.
- Ramist, L., Lewis, C., & McCauley, J. C. (2002). Validity of the SAT II Science Tests. *Science Insights*, 6(5) 25-29. Retrieved from <http://www.nas.org>.
- Rethinasamy, S., & Chuah, K. M. (2011). The Malaysian University English Test (MUET) and its use for placement purposes: A Predictive validity study. *Electronic Journal of Foreign Language Teaching*, 8(2), 234–24.
- Rethinasamy, S. (2005). *The effects of different rater training procedures on ESL Essay Rating Judgement* (Unpublished doctoral thesis). Roehampton University, England.
- Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic. *Language Testing and Assessment*, 2(2), 1–27.
- Roche, T., & Harrington, M. (2013). Recognition vocabulary skill as a predictor of academic English performance and academic achievement in English. *Language Testing in Asia*, 3(12), 133-157.
- Rudman, H. C. (1989). Integrating testing with teaching. *Practical assessment, research and evaluation*, 1(6), 1-4. Retrieved April 16, 2014 from <http://Pareonline.net/getvn.asp?v=1&n=6>
- Sadeghi, B., Kashanian, N. M., Maleki, A., & Haghdoost, A. (2013). English language proficiency as a predictor of academic achievement among medical students in Iran. *Theory and Practice in Language Studies*, 3(12), 2315-2321.

- Sahragard, R., Baharloo, A., & Soozandehfar, S. M. A. (2011). A closer look at the relationship between academic achievement and language proficiency among Iranian EFL Students. *Theory and Practice in Language Studies*, 1(12), 1740-1748.
- Samad, A.A., Rahman, S. Z. S., & Yahya, S.N. (2008). Refining English language tests for university admission: A Malaysian example. *Asian Journal University of Education*, 4(1). 57–68.
- Saville, N. (2009). *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider* (Unpublished doctoral thesis). University of Bedfordshire, England.
- Schoepp, K., & Garinger, D. (2016). IELTS and academic success in higher education: A UAE perspective. *International Journal of Applied Linguistics and English Literature*, 5(3), 145-153.
- Shaw, S. & C. J. Weir. (2007). *Examining writing: Research and practice in assessing second language writing*. *Studies in language testing* 26. Cambridge: Cambridge University Press.
- Shih, C. (2010). The washback of the general English proficiency test on University Policies: A Taiwan Case Study. *Language Assessment Quarterly*, 7(3), 234–254.
- Siddiek, A. G. (2010). The impact of test content validity on language teaching and learning, Shaqra University, *Asian Social Science Journal*, 6(12), 133-140.

- Sim, J. M. (2015). A valid and reliable English proficiency exam: A model from a University Language Program in Taiwan. *English as a Global Language Education (EaGLE) Journal*, 1(2), 91-125.
- Sivaramana, I., Al Balushib, A., & Rao, D. H. (2014). Understanding Omani students' (University) English language problems. *International Journal of Sciences: Basic and Applied Research*, 13(1), 28-35.
- So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly*, 11(3), 283–303.
- Social research method website. (2006). Illustrations of relationship between reliability and validity . Retrieved from <http://www.socialresearchmethods.net/kb/relandval.php>.
- Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, 9 (1), 5–29.
- Stevenson, R. (1994). *Global communication in the twenty-first century*. New York, U.S: Longman.
- Stevenson, D. K. (1985). Authenticity, validity and a tea party. *Language Testing*, 2(1), 41-47.
- Straus, M. A., Douglas, E., & Medeiros, R. A. (2004). *The primordial violence: Spanking Children, Psychological Development, Violence and Crime* . New York, U.S: Alta Mira Press.

- Szombatová, V. (2016). *The semi-structured in foreign language education research*. Proceedings of The International Conference on Language and Literature in Education and Research. Prague.
- Tatsuoka, M. M. (1997). Regression analysis of quantified data. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An international handbook* (pp. 648-657). Cambridge University Press, Cambridge.
- Thiel, T. (1995). *An analysis of the evolution of the IELTS test and an investigation of its face validity* (Unpublished doctoral thesis). University of Tasmania, U.S.
- Tran, H. P., Griffina, P., & Nguyen, C. (2010). Validating the university entrance English test to Vietnam National University: A conceptual framework and methodology. *Science Direct*, 2(2010), 1295 -1304.
- Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653-673). Mahwah, New Jersey: Erlbaum.
- Ughamadu, K. A. Onwuegbu, O. C. & Osunde, A. U. (1991). *Measurement of Evaluation in Education*. Nigeria: World of Book Publishers.
- Van der Walt, J.L. & Steyn, H. S. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics*, 38(2), 191-204.

- Van, H. V. (2007). The current situation and issues of the teaching of English in Vietnam, *Teaching English to speakers of other languages in the internationalization of higher education in Vietnam*, 3(4),1-12.
- Weir, C. (1993). *Understanding and developing language tests*. New York, U.S: Prentice Hall.
- Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.
- Weir, C. J., Chan, S. H. C., & Nakatsuhara, F. (2013). Examining the criterion-related validity of the GEPT advanced reading and writing tests: comparing GEPT with IELTS and real-life academic performance'. *LTTC-GEPT Research Report*, 1, 1-43.
- Wiersma, W., & Stephen, G. J. (1990). *Educational Measurement and Testing* (2nd ed). Boston, U.S: Allyn and Bacon.
- Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing*, 15(2), 263-287.
- Wilson, K. M. (1999). *Validating a test designed to assess ESL proficiency at lower developmental levels*, Princeton, New Jersey: Educational Testing Service.
- Wilson, R. G. (2016). *The PTE Academic and outer circle students: Assessing proficiency in English, ownership of English, and academic performance at UK Universities* (Unpublished doctoral thesis). University of Warwick, England.

- Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, 45(4), 628–660.
- Woodrow, L. (2006). Academic success of international postgraduate education students and the role of English proficiency. *University of Sydney papers in TESOL*, 1, 51-70.
- Wu, J. (2005). *Task difficulty in semi-direct speaking tests* (Unpublished doctoral thesis). Reading University, England.
- Wu, R. Y. F (2011). *Establishing the validity of the general English proficiency test reading component through a critical evaluation on alignment with the Common European Framework of Reference* (Unpublished doctoral thesis). University of Bedfordshire, England.
- Xi, X. (2008). *Encyclopedia of Language and Education : Methods of test validation*. Princeton: Springer.
- Yanagawa, K. (2012). *A partial validation of the contextual validity of the centre listening test in Japan* (Unpublished doctoral thesis). University of Bedfordshire, England.
- Yen, D., & Kuzma, J. (2009). Higher IELTS Score, Higher Academic Performance? The Validity of IELTS in Predicting the Academic Performance of Chinese Students. *Worcester Journal of Learning and Teaching*, 4(3). 1-7.
- Yixin, W., & Daller, M. (2015). *Predicting International Chinese Students Academic Success*. Proceedings of 47th Annual Conference of the British Association for Applied Linguistics (BAAL), University of Warwick, Coventry.

APPENDICES

Appendix 1: Letter of Request for AEP students 'test scores (AEPET, IELTS and CGPA)



**UNIVERSITI MALAYSIA
SARAWAK
Centre for Language Studies
94300 Kota Samarahan
Sarawak, Malaysia**

PERMISSION LETTER TO COLLECT DATA FOR RESEARCH

January, 2015

Dear Sir/Madam:

Warmest Greetings!

