***PUPIL: International Journal of Teaching, Education and Learning***
***ISSN 2457-0648***

*Supalak Nakhornsri, 2020*

*Volume 4 Issue 2, pp. 264-286*

*Date of Publication: 3rd October, 2020*

*DOI- https://doi.org/10.20319/pijtel.2020.42.264286*

*This paper can be cited as: Nakhornsri, S. (2020). Establishing Performance Indicators for Academic English Proficiency: A Case for EFL University Students. PUPIL: International Journal of Teaching, Education and Learning, 4(2), 264-286.*



# ESTABLISHING PERFORMANCE INDICATORS FOR ACADEMIC ENGLISH PROFICIENCY: A CASE FOR EFL UNIVERSITY STUDENTS

**Supalak Nakhornsri**
King Mongkut's University of Technology North Bangkok, Thailand
*supalak.n@arts.kmutnb.ac.th*

## Abstract

*In Thailand, the B1 level of the CEFR was specified as the goal of English proficiency for high school or vocational certificate graduates. Therefore, for university levels, students' English ability equal to the B1 level or above is expected. English proficiency tests are needed to fulfill this assessment purpose. Moreover, one of the considering points is that the test results are generally reported quantitatively without meaningful interpretations. This study attempts to investigate the Target Language Use (TLU) domain of academic English proficiency for EFL university students so that performance indicators can be established. The quality of this developed test was evaluated and the concurrent validity was examined to demonstrate how well the developed test correlated with a well-established test. The sample of the study included 39 informants that provided information about the TLU domain and 30 test takers for the main study. The instruments consisted of a questionnaire asking about the TLU domain and the developed academic English proficiency test. The arithmetic mean and Standard Deviation (S.D.) were employed for the analysis of the TLU domain. Construct validity, reliability, and item analysis were analyzed for test quality. Finally, Pearson's product-moment was implemented in order to examine concurrent validity. Besides,*

*regression analysis was used to demonstrate if the obtained scores could predict the scores from the standardized test. The findings are significant in several ways. First, these findings are able to provide a clearer concept of the performance indicators of academic English proficiency. The developed test can be a useful instrument for several purposes, for example for placement or achievement. Finally, the assessment of concurrent validity can be useful in terms of test score interpretations.*

**Keywords**

Ability Bands, Academic English, Concurrent Validity, English Proficiency

## 1. Introduction

Due to global challenges, universities or educational institutes possess an important and undeniable task to secure the future labor force which can be done through fostering knowledge, analytic thinking, broad capabilities, and technical skills on the part of students. A key priority is ensuring that students be prepared and supported to effectively make their own decisions in a rapidly changing world (O'Connell, 2016). English proficiency is considered a key for academic and occupational success, and for this reason, universities are essentially required to prepare their graduates with English proficiency.

In the educational context of Thailand, the Office of the Basic Education Commission and the Institute of the English Language have applied the concepts of the Common European Framework of Reference for Languages (CEFR), which is an international standard for describing language ability, as the model to reform English teaching. The CEFR meaningfully explains language ability on a six-point scale, from A1 for beginners, up to C2 for those that have mastered a language (Cambridge, 2020). Level B1 was specified as the goal of English proficiency for Mathayomsuksa 6 (high school) or vocational certificate graduates in Thailand (Sornkam, Person, & Yordchim, 2018). Therefore, for higher educational levels, students' English ability, which is equal to the B1 level or above, should be expected.

According to this, major tests of English proficiency designed to measure people's language ability (Yuyun, Meyling, Laksana, & Abenedgo, 2018) are needed in order to fulfill this assessment purpose. However, language teachers have neglected testing theories for decades in their tests. A lack of widely-accepted theories is the main cause of neglect. Buck (1991: 67) states that practically test constructors tend to create tests by following their instincts or by their judgment. This could

raise the awareness of the validity of a language test. Moreover, test results are generally reported quantitatively or in a form of numeric data without meaningful interpretation, and most test results are not likely to be used further as beneficial information for educational development.

In terms of EFL students' English proficiency, a number of studies reported that many EFL students possessed insufficient academic English skills in order to cope with academic study in English-medium universities. Importantly, the students required English ability to understand English lectures and express their opinions and comments (Bamford, 2006). Due to the lack of English proficiency, EFL students reported low satisfaction with their infrequent participation in interactive activities. For the required English skills, Han (2007) reported that additional oral and aural skill training courses were needed since these skills would be able to enhance their listening comprehension, conversation, and presentation skills. According to this, the improvement of English skills may allow the high possibility to increase their participation level.

In recent years there has been an increased focus on academic English proficiency, especially in terms of assessment, in order to have a valid test to recruit students into English-medium programs (Read, 2015). Hence, having a valid and reliable academic English proficiency test is necessary since the scores obtained from it can be beneficial for recruiting students to the proper programs, allowing institutes to provide preparation or remedial courses before or during the study in the English-medium programs or regular classes where English is required.

With regard to creating a valid and reliable test, there are many aspects to be considered. Although the concept of academic English proficiency has been accepted, an adequate definition of such proficiency is still ambiguous (Brindley, 1998). This might be caused by a lot of different processes and aspects involved in developing academic English proficiency tests, perhaps making a global, comprehensive definition impossible. Bachman and Palmer (1996) suggest the realization of language use while specific language use tasks are being performed. Therefore, constructing a language test should be seen as a combination of language ability and task characteristics.

Initially, a test aims to predict the academic performance of skills outside the typical academic setting. The trend of theory-based assessments holds several advantages. Examiners could get the opportunity to assess specific and accurate diagnoses about an individual's ability through the implementation of assessments based on research-based concepts in contradistinction to conventional tests from earlier periods. As a means of developing effective learning interventions, tasks from new assessment methods have been designed in accordance with the skills of academic

achievement. Ansastasi & Urbina (1997) suggested that an instrument should be assessed through the use of validity studies since validity can be considered as the most important aspect in test evaluation (Drummond, 1996). The main purpose of validity studies is to gain a better understanding of assessing ability processes (Sattler, 1992).

One kind of validity is concurrent validity, which is an important type of validity that can be a good guide for new testing procedures. Concurrent validity can be calculated by correlation analysis. The obtained correlation coefficient of the scores from the two sets of measurements is taken by the same target group. These two sets of measurements are the newly developed instruments and standard instruments (Craighead & Nemeroff, 2004).

According to this, the language skills used in certain situations in terms of task characteristics, language ability, and topical knowledge are necessary to be defined when designing and using a test (Bachman & Palmer, 1996).

As suggested above, particular Target Language Use (TLU) domains have to be established and included in the test construct in order to increase the validity of an English proficiency test. The purposes of this study are therefore 1) to develop an academic English proficiency test through the TLU domain obtained from the stakeholders, 2) to evaluate the quality of the developed test, 3) to assess concurrent validity by correlating the scores obtained from the developed test with the standardized one and 4) to create the ability bands for describing the English levels.

**1.1 Research Questions**

1. What is the TLU domain of academic English proficiency for EFL university students?
2. What is the quality of the developed academic English proficiency test?
3. What is the concurrent validity value of the developed academic English proficiency test when correlated with the standardized test?
4. What are the ability bands for describing the English levels obtained from the developed academic English test?

**1.2 Significance of the Study**

The findings of this study are expected to be significant in several ways. First, in terms of theoretical significance, the findings can provide clearer concepts of the TLU domain of academic English proficiency tests. The information on what constitutes an academic English proficiency test can be applied to the design and development of standardized language tests. The developed test has the potential to be a useful instrument in assessing the students' academic English proficiency used

when studying in English-medium universities, international programs, or regular classes where academic English is required.

Finally, the assessment of concurrent validity can be useful for test score interpretations and the creation of the ability bands since the value of concurrent validity is from the correlation analysis of scores obtained from the newly created instruments and the standard instruments. Hence, the scores from the developed test can be compared with the standardized test. It can be said that concurrent validity can be a good guide for new testing procedures.

## 2. Literature

To gain insights into English proficiency and concurrent validity, the literature relevant to these topics is reviewed in the following.

### 2.1 The Concepts of English Proficiency

TLU domains are essential for specifying the construct of tests, data, or information from English proficiency theories, needs analysis or stakeholders' opinions are required.

Frameworks for English language proficiency standards are similar to those used for the classroom and large-scale assessment. The standards can reflect the social and academic dimensions of acquiring a second language. A specific context for language acquisition (social and instructional settings) is normally addressed in English proficiency standards. When considering the perspectives of language proficiency, Spolsky's theory of second language learning (1989) illustrates a set of conditions shaping the acquisition process. There is a recognition that individual language learners vary in their productive and receptive skills. In terms of language skill development, receptive language skills (listening and reading) generally develop prior to and to a higher level than productive language skills (speaking and writing).

From the conceptual frameworks, Read (2015) connected the frameworks of language proficiency with the proficiency assessment. These frameworks have long been acknowledged as one of the basic purposes of language assessment since there are also various purposes of implementing proficiency tests. Additionally, Read (2015) stated that the focus of language proficiency assessment is to assess how well examinees can use the language for functional communication.

In the 1950s, the focus of language proficiency shifted in conceptual frameworks regarding the nature of the constructs that are considered as the basis for designing a proficiency test

(Hawkey, 2004). The conventional tests directly measure the test takers' language competence rather than his or her practical knowledge for communicative purposes.