I would like to ask your permission to allow me to collect students' Advanced Educational Program English Test (AEPET) and International English Language Test System (IELTS) scores, cumulative Grade Point Average (CGPA) from your University Database. This is one part of my thesis, entitled, "Investigating the Validity of the AEPET at [REDACTED] [REDACTED]". The purpose of this collection is to examine the relationship between students' AEPET and IELTS scores as well as to determine to what extent the AEPET can predict students' academic achievement measured by their CGPA

All information provided will be kept in utmost confidentiality and would be used only for academic purposes. After the data have been analyzed, you will receive a copy of the executive

summary. If you would be interested in greater detail, an electronic copy (e.g. PDF) of the entire thesis can be made available to you.

If you agree, kindly sign below acknowledging your consent and permission for me to collect students' test scores at your University and return the signed form on an enclosed envelope.

Your approval to conduct this study will be greatly appreciated. Thank you in advance for your interest and assistance with this research.

Sincerely

Nong Thi Hien Huong
Ph.D Student, UNIMAS

Noted by

Asso. Prof. Dr. Soubakeavathi Rethinasamy
Research Supervisor, UNIMAS

Approved by



Appendix 2: Letter of Request for conducting a survey questionnaire



**UNIVERSITI MALAYSIA
SARAWAK**
Centre for Language Studies
94300 Kota Samarahan
Sarawak, Malaysia

PERMISSION LETTER TO CONDUCT RESEARCH

February , 2017

Dear Sir/Madam:

Warmest Greetings!

I would like to ask your permission to allow me to conduct a survey questionnaire and interview among the English language lecturers in your University. This is the title of my thesis, entitled, “Investigating the Validity of the AEPET at [REDACTED] [REDACTED]. The purpose of this collection is to examine the content validity of the AEPET to determine to what extent the whether the AEPET content reflects the knowledge/ skills required to do in the AEP Course syllabus

The survey would last 20-30 minutes and would be arranged at a time convenience to the lecturers’ schedule (e.g. during break). Participation in the survey is entirely voluntary and there are no anticipated risks to participation in this study. All information provided will be kept in

utmost confidentiality and would be used only for academic purposes. After the data have been analyzed, you will receive a copy of the executive summary. If you would be interested in greater detail, an electronic copy (e.g. PDF) of the entire thesis can be made available to you.

If you agree, kindly sign below acknowledging your consent and permission for me to conduct this survey at your school and return the signed form on an enclosed envelope. Your approval to conduct this study will be greatly appreciated. Thank you in advance for your interest and assistance with this research.

Sincerely

Nong Thi Hien Huong
Ph.D Student, UNIMAS

Noted by

Asso. Prof. Dr. Soubakeavathi Rethinasamy
Research Supervisor, UNIMAS

Approved by



Appendix 3: Content Validity Questionnaire

Date:

Dear Participant,

I am a postgraduate student at University Malaysia Sarawak, studying Ph.D. in Applied Linguistics at the Centre for Language Studies. My PhD research focuses on the content validity of the Advanced Educational Program English Test (AEPET) at a public university in Vietnam.

I would like to invite you to kindly participate in this study by completing the enclosed questionnaire. The survey should take about 10-15 minutes to complete. Your participation in this study is voluntary and you are free to withdraw your participation from this study at any time. There are no risks associated with participating in this study. All responses in the survey will be recorded anonymously and used for the purpose of this research only.

If you agree to participate in this study, please respond to the items in the questionnaire. Your honest opinions are very crucial for this research.

If you have any questions regarding the survey or the research in general, please feel free to contact Nong Thi Hien Huong at hhuong04052002@yahoo.com

Thank you for your cooperation and time.

Sincerely,

Nong Thi Hien Huong
Faculty of Language and Communication
University Malaysia Sarawak
Email: hhuong04052002@yahoo.com
Cellphone number: +84.984.888.345

Section A : Background Information				
For Section A, please fill up the space provided with relevant information.				
1.	Full Name	:		
2.	Gender	:	☐ Male	☐ Female
3.	Age in 2017	:		
4.	Email	:		
5.	Cellphone number	:		
6.	What is your highest academic qualification?	:	☐ Ph.D. ☐ Master Degree ☐ Bachelor Degree	
7.	How long have you been teaching English?	:	☐ Less than 5 years ☐ 5 to 10 years ☐ 11 to 15 years ☐ More than 15 years	
8	What English skills do you teach? (You can tick more than one option)	:	☐ Listening ☐ Reading ☐ Writing ☐ Speaking	
9	Have you attended course on language testing	:	☐ Yes ☐ No	
10	If Yes in Question 9, please specify The training course was on	:	☐ Listening ☐ Reading ☐ Writing ☐ Speaking	

Section B: AEPET Components

This section aimed to gather information on the AEPET components which are Listening, Reading, Writing and Speaking. A summary of the AEPET format and weighting is presented in this table below.

Test Components	Time allowance (Minutes)	Sections/ Parts	Questions	Maximum Possible Scores	Weightage (%)
Listening	30	4	50	10	20
Reading	60	4	50	10	20
Writing	60	2	2	10	40
Speaking	11-14	3	20	10	20

From questions 11 to 38, please circle only **ONE** of the options given

SD D N A A
Strongly disagree Disagree Neutral Agree Strongly agree

Question 11 - 17: Listening Test

11	Weightage for the listening test was appropriate.	SD	D	N	A	SA
12	Time allocation of the listening test was sufficient.	SD	D	N	A	SA
13	Listening skills taught were sufficiently represented in the listening test.	SD	D	N	A	SA
14	Topics taught were sufficiently represented in the listening test.	SD	D	N	A	SA
15	Questions in the listening test were clear.	SD	D	N	A	SA
16	Instructions explaining what to do in each section of the listening test were clear	SD	D	N	A	SA
17	Marks allocated for each section of the listening test were stated clearly.	SD	D	N	A	SA

Question 18- 24: Reading Test

18	Weightage for the reading test was appropriate.	SD	D	N	A	SA
19	Time allocation of the reading test was sufficient.	SD	D	N	A	SA
20	Reading skills taught were sufficiently represented in the reading test.	SD	D	N	A	SA
21	Topics taught were sufficiently represented in the reading test.	SD	D	N	A	SA
22	Questions in the reading test were clear.	SD	D	N	A	SA
23	Instructions explaining what to do in each section of the reading test were clear	SD	D	N	A	SA
24	Marks allocated for each section of the reading test were stated clearly.	SD	D	N	A	SA

Question 25- 31: Writing Test

25	Weightage for the writing test was appropriate.	SD	D	N	A	SA
26	Time allocation of the writing test was sufficient.	SD	D	N	A	SA
27	Writing skills taught were sufficiently represented in the writing test.	SD	D	N	A	SA
28	Topics taught were sufficiently represented in the writing test.	SD	D	N	A	SA
29	Questions in the writing test were clear.	SD	D	N	A	SA
30	Instructions explaining what to do in each section of the writing test were clear.	SD	D	N	A	SA
31	Marks allocated for each section of the writing test were stated clearly.	SD	D	N	A	SA

Question 32- 38: Speaking Test

32	Weightage for the speaking test was appropriate.	SD	D	N	A	SA
----	--	----	---	---	---	----

33	Time allocation of the speaking test was sufficient.	SD	D	N	A	SA
34	Speaking skills taught were sufficiently represented in the speaking test.	SD	D	N	A	SA
35	Topics taught were sufficiently represented in the speaking test.	SD	D	N	A	SA
36	Questions in the speaking test were clear.	SD	D	N	A	SA
37	Instructions explaining what to do in each section of the speaking test were clear	SD	D	N	A	SA
38	Marks allocated for each section of the speaking test were stated clearly.	SD	D	N	A	SA

Section C: Test Preparation

This section aims to gather information on the process of AEPET preparation encompasses the four components: Listening, Reading, Writing and Speaking.