From the tentative changes of the assessment, assessing the use of the language for communicative purposes is considered essential. This means that the focus of assessment should be moving toward performance. Definitions of performance allow a global overview of the process of language acquisition. A summary and synthesis of the model performance indicators for each language proficiency level can be created through these definitions.

**2.2 Concurrent Validity**

Concurrent validity is an approach estimating an examinee's performance by using different tests. Lin and Yao (2019) explain that it describes the processes when a test is able to effectively assess an examinee's performance by using certain outcome measures. The outcome measure can be beneficial if it can provide a precise prediction of the criteria. It is similar to predictive validity since the interpretation is based on correlations between a test and the relevant criteria (McIntire & Miller, 2005). Once some types of tests are created, validity needs to be examined concerning whether it measures what it is previously designed. Evaluation is required to evaluate if a new test can be compared with a well-established one. It can also refer to the practice of concurrently testing two groups at the same time, or asking two different groups of people to take the same test. Glen (2015) suggests the advantages of concurrent validity—that it provides a simple and quick method to validate data and it is a highly appropriate way to validate personal attributes such as strengths and weaknesses.

Concurrent validity can be calculated by carrying out the correlation between a new test and an existing one in order to prove whether the new test correlates with the existing one (Murphy & Davidshofer, 1998). According to this, the result obtained from the correlation analysis is the concurrent validity coefficient. Hence, the concepts of English proficiency should be well defined in order to establish clear and detailed performances, which can allow the expected responses obtained from the tests to beneficially reflect authentic tasks. Moreover, once the tests are developed, concurrent validity, which is a kind of test quality, can reflect how well the test is designed and it can prove if the test construct corresponds with the existing standardized tests.

## 3. Research Methodology

The implementation of this part is as follows:

### 3.1 Research Design

Research and Development (R&D) aims to create an academic English proficiency test for EFL university students. Concurrent validity, which allows the meaningful interpretation and explanation of the scores, is also the focus.

### 3.2 Participants

According to the purposes of this study, the sample of the study consisted of three groups.

**3.2.1 Group 1**: 39 informants providing information to create a TLU domain for academic English proficiency for EFL university students consisting of English language instructors, currently-enrolled students, and the instructors of other faculties.

**3.2.2 Group 2**: 30 test takers participating in the pilot study, including undergraduate and postgraduate students with heterogeneous English abilities.

**3.2.3 Group 3**: 30 test takers for the main study consisting of students studying at the undergraduate and postgraduate levels in Thai universities. Dornyei (2011: 99-100) suggested that the optimum number for correlational research should be at least 15, and so the researcher included 30 students, which was more than the suggested number. The purposive sampling technique was used.

### 3.3 Research Instruments

The main instruments consisted of the questionnaire, asking about the TLU domain for the academic English proficiency of EFL students and the academic English proficiency test. These instruments were developed through the following processes.

#### 3.3.1 The Questionnaire on the TLU Domain for the Academic English Proficiency of EFL Students

1. The lists of language tasks included in the questionnaire were obtained by interviewing two English instructors, two instructors of other subjects, and four undergraduate and postgraduate students regarding the academic language tasks implemented through three language skills. Moreover, the language tasks specified in the Common European Framework (CEFR, 2018) were added.

2. After creating the questionnaire, three experts were asked to evaluate the validity of the contents by considering the framework of CEFR and the authenticity of the language tasks. The Item Objective Congruence (IOC) form was used.

3. The revised questionnaire was distributed online and 39 informants responded. The findings from the online questionnaire were summarized and used to create the TLU domain.

**3.3.2 The Academic English Proficiency Test**

1. The findings from the questionnaire on the TLU domain were used to develop the TLU domain for the academic English of the EFL students. As suggested by Bachman and Palmer (1996), the TLU domain should be appropriately modified so that it can be applied to the test design. Thus, the researcher modified the TLU domain containing the characters practically used to develop the test tasks.

2. Once the modified TLU domain was obtained, the test specification of the English Proficiency Test was developed and the ability descriptors of the CEFR were used for setting the difficulty levels of the test items.

Since the Institute of the English Language applied the concepts of the CEFR as the model for reforming English teaching. Level B1 was specified as the goal of English proficiency for the students that graduated from Mathayomsuksa 6 (high school) or that had a vocational certificate (Sornkam, Person, & Yordchim, 2018). According to this, this test, which focuses on the study at the university level, measures the lowest acceptable language ability level from Level B1, which is equivalent to IELTS Band 4 (intermediate). Consequently, the construct of academic English proficiency is as follows.

- The ability descriptors specified in the IELTS bands ranging from Band 4 to Band 8 are used as the measured behaviors and reflect the levels of measured English ability. Therefore, the IELTS band descriptors ranging from Band 4 to Band 8 were employed as the theoretical definitions of this test construct.