From questions 39 to 63, please circle only **ONE** of the options given

SD D N A A
 Strongly disagree Disagree Neutral Agree Strongly agree

39	I am given information about AEPET (test specification) before constructing test items.	SD	D	N	A	SA
40	When constructing the test items, I refer to the test specification .	SD	D	N	A	SA
41	When constructing the test items, I refer to the test specification to determine the topics to be covered.	SD	D	N	A	SA
42	When constructing the test items, I refer to the test specification to determine the time allocation .	SD	D	N	A	SA
43	When constructing the test items, I refer to the test specification to determine the number of test items .	SD	D	N	A	SA
44	When constructing the test items, I refer to the test specification to determine the difficulty level .	SD	D	N	A	SA
45	When constructing the test items, I refer to the test specification to determine the marks allocation (weightage) .	SD	D	N	A	SA
46	I am given sample AEPET papers before constructing test items	SD	D	N	A	SA
47	When constructing the test items, I refer to the sample test paper to determine the topics to be covered.	SD	D	N	A	SA
48	When constructing the test items, I refer to the sample test paper to determine the time allocation .	SD	D	N	A	SA
49	When constructing the test items, I refer to the sample test paper to determine the number of test items .	SD	D	N	A	SA
50	When constructing the test items, I refer to the sample test paper to determine the difficulty level .	SD	D	N	A	SA
51	When constructing the test items, I refer to the sample test paper to determine the marks allocation (weightage) .	SD	D	N	A	SA
52	I refer to English course syllabus when constructing the test items	SD	D	N	A	SA
53	When constructing the test items, I refer to English course syllabus to determine the topics to be covered.	SD	D	N	A	SA
54	When constructing the test items, I refer to English course syllabus to determine the time allocation .	SD	D	N	A	SA
55	When constructing the test items, I refer to English course syllabus to	SD	D	N	A	SA

	determine the number of test items .					
56	When constructing the test items, I refer to English course syllabus to determine the difficulty level .	SD	D	N	A	SA
57	When constructing the test items, I refer to English course syllabus to determine the marks allocation (weightage) .	SD	D	N	A	SA
58	The answer keys for the objective questions are prepared during the item construction.	SD	D	N	A	SA
59	The marking scheme (rubrics) is prepared before the examination is administered.	SD	D	N	A	SA
60	The questions for the test are vetted in a meeting.	SD	D	N	A	SA
61	The answer keys for the objective questions are vetted in a meeting.	SD	D	N	A	SA
62	The marking scheme (rubrics) is vetted in a meeting.	SD	D	N	A	SA
63	I attend rater training session (moderation) before assessing the answer scripts assigned to me.	SD	D	N	A	SA

Thank you for your cooperation!

Appendix 4: Sample AEPET Listening Test

ADVANCED EDUCATIONAL PROGRAM ENGLISH TEST

Subject: Listening

Place:

Time: 30'

Date:

Test code: 127

Full name:..... Candidate No:.....

INSTRUCTIONS

Do not open the question paper until you are told to do so

Write your name and your candidate number on your answer sheet if they are not already there

Read **ALL** the instructions carefully

Answer **ALL** the questions

Read the instructions on the answer sheets

Write your answers on the answer sheets. Use a pencil

The recordings will be heard only **ONCE**

You must complete the answer sheets within the time limit

INFORMATION

Each correct answer carries two marks

Do not accept misspelling

SECTION 1 Questions 1-10

Complete the notes. Use **NO MORE THAN THREE WORDS** for each answer.

Example	Answer
Name of agent:	Flagstone
Areas dealt with:	1 north suburbs
Rent:	from 2 £ to £ a month
Depends on:	the area availability of 3 garage
Properties available:	West Park Road rent 4 £ a month including 5 Tithe Road rent £380 a month including 6 rental
Viewing arrangements:	meet at office on 7 at 5.00 pm
Need:	letter from bank reference from 8
Must:	give 9 notice of moving in give deposit of 10 pay for contract

SECTION 3 Questions 21-32

Questions 21-24

Circle the correct answer.

- 21** At first Fiona thinks that Martin's tutorial topic is
- A** inappropriate.
 - B** dull.
 - C** interesting.
 - D** fascinating.
- 22** According to Martin, the banana
- A** has only recently been cultivated.
 - B** is economical to grow.
 - C** is good for your health.
 - D** is his favourite food.
- 23** Fiona listens to Martin because she
- A** wants to know more about bananas.
 - B** has nothing else to do today.
 - C** is interested in the economy of Australia.
 - D** wants to help Martin.
- 24** According to Martin, bananas were introduced into Australia from
- A** India.
 - B** England.
 - C** China.
 - D** Africa.

SECTION 2 Questions 11-20

Complete the notes below. Use **NO MORE THAN THREE WORDS** for each answer.

There are many kinds of bicycles available:

racing

touring

(11)

ordinary



They vary in price and (12)

Prices range from \$50.00 to (13)

Single speed cycles are suitable for (14)

Three speed cycles are suitable for (15)

Five and ten speed cycles are suitable for longer distances, hills and (16)

Ten speed bikes are better because they are (17) in price but (18)

Buying a cycle is like (19)

The size of the bicycle is determined by the size of the (20)

SECTION 3 *Questions 21-32*

Questions 21-24

Circle the correct answer.

- 21** At first Fiona thinks that Martin's tutorial topic is
- A** inappropriate.
 - B** dull.
 - C** interesting.
 - D** fascinating.
- 22** According to Martin, the banana
- A** has only recently been cultivated.
 - B** is economical to grow.
 - C** is good for your health.
 - D** is his favourite food.
- 23** Fiona listens to Martin because she
- A** wants to know more about bananas.
 - B** has nothing else to do today.
 - C** is interested in the economy of Australia.
 - D** wants to help Martin.
- 24** According to Martin, bananas were introduced into Australia from
- A** India.
 - B** England.
 - C** China.
 - D** Africa.

Questions 25-30

Complete Martin's notes Use **NO MORE THAN THREE WORDS** for each answer.

Commercially grown
banana plant



Each banana tree produces
(25)
of bananas.

On modern plantations in tropical
conditions a tree can bear fruit after
(26)

Banana trees prefer to grow (27) and they require
rich soil and (28) The fruit is often protected by
(29)

Ripe bananas emit a gas which helps other (30)

Questions 31 and 32

Circle the **TWO** correct boxes.

Consumption of Australian bananas



A Europe

B Asia

C New Zealand

D Australia

E Other

SECTION 4 Questions 33-41**Questions 33-35***Circle the correct answer*

According to the first speaker:

33 The focus of the lecture series is on

- A** organising work and study.
B maintaining a healthy lifestyle.

- C** coping with homesickness.
D settling in at university.

34 The lecture will be given by

- A** the president of the Union.
B the campus doctor.

- C** a sports celebrity.
D a health expert.

According to the second speaker:

35 This week's lecture is on

- A** campus food.
B dieting.

- C** sensible eating.
D saving money.

Questions 36-39*Complete the notes. Write NO MORE THAN THREE WORDS for each answer.***A balanced diet**

A balanced diet will give you enough vitamins for normal daily living.

Vitamins in food can be lost through **(36)**

Types of vitamins:

(a) Fat soluble vitamins are stored by the body.

(b) Water soluble vitamins - not stored, so you need

a **(37)****Getting enough vitamins**Eat **(38)** of foods.

Buy plenty of vegetables and store them in

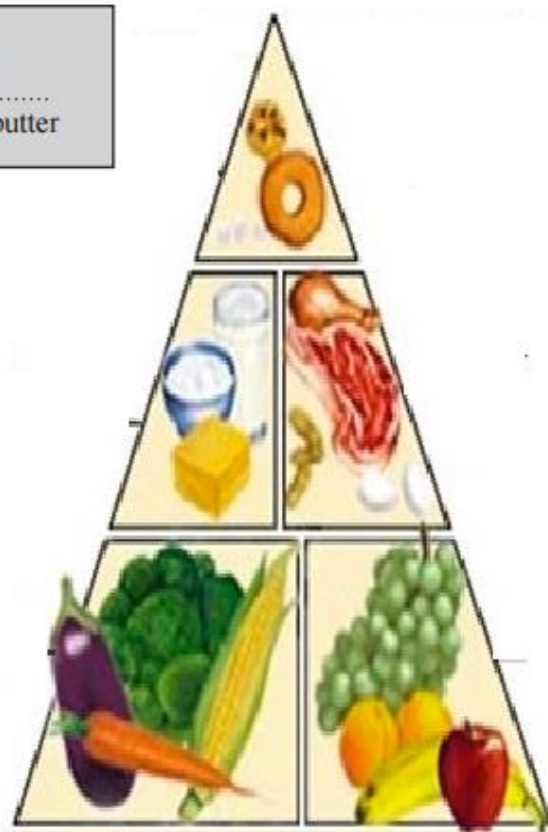
(39)

Questions 40-41

Complete the diagram by writing **NO MORE THAN THREE WORDS** in the boxes provided.

Example
Try to avoid
.....
sugar, salt and butter

40
.....
milk, lean meat, fish,
nuts, eggs



41
.....
bread, vegetables and
fruit

SECTION 5 Questions 42-50*Listening**Complete the notes. Write NO MORE THAN THREE WORDS for each answer.*

KATE	
Her first impressions of the town	<i>Example Quiet</i>
Type of accommodation	(42)
Her feelings about the accommodation	(43)
Her feelings about the other students	(44)
Name of course	<i>Environmental Studies</i>
Difficulties experienced on the course	(45)
Suggestions for improving the course	(46)
LUKI	
First type of accommodation	(47)
Problem with the first accommodation	(48)
Second type of accommodation	(49)
Name of course	(50)
Comments about the course	<i>Computer room busy</i>

The End

[illegible]

Then write your six digit Candidate number in the boxes and shade the number in the grid on the right.

Day: 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Month: 01 02 03 04 05 06 07 08 09 10 11 12 Year (last 2 digits): 09 10 11 12 13 14 15 16 17 18

		Mariner use only			Mariner use only
1		☐ ☐	26		☐ ☐
2		☐ ☐	27		☐ ☐
3		☐ ☐	28		☐ ☐
4		☐ ☐	29		☐ ☐
5		☐ ☐	30		☐ ☐
6		☐ ☐	31		☐ ☐
7		☐ ☐	32		☐ ☐
8		☐ ☐	33		☐ ☐
9		☐ ☐	34		☐ ☐
10		☐ ☐	35		☐ ☐
11		☐ ☐	36		☐ ☐
12		☐ ☐	37		☐ ☐
13		☐ ☐	38		☐ ☐
14		☐ ☐	39		☐ ☐
15		☐ ☐	40		☐ ☐
16		☐ ☐	41		☐ ☐
17		☐ ☐	42		☐ ☐
18		☐ ☐	43		☐ ☐
19		☐ ☐	44		☐ ☐
20		☐ ☐	45		☐ ☐
21		☐ ☐	46		☐ ☐
22		☐ ☐	47		☐ ☐
23		☐ ☐	48		☐ ☐
24		☐ ☐	49		☐ ☐
25		☐ ☐	50		☐ ☐

Listening
Total

Appendix 5: Sample AEPET Reading Test

READING

READING PASSAGE 1

You should spend about 20 minutes on Questions 1-13, which are based on Reading Passage 1 on the following pages.

Questions 1-5

Reading Passage 1 has five marked paragraphs. A-E.

Choose the correct heading for each paragraph from the list of headings below.

Write the correct number, i-viii, in boxes 1-5 on your answer sheet

List of headings

- i Avoiding an overcrowded centre
- ii A successful exercise in people power
- iii The benefits of working together in cities
- iv Higher incomes need not mean more cars
- v Economic arguments fail to persuade
- vi The impact of telecommunications on population distribution
- viii Responding to arguments against public transport

- 1 Paragraph A
- 2 Paragraph B
- 3 Paragraph C
- 4 Paragraph D
- 5 Paragraph E

Advantages of public transport



A new study conducted for the World Bank by Murdoch University's Institute for Science and Technology Policy (ISTP) has demonstrated that public transport is more efficient than cars. The study compared the proportion of wealth poured into transport by thirty-seven cities around the world. This included both the public and private costs of building, maintaining and using a transport system.

The study found that the Western Australian city of Perth is a good example of a city with minimal public transport. As a result, 17% of its wealth went into transport costs. Some European and Asian cities, on the other hand, spent as little as 5%. Professor Peter Newman, ISTP Director, pointed out that these more efficient cities were able to put the difference into attracting industry and jobs or creating a better place to live.

According to Professor Newman, the larger Australian city of Melbourne is a rather unusual city in this sort of comparison. He describes it as two cities: 'A European city surrounded by a car-dependent one'. Melbourne's large tram network has made car use in the inner city much lower, but the outer suburbs have the same car-based structure as most other Australian cities. The explosion in demand for accommodation in the inner suburbs of Melbourne suggests a recent change in many people's preferences as to where they live.

Newman says this is a new, broader way of considering public transport issues. In the past, the case for public transport has been made on the basis of environmental and social justice considerations rather than economics. Newman, however, believes the study demonstrates that 'the auto-dependent city model is inefficient and grossly inadequate in economic as well as environmental terms'.

Bicycle use was not included in the study but Newman noted that the two most 'bicycle friendly'¹ cities considered - Amsterdam and Copenhagen - were very efficient, even though their public transport systems were 'reasonable but not special'.

It is common for supporters of road networks to reject the models of cities with good public transport by arguing that such systems would not work in their particular city. One objection is climate. Some people say their city could not make more use of public transport because it is either too hot or too cold. Newman rejects this, pointing out that public transport has been successful in both Toronto and Singapore and, in fact, he has checked the use of cars against climate and found 'zero correlation'.

When it comes to other physical features, road lobbies are on stronger ground. For example, Newman accepts it would be hard for a city as hilly as Auckland to develop a really good rail network. However, he points out that both Hong Kong and Zurich have managed to make a success of their rail systems, heavy and light respectively, though there are few cities in the world as hilly.

- A in fact Newman believes the main reason for adopting one sort of transport over another is politics. The more democratic the process, the more public transport is favored. He considers Portland Oregon, a perfect example of this. Some years ago federal money was granted to build a new road. However, local pressure groups forced a referendum over whether to spend the money on light rail instead. The rail proposal won and the railway worked spectacularly well. In the years that have followed, more and more rail systems have been put in, dramatically changing the nature of the city. Newman notes that Portland has about the same population as Perth and had a similar population density at the time.
- B In the UK, travel times to work had been stable for at least six centuries, with people avoiding situations that required them to spend more than half an hour travelling to work. Trains and cars initially allowed people to live at greater distances without taking longer to reach their destination. However, public infrastructure did not keep pace with urban sprawl, causing massive congestion problems which now make commuting times far higher.
- C There is a widespread belief that increasing wealth encourages people to live farther out where cars are the only viable transport. The example of European cities refutes that. They are often wealthier than their American counterparts but have not generated the same level of car use. In Stockholm, car use has actually fallen in recent years as the city has become larger and wealthier. A new study makes this point even more starkly. Developing cities in Asia, such as Jakarta and Bangkok, make more use of the car than wealthy Asian cities such as Tokyo and Singapore. In cities that developed later, the World Bank and Asian Development Bank discouraged the building of public transport and people have been forced to rely on cars -creating the massive traffic jams that characterize those cities.
- D Newman believes one of the best studies on how cities built for cars might be converted to rail use is *The Urban Village* report, which used Melbourne as an example. It found that pushing everyone into the city centre was not the best approach. Instead, the proposal advocated the creation of urban villages at hundreds of sites, mostly around railway stations.
- E It was once assumed that improvements in telecommunications would lead to more dispersal in the population as people were no longer forced into cities. However, the ISTP team's research demonstrates that the population and job density of cities rose or remained constant in the 1980s after decades of decline. The explanation for this seems to be that it is valuable to place people working in related fields together. 'The new world will largely depend on human creativity, and creativity flourishes where people come together face-to-face.'