- Regarding the content coverage, this evidence is demonstrated in the characteristics of the TLU domain, construct definitions, and characteristics of the test task portions of the design statement.

- For the relevance of the construct to the purpose and use of the test or content relevance to justify the test used for the intended purpose, this quality was evaluated by content experts. They were asked to validate if the test items reflected the measured behaviors (IELTS band & CEFR descriptors) and the TLU domain.

- In terms of the unbiased test scores, this was designed for EFL university students so the tasks did not include culture-specific items and the tests did not give an unfair advantage or disadvantage to students from different cultural groups.

- For the domain of generalization, the test focuses on the domain of generalization on the TLU domain of university study in English-medium programs or regular programs where academic English is required.

3. The test specification was validated regarding task authenticity by two English lecturers and lecturers from other fields. Their comments were used to revise the test specification. Lastly, the test specification was finalized.

4. The test was created according to the revised specification. After its development, the test was evaluated for content relevance and appropriateness by three content experts in the applied linguistics field. Then, it was revised and tried out in the pilot study. Lastly, the test was revised again before being used in the main study.

5. Since the responses of the task tasks are short answers and paragraphs, intra-rater reliability was employed to specify the test reliability.

6. Once the test had been developed, the ability bands for describing English ability were created. Due to the fact that the CEFR descriptors and the TLU domains were applied to the test specification, the ability bands were developed.

## 4. Findings

The presentation of the findings follows the research questions respectively.

### 4.1 Research question 1: What is the TLU Domain of Academic English Proficiency for EFL University Students?

The data obtained from the 4-point Likert scale questionnaire asking about the TLU domain of EFL university students were calculated for a mean ( $\overline{X}$ **)** and standard deviation (S.D.). The findings can be presented in Table 1.

**Table 1:** *Descriptive Statics for the TLU Domain of the EFL University Students*

| Language skill | $\overline{X}$ | S. D | Interpretation |
|---|---|---|---|
| **Listening Purposes for Academic English Use** | | | |
| 1. To understand the lectures or the lessons in a classroom | 3.36 | .71 | **Always** |
| 2. To understand the interlocutor's speech for communication purposes | 3.33 | .66 | **Always** |
| 3. To understand academic content, especially in a conference | 2.97 | .90 | Usually |
| 4. To collect information or data | 3.46 | .64 | **Always** |
| 5. To understand information from visual representations such as a diagram (e.g. a piece of equipment), a set of pictures, a plan | 3.31 | .69 | **Always** |

| Language skill | $\overline{X}$ | S. D | Interpretation |
|---|---|---|---|
| (e.g. of a building), a map (e.g. parts of a town) | | | |
| 6. To understand the ideas/facts on a form: often used to record factual details such as names | 3.15 | .78 | Usually |
| 7. To understand the ideas/facts in a set of notes: used to summarize any types of information using the layout to show how different items relate to one another | 3.10 | .75 | Usually |
| 8. To understand the ideas/facts in a table: used as a way of summarizing information related to clear categories, e.g. place/time/price | 3.13 | .73 | Usually |
| 9. To understand the ideas/facts in a flow-chart: used to summarize a process that has clear stages, with the direction of the process shown by arrows | 3.15 | .75 | Usually |
| **Reading Purposes for Academic English Use** | | | |
| 1. To understand the content from a textbook in a classroom | 3.59 | .55 | **Always** |
| 2. To understand news or information from newspapers or websites | 3.23 | .71 | Usually |
| 3. To understand the meeting minutes at a meeting | 2.95 | .83 | Usually |
| 4. To understand a research article from a journal | 3.05 | .76 | Usually |
| 5. To understand general information from any reading texts | 3.21 | .57 | Usually |
| 6. To understand the information presented in a visual presentation such as a diagram, a flowchart, etc. | 3.08 | .74 | Usually |
| **Writing Purposes for Academic English Use** | | | |
| 1. To write a report or a research report | 3.15 | .81 | Usually |
| 2. To write processes or procedures | 2.97 | .81 | Usually |
| 3. To write a letter | 3.08 | .96 | Usually |
| 4. To summarize the contents learned in a classroom | 3.21 | .77 | Usually |
| 5. To fill out a form, i.e. an application form | 3.28 | .79 | **Always** |
| 6. To take notes | 3.21 | .89 | Usually |

According to the table, the purposes of academic English use can be summarized as follows.

**Listening Purposes:** the purposes that were rated as always implemented while studying were the following: 1. to understand lectures or lessons in a classroom ( $\overline{X}$ = 3.36, S.D. = .71); 2. to understand an interlocutor's speech for communication purposes ( $\overline{X}$ = 3.33, S.D. = .66); 4. to collect information or data ( $\overline{X}$ = 3.46, S.D. = .64); and 5. to understand the information from a visual representation ( $\overline{X}$ = 3.31, S.D. = .69).