Questions 6-10

Do the following statements agree with the information given in Reading Passage 1?

In boxes 6-10 on your answer sheet, write

TRUE if the statement agrees with the information

FALSE if the statement contradicts the information

NOT GIVEN if there is no information on this

- 6 The ISTP study examined public and private systems in every city of the world.
- 7 Efficient cities can improve the quality of life for their inhabitants.
- 8 An inner-city tram network is dangerous for car drivers.
- 9 In Melbourne, people prefer to live in the outer suburbs.
- 10 Cities with high levels of bicycle usage can be efficient even when public transport is only averagely good.

Questions 11-13

Look at the following cities (Questions 11-13) and the list of descriptions below.

Match each city with the correct description, A-F.

Write the correct letter, A-F, in boxes 11-13 on your answer sheet.

- 11 Perth
- 12 Auckland
- 13 Portland

List of Descriptions

- A successfully uses a light rail transport system in hilly environment
- B successful public transport system despite cold winters
- C profitably moved from road to light rail transport system
- D hilly and inappropriate for rail transport system
- E heavily dependent on cars despite widespread poverty
- F inefficient due to a limited public transport system

READING PASSAGE 2

You should spend about 20 minutes on Questions 14-26, which are based on Reading Passage 2 below.

GREYING POPULATION STAYS IN THE PINK

Elderly people are growing healthier, happier and more independent, say American scientists. The results of a 14-year study to be announced later this month reveal that the diseases associated with old age are afflicting fewer and fewer people and when they do strike, it is much later in life.

In the last 14 years, the National Long-term Health Care Survey has gathered data on the health and lifestyles of more than 20,000 men and women over 65. Researchers, now analysing the results of data gathered in 1994, say arthritis, high blood pressure and circulation problems - the major medical complaints in this age group - are troubling a smaller proportion every year. And the data confirms that the rate at which these diseases are declining continues to accelerate. Other diseases of old age - dementia, stroke, arteriosclerosis and emphysema - are also troubling fewer and fewer people.

'It really raises the question of what should be considered normal ageing,' says Kenneth Manton, a demographer from Duke University in North Carolina. He says the problems doctors accepted as normal in a 65-year-old in 1982 are often not appearing until people are 70 or 75.

Clearly, certain diseases are beating a retreat in the face of medical advances. But there may be other contributing factors. Improvements in childhood nutrition in the first quarter of the twentieth century, for example, gave today's elderly people a better start in life than their predecessors,

On the downside, the data also reveals failures in public health that have caused surges in some illnesses. An increase in some cancers and bronchitis may reflect changing smoking habits and poorer air quality, say the researchers. These may be subtle influences 'says Manton, 'but our subjects have been exposed to worse and worse pollution for over 60 years. It's not surprising we see some effect.'

One interesting correlation Manton uncovered is that better-educated people are likely to live longer. For example, 65-year-old women with fewer than eight years of schooling are expected on average, to live to 82. Those who continued their education live an extra seven years. Although some of this can be attributed to a higher income, Manton believes it is mainly because educated people seek more medical attention.

The survey also assessed how independent people over 65 were, and again found a striking trend. Almost 80% of those in the 1994 survey could complete everyday activities ranging from eating and dressing unaided to complex tasks such as cooking and managing their finances. That represents a significant drop in the number of disabled old people in the population. If the trends apparent in the United States 14 years ago had continued,

researchers calculate there would be an additional one million disabled elderly people in today's population. According to Manton, slowing the trend has saved the United States government's Medicare system more than \$200 billion, suggesting that the greying of America's population may prove less of a financial burden than expected.

The increasing self-reliance of many elderly people is probably linked to a massive increase in the use of simple home medical aids. For instance, the use of raised toilet seats has more than doubled since the start of the study, and the use of bath seats has grown by more than 50%. These developments also bring some health benefits, according to a report from the Macarthur Foundation's research group on successful ageing. The group found that those elderly people who were able to retain a sense of independence were more likely to stay healthy in old age.

Maintaining a level of daily physical activity may help mental functioning, says Carl Cotman, a neuroscientist at the University of California at Irvine. He found that rats that exercise on a treadmill have raised levels of brain-derived neurotrophic factor coursing through their brains. Cotman believes this hormone, which keeps neurons functioning, may prevent the brains of active humans from deteriorating.

As part of the same study, Teresa Seeman, a social epidemiologist at the University of Southern California in Los Angeles, found a connection between self-esteem and stress in people over 70. In laboratory simulations of challenging activities such as driving, those who felt in control of their lives pumped out lower levels of stress hormones such as cortisol. Chronically high levels of these hormones have been linked to heart disease.

But independence can have drawbacks. Seeman found that elderly people who felt emotionally isolated maintained higher levels of stress hormones even when asleep. The research suggests that older people fare best when they feel independent but know they can get help when they need it.

'Like much research into ageing, these results support common sense,' says Seeman. They also show that we may be underestimating the impact of these simple factors. 'The sort of thing that your grandmother always told you turns out to be right on target,' she says.

Questions 14-22

Complete the summary using the list of words, A-Q, below.

Write the correct letter, A-Q, in boxes 14-22 on your answer sheet.

Research carried out by scientists in the United States has shown that the proportion of people over 65 suffering from the most common age-related medical problems is 14 and that the speed of this change is 15..... It also seems that these diseases *ere* affecting people 16 in life than they did in the past. This is largely due to developments in 17, but other factors such as improved 18..... may also be playing a part. Increases in some other illnesses may be due to changes in personal habits and to 19..... The research establishes a link between levels of 20 and life expectancy. It also shows that there has been a considerable reduction in the number of elderly people who are 21 which means that the 22 involved in supporting this section of the population may be less than previously predicted.

A cost	B falling	C technology
D undernourished	E earlier	F later
G disabled	H more	I increasing
J nutrition	K education	L constant
M medicine	N pollution	O environmental
P health	Q independent	

Questions 23-26

Complete each sentence with the correct ending, A-H, below. Write the correct letter, A-H, in boxes 23-26 on your answer sheet.

- 23 Home medical aids
- 24 Regular amounts of exercise
- 25 Feelings of control over life
- 26 Feelings of loneliness

- A** may cause heart disease.
- B** can be helped by hormone treatment.
- C** may cause rises in levels of stress hormones.
- D** have cost the United States government more than \$200 billion.
- E** may help prevent mental decline.
- F** may get stronger at night.
- G** allow old people to be more independent.
- H** can reduce stress in difficult situations.

READING PASSAGE 3

Numeration

One of the first great intellectual feats of a young child is learning how to talk, closely followed by learning how to count. From earliest childhood we are so bound up with our system of numeration that it is a feat of imagination to consider the problems faced by early humans who had not yet developed this facility. Careful consideration of our system of numeration leads to the conviction that, rather than being a facility that comes naturally to a person, it is one of *the* great and remarkable achievements of the human race.