**Reading Purposes:** the only purpose rated as always implemented while studying was to understand the content of a textbook in a classroom ( $\overline{X}$ = 3.59, S.D. = .55).

**Writing Purposes:** the only purpose rated as always implemented while studying was to fill out a form, i.e. an application form ( $\overline{X}$ = 3.28, S.D. = .79).

The findings from the TLU domain were applied to create the test specification. The table below is a summary of the test tasks in the developed test.

**Table 2:** *Summary of the Test Tasks in the Academic English Proficiency Test*

| Skill | Task Characteristic | Number of Item | Score |
|---|---|---|---|
| **Academic Listening** | **Task 1:** Listening as a member of a live audience | 15 | 15 |
| | **Task 2:** Listening to announcements and instructions | 15 | 15 |
| | **Task 3:** Understanding conversations between other speakers | 10 | 15 |
| | ***Total*** | ***40*** | ***40*** |
| **Academic Reading** | **Task 1:** Reading for information and argument | 10 | 10 |
| | **Task 2:** Reading instructions | 10 | 10 |
| | **Task 3:** Reading for orientation | 10 | 10 |
| | **Task 4:** Reading correspondence | 10 | 10 |
| | ***Total*** | ***40*** | ***40*** |
| **Academic Writing** | **Task 1:** Writing an abstract | 1 | 9 |
| | **Task 2:** Writing the details of a complex process | 1 | 9 |
| | **Task 3:** Writing an academic article that includes an introduction, body, and conclusion | 1 | 9 |
| | ***Total*** | ***3*** | ***27*** |
| **Overall** | | **83** | **107** |

Table 2 shows that this academic English proficiency test included three skills: listening, reading, and writing. There were three tasks with 40 short answer items for the academic listening part and four tasks with 40 short answer items for the academic reading part. Finally, there were three tasks in the academic writing part. In terms of scoring, one point was given for a correct answer in two parts: academic listening and reading. This means that each part had 40 points as the full score. For the academic writing part, the marking rubrics were used to assign the scores. The full score for each writing task was 9. Therefore, the total score for this academic writing part is 27.

**4.2 Research question 2:** What is the quality of the developed academic English proficiency test?

To evaluate the quality of the developed test, the test validity and reliability values were assessed. This is according to Darr (2005), where the validity assessment can be used to prove if

tests can measure what they are designed to measure. Moreover, the reliability, which refers to the equivalent assessments, can provide consistent results.

The results of the test validity analysis: Three experts were asked to consider the validity of the contents by considering a framework of language task characteristics and the written items. They evaluated and rated the congruence of the test items by using the Item Objective Congruence (IOC) form. For the estimation results, it was found that the IOC values of all the items were higher than 0.5. The results suggest that all of the items agreed and possessed the agreed levels of content validity.

The results of the test reliability analysis: The students' responses from the pilot study were analyzed for test reliability. Since short answers were the item type used in this developed test, the scores obtained were categorized as polytomous or scoring the written responses by using rating scales (Stevens, n.d.). All of the students' responses to the test items were analyzed for intra-rater reliability. Pearson's product-moment was employed to analyze the consistency of the rater when marking responses twice. The intra-rater reliability coefficients and the interpretations of the three parts are reported in the table below.

**Table 3:** *The Test Reliability Values for Academic Listening, Reading and Writing*

| Part | Intra-Rater Reliability Coefficient | Interpretation |
|---|---|---|
| 1. Academic Listening | .95 | Excellent |
| 2. Academic Reading | .96 | Excellent |
| 3. Academic Writing | .98 | Excellent |

It can be seen that the test has very high reliability (.95, .96, and .98 respectively). Therefore, test results are stable and consistent.

**The Results of the Test Item Analysis:** Item analysis is a process that examines the test takers' responses to individual test items and it was used to check the quality of each test item. The item analysis reported the values of item difficulty levels and the item discrimination index.

**1. Item Difficulty Levels**

Due to the short answer items, which were polytomous, scoring rubrics were employed. The formula for calculating the item difficulty for the polytomous items was:

$$\frac{\text{The average score on that item}}{\text{The highest number of points for anyone alternative}}$$ (Office of Educational Assessment, 2020)

Table 4 summarizes the ranges of difficulty levels from the three parts: academic listening, reading, and writing.