It is impossible to learn the sequence of events that led to our developing the concept of number. Even the earliest of tribes had a system of numeration that, if not advanced, was sufficient for the tasks that they had to perform. Our ancestors had little use for actual numbers; instead their considerations would have been more of the kind *is this enough* rather than *How many* when they were engaged in food gathering, for example. However when early humans first began to reflect on the nature of things around them, they discovered that they needed an idea of number simple to keep their thoughts in order. As they began to settle, grow plants and herd animals, the need for a sophisticated number system became paramount. It will never be known how and when this numeration ability developed, but it is certain that numeration was well developed by the time humans had formed even semi-permanent settlements.

Evidence of early stages of arithmetic and numeration can be readily found. The indigenous peoples of Tasmania were only able to count one, two, many; those of South Africa counted one, two, two and one, two twos and one, and so on. But in real situations the number and words are often accompanied by gestures to help resolve any confusion. For example, when using the one, two, many type of system, the word *many* would mean, *Look at my hands and see how many fingers I am showing you*. This basic approach is limited in the range of numbers that it can express, but this range will generally suffice when dealing with the simpler aspects of human existence.

The lack of ability of some cultures to deal with large numbers is not really surprising. European languages, when traced back to their earlier version, are very poor in number words and expressions. The ancient Gothic word for ten, *tachund*, is used to express the number 100 as *tachund tachund*. By the seventh century, the word *teon* had become interchangeable with the *tachund* or *hund* of the Anglo-Saxon language, and so 100 was denoted as *hund teontig*, or ten times ten. The average person in the seventh century in Europe was not as familiar with numbers as we are today. In fact, to qualify as a witness in a court of law a man had to be able to count to nine!

Perhaps the most fundamental step in developing a sense of number is not the ability to count, but rather to see that a number is really an abstract idea instead of a simple attachment to a group of particular objects. It must have been within the grasp of the earliest humans to conceive that four birds are distinct from two birds; however, it is not an elementary step to associate the number 4, as connected with four birds, to the number 4, as connected with four rocks. Associating a number as one of the qualities of a specific object is a great hindrance to the development of a true number sense. When the number 4 can be registered in the mind as a specific word, independent of the object being referenced, the individual is ready to take the first step toward the development of a notational system for numbers and, from there, to arithmetic.

Traces of the very first stages in the development of numeration can be seen in several living languages today. The numeration system of the Tsimshian language in British Columbia contains seven distinct sets of words for numbers according to the class of the item being counted: for counting flat objects and animals, for round objects and time, for people, for long objects and trees, for canoes, for measures, and for counting when no particular object is being numerated. It seems that the last is a later development while the first six groups show the relics of an older system. This diversity of number names can also be found in some widely used languages such as Japanese.

Intermixed with the development of a number sense is the development of an ability to count. Counting is not directly related to the formation of a number concept because it is possible to count by matching the items being counted against a group of pebbles, grains of corn, or the counter's fingers. These aids would have been indispensable to very early people who would have found the process impossible without some form of mechanical aid. Such aids, while different, are still used even by the most educated in today's society due to their convenience. All counting ultimately involves reference to something other than the things being counted. At first it may have been grains or pebbles but now it is a memorised sequence of words that happen to be the names of the numbers.

Questions 27-31

Complete each sentence with the correct ending, A-G, below. Write the correct letter, A-G, in boxes 27-31 on your answer sheet.

- 27 A developed system of numbering
- 28 An additional hand signal
- 29 In seventh-century Europe, the ability to count to a certain number
- 30 Thinking about numbers as concepts separate from physical objects
- 31 Expressing number differently according to class of item

A	was necessary in order to fulfil a civic role.
B	was necessary when people began farming
C	was necessary for the development of arithmetic
D	persists in all societies
E	was used when the range of number words was restricted
F	can be traced back to early European languages.
G	was a characteristic of early numeration systems

Questions 32-40

Do the following statements agree with the information given in Reading Passage 3? In boxes 32-40 on your answer sheet, write

TRUE	if the statement agrees with the information
FALSE	if the statement contradicts the information
NOT GIVEN	if there is no information on this

- 32 For the earliest tribes, the concept of sufficiency was more important than the concept of quantity.
- 33 Indigenous Tasmanians used only four terms to indicate numbers of objects.
- 34 Some peoples with simple number systems use body language to prevent misunderstanding of expressions of number.
- 35 All cultures have been able to express large numbers clearly.
- 36 The word 'thousand'¹ has Anglo-Saxon origins.
- 37 In general, people in seventh-century Europe had poor counting ability.
- 38 In the Tsimshian language, the number for long objects and canoes is expressed with the same word.
- 39 The Tsimshian language contains both older and newer systems of counting.
- 40 Early peoples found it easier to count by using their fingers rather than a group of pebbles.

Questions 41 – 50

Look at the sentences below about a comedy club for young people, called *Comedy Kids*.

Read the text on the opposite page to decide if each sentence is correct or incorrect.

If it is correct, mark **A** on your answer sheet.

If it is not correct, mark **B** on your answer sheet.

- 41 *Comedy Kids* operates a number of internationally-based clubs for young people.
- 42 *Comedy Kids* comedians have had as much experience working with children as with adults.
- 43 According to John Winterton, he is the first ever young people's stand-up comedian.
- 44 John looks like most people's idea of a typical entertainer for young people.
- 45 *Comedy Kids* will hold special parties at their clubs for young people if requested.
- 46 A lot of *Comedy Kids* comedians have the right qualities for performing at young people's parties.
- 47 Guests at *Comedy Kids* parties are invited to perform if they wish.
- 48 *Comedy Kids* have discovered lots of good young comedians who have not been trained.
- 49 In a year's time there'll be more classes available for young people wanting to be comedians.
- 50 *Comedy Kids* prefer it if the young people who join them are already quite confident.

Pencil must be used to complete this sheet.

Please write your full name in CAPITAL letters on the line below:

Then write your six digit Candidate number in the boxes and shade the number in the grid on the right.

A staircase diagram with 6 steps, each labeled with numbers 0 through 9. A pencil icon is on the right.

Test date (shade ONE box for the day, ONE box for the month and ONE box for the year):

Day: 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Month: 01 02 03 04 05 06 07 08 09 10 11 12 Year (last 2 digits): 09 10 11 12 13 14 15 16 17 18

Student name only			Teacher name only		
1		☐	26		☐
2		☐	27		☐
3		☐	28		☐
4		☐	29		☐
5		☐	30		☐
6		☐	31		☐
7		☐	32		☐
8		☐	33		☐
9		☐	34		☐
10		☐	35		☐
11		☐	36		☐
12		☐	37		☐
13		☐	38		☐
14		☐	39		☐
15		☐	40		☐
16		☐	41		☐
17		☐	42		☐
18		☐	43		☐
19		☐	44		☐
20		☐	45		☐
21		☐	46		☐
22		☐	47		☐
23		☐	48		☐
24		☐	49		☐
25		☐	50		☐

Listening	
Total	

Appendix 6: Sample AEPET Writing Test

ADVANCED EDUCATIONAL PROGRAM ENGLISH TEST

Subject: Writing

Place:

Time: 60'

Date:

Test code: 503

Full name:..... Candidate No:.....