**Table 4:** *Summary of the Ranges of Difficulty Levels*

| Part | Range of difficulty levels |
|---|---|
| 1. Academic Listening: 40 items | 0.20 – 0.79 |
| 2. Academic Reading: 40 items | 0.20 – 0.80 |
| 3. Academic Writing: 3 items | 0.38 – 0.47 |

According to Table 4, the difficulty level values passed the acceptable ranges since the acceptable difficulty level values were between 0.20-0.80 (Bachman, 2004: 138).

**2. Item Discrimination Index**

The discrimination index means the ability of an item to differentiate test takers with high ability and low ability (Office of Educational Assessment, 2020). An item with a discrimination value greater than 0.20 is acceptable (Ebel, 1979). The summary of the ranges of the discrimination index from the three parts is shown in Table 5.

**Table 5:** *Summary of the Ranges of the Discrimination Index*

| Part | Range of Discrimination Index |
|---|---|
| 1. Academic Listening: 40 items | 0.94 – 0.95 |
| 2. Academic Reading: 40 items | 0.95 – 0.96 |
| 3. Academic Writing: 3 items | 0.67 – 0.98 |

According to Ebel (1979), an item with a discrimination value greater than 0.20 is acceptable and items with values higher than 0.40 are considered very good.

**4.3 Research Question 3: What is the Concurrent Validity Value of the Developed Academic English Proficiency Test when Correlating it with the Standardized Test?**

The scores obtained from the developed test and the IELTS scores were analyzed by means of Pearson's product-moment and Regression analysis. The main purpose of the analysis was to indicate concurrent validity. A level of significance at $\alpha = 0.05$ was determined (Hinkle et. al, 1998).

The hypotheses were set to investigate the relationship between the scores obtained from each part of the academic English proficiency test (academic listening, academic reading, and academic writing) and each part of the IELTS. The three hypotheses were as follows:

Hypothesis 1: There is a significant relationship between the scores obtained from the academic listening part and the IELTS listening scores. (H1: $r_{\mathbf{x(listening).y(IELTS\ listening)}} \neq 0$)

Hypothesis 2: There is a significant relationship between the scores obtained from the academic reading part and the IELTS reading scores. (H1: $r_{\mathbf{x(reading).y(IELTS\ reading)}} \neq 0$)

Hypothesis 3: There is a significant relationship between the scores obtained from the academic writing part and the IELTS writing scores. (H1: $r_{\mathbf{x(writing).y(IELTS\ writinging)}} \neq 0$)

Table 6 presents the results of the hypothesis testing.

**Table 6:** *The Correlation Coefficients of the Scores obtained from each part of the Academic English Proficiency Test and each part of the IELTS*

| Pair Correlation | Correlation Coefficient (r) |
|---|---|
| Academic Listening – IELTS Listening | .59** |
| Academic Reading – IELTS Reading | .65** |
| Academic Writing – IELTS Writing | .86** |

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

Table 6 illustrates the correlation coefficient between the academic listening scores and the IELTS listening scores which were 0.59. This coefficient value was considered as a positive moderate correlation. Similarly, the correlation coefficient between the academic reading scores and the IELTS reading scores was 0.65. This coefficient value was also considered as a positive moderate correlation. Regarding the writing skill, it shows that the correlation coefficient between the academic writing scores and the IELTS writing scores was 0.86. This coefficient value was considered as a positive high correlation.

In terms of predictability, the regression analysis was implemented in order to ascertain if academic listening, reading, and writing could predict the three skills on the IELTS. The results showed that the three parts of the developed test could predict the three skills on the IELTS and the regression equations could then be written as follows:

*IELTS listening score = 10.63 + 0.33 (The academic listening score)*

*IELTS reading score = 7.44 + 0.42 (The academic reading score)*

*IELTS writing score = 0.99 + 0.74 (The academic writing score)*

**4.4 Research question 4: What are the Ability Bands for Describing the English Levels Obtained from the Developed Academic English Test?**

By means of the three equations obtained for the three parts of the developed test, the ranges of the academic scores converted to IELTS scores could be obtained. Moreover, the creation of the test specification included the applications of the TLU domain of the EFL university students, the structures of the IELTS, and the descriptors of the CEFR, so the developed ability bands included the band descriptors that represented the performance indicators of the academic English proficiency of the EFL university students. Table 7 shows the sample of the created ability bands.

**Table 7:** *Sample of Ability Bands*

| Academic English Proficiency Test | IELTS | Ability Descriptors |
|---|---|---|
| **Level: Academic English user** ***"13 - 18 items"*** | **Bands 4-5 (CEFR level: B1)** | **<u>Listening</u>**<br>**Context: Listening for information about courses, assignments, projects, facilities**<br>**Task: Listening as a member of a live audience**<br>**Ability:** Can follow a lecture or talk within his/her field, provided the subject matter is familiar and the presentation is straightforward and structured<br>Can distinguish between main ideas and supporting details in standard lectures on familiar subjects, provided these are delivered in clearly articulated standard speech<br>Can follow a straightforward conference presentation or demonstration with visual support (e.g. slides, handouts) on a topic or product within his/her field, understanding explanations given (Cambridge, 2020) |

## 5. Discussion/Recommendations

This section presents a discussion and recommendations for further studies.

**5.1 Discussion:** Based on the research questions, the findings were discussed as follows.

**5.1.1 The TLU Domain of Academic English Proficiency for EFL University Students**

The study demonstrates the TLU domain derived from the task analysis, which was conducted with English teachers, teachers of other subjects, and undergraduate and postgraduate students. Moreover, the lists of the performances included in the research instrument were from various sources, such as interviews, previous studies, theories, and the CEFR descriptors, and the TLU domain signified the expected academic English performances so the performances could be comparable to standard benchmarks such as those of the CEFR and IELTS.

Turning to the details of the derived performances of the three English skills—the performances were in agreement with the dimensions of the functions of academic English suggested by Scarcella (2003). These dimensions include linguistic, cognitive, sociocultural, and psychological. This implies that in order to assess students' academic English proficiency, linguistic or language competence should not be the only focus. Other dimensions, such as cognitive or knowledge, and the socio-cultural and psychological dimensions, should be important focuses included in academic English tests. Additionally, McNamara and Roever (2006) has challenged classroom-based researchers to expand the notion of assessment. There are two specific areas suggested for further research and development: 1. more research relevant to the implementation of assessment schemes, and 2. more facilitative functions of assessment in classrooms involving the expansion of the notion regarding the assessment.

The above discussion reflects a significant shift in language assessment. Hawkey (2004) pointed out that conventional testing measures directly test the test taker's language knowledge rather than his/her ability to use the language for communicative purposes. However, the focus on language proficiency has shifted the thinking about the nature of the constructs that should be the basis for the design of a proficiency test. From the ongoing move of the assessment, assessing the use of the language for communicative purposes is considered essential. This means that the focus of the assessment is moving toward performance.

Accordingly, many universities worldwide are recognizing the importance of standard indicators that demonstrate the English proficiency level of their students, and to respond to this global challenge, universities are preparing their graduates with English proficiency tests. To conventionally assess students' English proficiency may not be valid, reliable, or accurate. The need to develop English language proficiency standards should develop from three sources: 1). pedagogy, 2). assessment, and 3). educational policy (CEFR, 2018).

Language proficiency assessment has to change teaching practices for English language learners. English language proficiency standards can be guides to develop test blueprints, task specifications, and English language proficiency measures. It can be said that language proficiency standards could increase the level of reliability and validity of assessment tools.

Regarding the establishment of performance indicators in language proficiency, especially academic English, the developing process of indicators and tests should include English language

instructors and instructors that possess the views of academics, with expertise across the various disciplines taught within the university.

In Thailand, there have been endeavors to set standards for English proficiency based on mapping or comparing tests with the CEFR. One group of Thai researchers was interested in setting a national standard. Hiranburana et al. (2018) introduced the Framework of Reference for English Language Education in Thailand – (FRELE-TH), which is based on the CEFR to be a shared basis for reflection and communication among different partners and practitioners in English language education in Thailand, including curriculum or syllabus planning, and textbook and course materials development.

**5.1.2 The Quality of the Developed Academic English Proficiency Test**

According to Bachman (1990), to describe the processes of the test development, a dominant framework includes five stages: 1. initial planning; 2. design; 3. operationalization; 4. trialing; and 5. assessment use. It can be seen that the trialing stage is a necessary phase to be implemented before the actual use of the test. The results from the trialing can signify test usefulness (Bachman and Palmer, 1996), a concept defined as the qualities of language tests.

In this study, construct validity, authenticity, interactiveness, and impact and practicality were evaluated by asking experts in the fields of language teaching and testing to consider these aspects. Since the TLU domain was emphasized, the test tasks corresponded to the authentic language use. This can reduce the differences between the artificial testing situation and how language will be used by students in the future. When these aspects have been investigated, it can reassure the teacher that the test is created for a certain group of students and it shows that the test designers already have gathered information about the target group. Statistically, the reliability or the consistency of the scoring has been calculated. The high values obtained can be good evidence supporting the idea that the developed test is consistent and stable in measuring what it was intended to measure.

According to the above, test development can also be seen as a circular and continual process. This includes feeding back the knowledge and experience gained at different stages of the process into a continuous re-assessment of a given test and each administration of it.

**5.1.3 The Concurrent Validity of the Developed Academic English Proficiency Test**

As concurrent validity is one approach of criterion validity. It can describe the efficiency of the estimation regarding an examinee's performance on some outcome measures. The test scores

can be useful if they provide a basis for the precise prediction of certain criteria (Lin & Yao, 2019). This means that once some types of tests are created, validity needs to be examined since it can measure how well a new test compares with a well-established one.