INSTRUCTIONS

Do not open the question paper until you are told to do so

Write your name and your candidate number on your answer sheet if they are not already there

Read **ALL** the instructions carefully

You must complete the answer sheets within the time limit

INFORMATION

Writing Part 1 carries 4 marks

Writing Part 2 carries 6 marks

Do not accept misspelling

WRITING

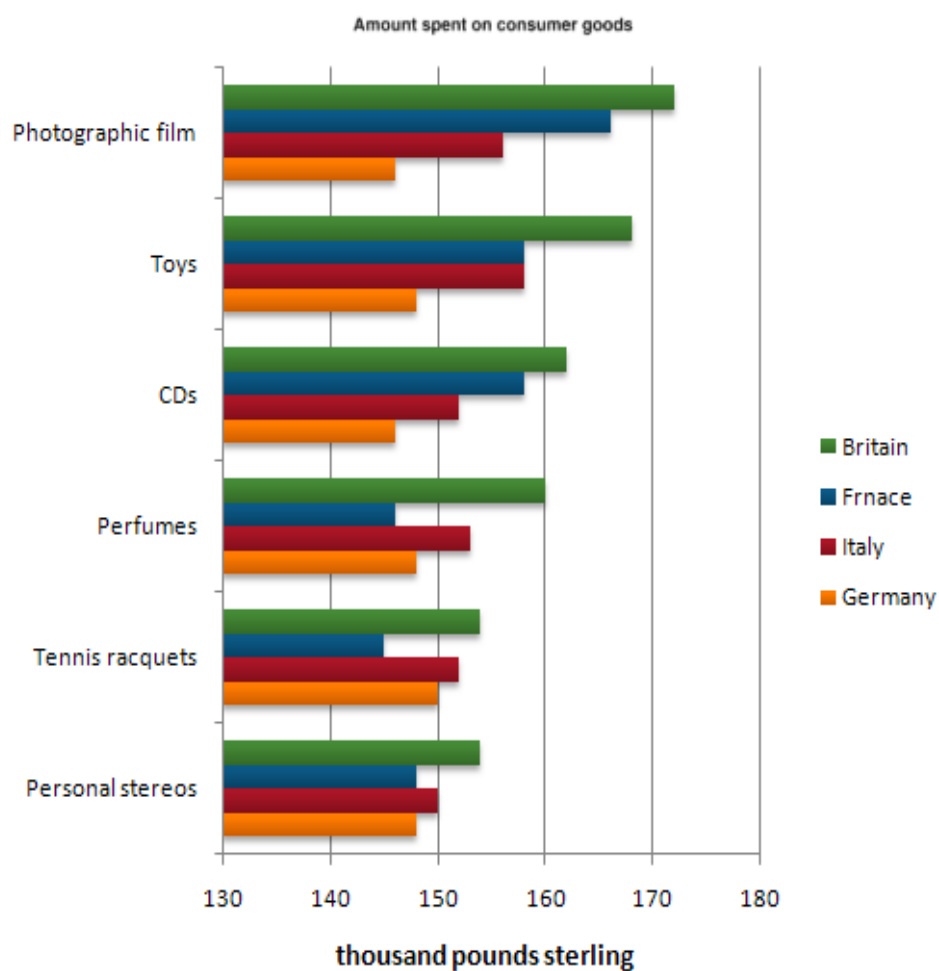
WRITING TASK 1

You should spend about 20 minutes on this task.

The chart below shows the amount spent on six consumer goods in four European countries.

Write a report for a university lecturer describing the information shown below.

You should write at least 150 words.



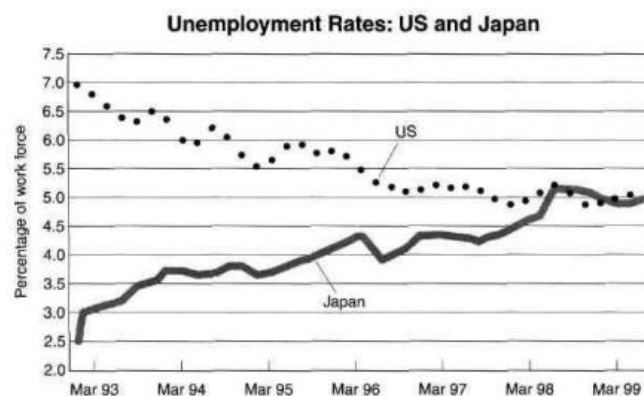
WRITING**WRITING TASK 1**

You should spend about 20 minutes on this task.

The graph below shows the unemployment rates in the US and Japan between March 1993 and March 1999.

Write a report for a university lecturer describing the information shown below.

You should write at least 150 words.



Appendix7: Sample AEPET Speaking Test

ENGLISH SPEAKING TEST
TIME LIMIT: 11-14 MINUTES / CANDIDATE
(HANDOUTS FOR EXAMINERS)

SPEAKING

PART 1

The examiner asks the candidate about him/herself, his/her home, work or studies and other familiar topics.

EXAMPLE

Traffic where you live

- . How do most people travel to work where you live?
- . What traffic problems are there in your area? [Why?]
- . How do traffic problems affect you?
- . How would you reduce the traffic problems in your area?

PART 2

Describe a game or sport you enjoy playing.

You should say:

- what kind of game or sport it is
- who you play it with
- where you play it and explain why you enjoy playing it.

You will have to talk about the topic for one to two minutes.

You have one minute to think about what you are going to say.

You can make some notes to help you if you wish.

PART 3

Discussion topics:

Children's games

Example questions:

How have games changed from the time when you were a child?

Do you think this has been a positive change? Why? Why do you think children like playing games?

Games and competition

Example questions:

Do you think competitive games are good or bad for children? In what ways?

How can games sometimes help to unite people?

Why is competition often seen as important in today's society?

Appendix 8: IELTS Writing Band Descriptors: Task 1 (Public version)

IELTS™		WRITING TASK 1: Band Descriptors (public version)	
Band	Task achievement	Coherence and cohesion	Lexical resource
9	- fully satisfies all the requirements of the task - clearly presents a fully developed response	- uses cohesion in such a way that it attracts no attention - skillfully manages paragraphing	uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'
8	- covers all requirements of the task sufficiently - presents, highlights and illustrates key features/ bullet points clearly and appropriately	- sequences information and ideas logically - manages all aspects of cohesion well - uses paragraphing sufficiently and appropriately	- uses a wide range of vocabulary fluently and flexibly to convey precise meanings - skillfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation - produces rare errors in spelling and/or word formation
7	- covers the requirements of the task (A) presents a clear overview of main trends, differences or stages (GT) presents a clear purpose, with the tone consistent and appropriate - clearly presents and highlights key features/bullet points but could be more fully extended	- logically organises information and ideas, there is clear progression throughout - uses a range of cohesive devices appropriately although there may be some under-/over-use	- uses a sufficient range of vocabulary to allow some flexibility and precision - uses less common lexical items with some awareness of style and collocation - may produce occasional errors in word choice, spelling and/or word formation
6	- addresses the requirements of the task (A) presents an overview with information appropriately selected (GT) presents a purpose that is generally clear; there may be inconsistencies in tone - presents and adequately highlights key features/ bullet points but details may be irrelevant, inappropriate or inaccurate	- arranges information and ideas coherently and there is a clear overall progression - uses cohesive devices effectively, but cohesion within and/or between sentences may be faulty or mechanical - may not always use referencing clearly or appropriately	- uses an adequate range of vocabulary for the task - attempts to use less common vocabulary but with some inaccuracy - makes some errors in spelling and/or word formation, but they do not impede communication
5	- generally addresses the task; the format may be inappropriate in places (A) recounts detail mechanically with no clear overview; there may be no data to support the description (GT) may present a purpose for the letter that is unclear at times; the tone may be variable and sometimes inappropriate - presents, but inadequately covers, key features/ bullet points; there may be a tendency to focus on details	- presents information with some organisation but there may be a lack of overall progression - makes inadequate, inaccurate or over-use of cohesive devices - may be repetitive because of lack of referencing and substitution	- uses a limited range of vocabulary, but this is minimally adequate for the task - may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader - may be faulty; errors can cause some difficulty for the reader
4	- attempts to address the task but does not cover all key features/bullet points; the format may be inappropriate (GT) fails to clearly explain the purpose of the letter; the tone may be inappropriate - may confuse key features/bullet points with detail; parts may be unclear, irrelevant, repetitive or inaccurate	- presents information and ideas but these are not arranged coherently and there is no clear progression in the response - uses some basic cohesive devices but these may be inaccurate or repetitive	- uses only a very limited range of structures with only rare use of subordinate clauses - some structures are accurate but errors predominate, and punctuation is often faulty
3	- fails to address the task, which may have been completely misunderstood - presents limited ideas which may be largely irrelevant/repetitive	- does not organise ideas logically - may use a very limited range of cohesive devices, and those used may not indicate a logical relationship between ideas	attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning
2	answer is barely related to the task	has very little control of organisational features	cannot use sentence forms except in memorised phrases
1	answer is completely unrelated to the task	fails to communicate any message	cannot use sentence forms at all
0	- does not attempt the task in any way - writes a totally memorised response		

Appendix 9: IELTS Writing Band Descriptors: Task 2 (Public version)

IELTS TM WRITING TASK 2: Band Descriptors (public version)			
Band	Task response	Coherence and cohesion	Lexical resource
9	- fully addresses all parts of the task - presents a fully developed position in answer to the question with relevant, fully extended and well supported ideas	- uses cohesion in such a way that it attracts no attention - skillfully manages paragraphing	uses a wide range of vocabulary with very natural and sophisticated control of lexical features; rare minor errors occur only as 'slips'
8	- sufficiently addresses all parts of the task - presents a well-developed response to the question with relevant, extended and supported ideas	- sequences information and ideas logically - manages all aspects of cohesion well - uses paragraphing sufficiently and appropriately	- uses a wide range of vocabulary fluently and flexibly to convey precise meanings - skillfully uses uncommon lexical items but there may be occasional inaccuracies in word choice and collocation - produces rare errors in spelling and/or word formation
7	- addresses all parts of the task - presents a clear position throughout the response - presents, extends and supports main ideas, but there may be a tendency to over-generalise and/or supporting ideas may lack focus	- logically organises information and ideas; there is clear progression throughout - uses a range of cohesive devices appropriately although there may be some under-over-use - presents a clear central topic within each paragraph	- uses a sufficient range of vocabulary to allow some flexibility and precision - uses less common lexical items with some awareness of style and collocation - may produce occasional errors in word choice, spelling and/or word formation
6	- addresses all parts of the task although some parts may be more fully covered than others - presents a relevant position although the conclusions may become unclear or repetitive - presents relevant main ideas but some may be inadequately developed/unclear	- arranges information and ideas coherently and there is a clear overall progression - uses cohesive devices effectively, but cohesion within and/or between sentences may be faulty or mechanical - may not always use referencing clearly or appropriately - uses paragraphing, but not always logically	- uses an adequate range of vocabulary for the task - attempts to use less common vocabulary but with some inaccuracy - makes some errors in spelling and/or word formation, but they do not impede communication
5	- addresses the task only partially; the format may be inappropriate in places - expresses a position but the development is not always clear and there may be no conclusions drawn - presents some main ideas but these are limited and not sufficiently developed; there may be irrelevant detail	- presents information with some organisation but there may be a lack of overall progression - makes inadequate, inaccurate or over-use of cohesive devices - may be repetitive because of lack of referencing and substitution - may not write in paragraphs, or paragraphing may be inadequate	- uses a limited range of vocabulary, but this is minimally adequate for the task - may make noticeable errors in spelling and/or word formation that may cause some difficulty for the reader
4	- responds to the task only in a minimal way or the answer is tangential; the format may be inappropriate - presents a position but this is unclear - presents some main ideas but these are difficult to identify and may be repetitive, irrelevant or not well supported	- presents information and ideas but these are not arranged coherently and there is no clear progression in the response - uses some basic cohesive devices but these may be inaccurate or repetitive - may not write in paragraphs or their use may be confusing	- uses only basic vocabulary which may be used repetitively or which may be inappropriate for the task - has limited control of word formation and/or spelling; errors may cause strain for the reader
3	- does not adequately address any part of the task - does not express a clear position - presents few ideas, which are largely undeveloped or irrelevant	- does not organise ideas logically - may use a very limited range of cohesive devices, and those used may not indicate a logical relationship between ideas	- uses only a very limited range of structures with only rare use of subordinate clauses - some structures are accurate but errors predominate, and punctuation is often faulty
2	- barely responds to the task - does not express a position - may attempt to present one or two ideas but there is no development	has very little control of organisational features	attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning
1	answer is completely unrelated to the task	fails to communicate any message	cannot use sentence forms except in memorised phrases
0	- does not attempt - does not attempt the task in any way - writes a totally memorised response	can only use a few isolated words	cannot use sentence forms at all

SPEAKING: Band Descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	- speaks fluently with only rare repetition or self-correction; - any hesitation is content-related rather than to find words or grammar - speaks coherently with fully appropriate cohesive features - develops topics fully and appropriately	- uses vocabulary with full flexibility and precision in all topics - uses idiomatic language naturally and accurately	- uses a full range of structures naturally and appropriately - produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	- uses a full range of pronunciation features with precision and subtlety - sustains flexible use of features throughout - is effortless to understand
8	- speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language - develops topics coherently and appropriately	- uses a wide vocabulary resource readily and flexibly to convey precise meaning - uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies - uses paraphrase effectively as required	- uses a wide range of structures flexibly - produces a majority of error-free sentences with only very occasional inaccuracies or less-than-systematic errors	- uses a wide range of pronunciation features - sustains flexible use of features, with only occasional lapses - is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	- speaks at length without noticeable effort or loss of coherence - may demonstrate language-related hesitation at times, or some repetition and/or self-correction - uses a range of connectives and discourse markers with some flexibility	- uses vocabulary resource flexibly to discuss a variety of topics - uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices - uses paraphrase effectively	- uses a range of complex structures with some flexibility - frequently produces error-free sentences, though some grammatical mistakes persist	shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	- is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation - uses a range of connectives and discourse markers but not always appropriately	- has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inaccuracies - generally paraphrases successfully	- uses a mix of simple and complex structures, but with limited flexibility - may make frequent mistakes with complex structures though these rarely cause comprehension problems	- uses a range of pronunciation features with mixed control - shows some effective use of features but this is not sustained - can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	- usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going - may over-use certain connectives and discourse markers - produces simple speech fluently, but more complex communication causes fluency problems	- manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility - attempts to use paraphrase but with mixed success	- produces basic sentence forms with reasonable accuracy - uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	- cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction - links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	- is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice - rarely attempts paraphrase	- produces basic sentence forms and some correct simple sentences but subordinate structures are rare - errors are frequent and may lead to misunderstanding	- uses a limited range of pronunciation features - attempts to control features but lapses are frequent - mispronunciations are frequent and cause some difficulty for the listener
3	- speaks with long pauses - has limited ability to link simple sentences - gives only simple responses and is frequently unable to convey basic message	- uses simple vocabulary to convey personal information - has insufficient vocabulary for less familiar topics	- attempts basic sentence forms but with limited success, or relies on apparently memorised utterances - makes numerous errors except in memorised expressions	shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	- pauses lengthily before most words - little communication possible	only produces isolated words or memorised utterances	cannot produce basic sentence forms	Speech is often unintelligible
1	no communication possible			
0	- no reliable language - does not attend			