Therefore, this study has attempted to evaluate the concurrent validity of the developed test by correlating the scores obtained from the test with those from the IELTS. The IELTS was included in this study since it is one of the standardized English tests that are widely accepted. The scores from the IELTS are commonly used as requirements for university admission. Thus, the developed test can share the same purpose as the IELST (academic module).

According to the results of the concurrent validity analysis, the overall coefficients from the three parts of the developed test show that they have a positive moderate correlation with the scores from the IELTS. Thus, it can be said that the concurrent validity of the developed test is at a moderate level. The benefit of this result is predictability. In other words, the developed test can be used to predict the scores obtained from the IELTS. Moreover, Glen (2015) suggests the advantages of concurrent validity—that it is a fast way to validate data, and it may not be able to provide accurate comparable scores with other standardized scores. Hence, if exact comparable scores are required, other techniques such as standard-setting should be implemented.

For standard-setting, Bejar (2008 cited in Wudthayagorn, 2018) defines that it is a mapping methodology for the levels of language proficiency. It includes cut-off scores corresponding to the respective proficiency level. It can be considered as challenging in terms of implementation to prove the test validity for the use of test scores. It is the application of socially moderated technical methodological approaches, which has its origin in certification testing and educational assessment. This standard-setting has recently begun to receive attention in language testing (Kenyon & Römhild, 2013). It can be concluded then that standard-setting is a challenging approach used to set test scores for describing performance indicators. Since there are many kinds of test tasks, the methods used for establishing standards are also various.

In Thailand, a number of researchers have been working on setting standards. Wudthayagorn (2018) for example investigated if the Chulalongkorn University Test of English Proficiency, or the CU-TEP, can be mapped to the Common European Framework of Reference (CEFR) by employing a standard-setting methodology. According to that study, score users know which CUTEP score range falls into which particular CEFR level. Moreover, they would also know what test-takers can do with the English language at a particular CEFR level.

Similary, Athiworakun, and Wudthayagorn (2018) mapped the English standardized test created by the Language and Academic Services Centre, International College for Sustainability Studies, Srinakharinwirot University, with the CEFR. As they understand and foresee the significance of using English in general and academic contexts, the tests parallel with international standardized tests are used to measure the test-takers' English proficiency. Therefore, standard-setting will be another challenging task to be implemented after the test construct and the test quality have been revised.

**5.1.4 The Creation of Ability Bands for Describing the English Levels Obtained from the Developed Academic English Test**

This study could initiate the ability of bands to describe the details of the academic English proficiency of EFL university students in terms of academic listening, reading, and writing. Due to the fact that the creation of the test specifications includes the application of the TLU domain of EFL university students, the structures of the IELTS, and the descriptors of the CEFR, the developed ability of the bands are according to experts' considerations. Therefore, the bands can represent performance indicators for the academic English proficiency of EFL university students.

The regression analysis is to prove if the developed test can predict the IELTS scores. The regression equations obtained were used to estimate the IELTS scores and the cut-off scores for the developed test. According to this, the ability bands included the estimated scores for the developed tests when compared to the IELTS scores and CEFR levels together with the ability descriptors.

The advantages of the obtained ability bands are in accordance with Read's suggestions (2015), as he suggests that the conceptual frameworks of English proficiency can be connected to the proficiency testing and this has been recognized as one of the basic purposes of language assessment. Moreover, Read (2015) stated that the focus of English proficiency assessment is on the learners' ability to use the language for functional communication. As a result, the results obtained from the proficiency tests can be used for various purposes. Since the model performance indicators are functional, the measurable indices of the language domains are aimed at the targeted age/developmental levels of English language learners, and the setting of performance standards can be considered as one of the most challenging matters confronting language test developers.
Moreover, these indicators can be applied to the creation of standards or standard-setting.

Standard-setting or mapping test scores to the descriptions of language skills expressed within a scale of levels of competencies is a way of attributing meaning to test scores and of

translating test scores into the descriptions of test-takers' abilities (Kane 2012 cited in Knoch & Frost, 2017).

### 5.2 Recommendations

With the attempts of Thai researchers to create national standards for English proficiency, many educational institutes have been creating and developing their own English proficiency tests. The main purpose of this study aimed at exploring the academic English performances of EFL university students. The findings were mainly expected to assist in the design and development of a test that could be used as a valid and reliable indicator in terms of academic English proficiency. However, the generalizability of the results is limited by the number of participants, which cannot be representative of the whole population of EFL university students in Thailand. Further studies that include repetitiveness and sufficient population size are highly recommended.

Moreover, the following recommendations can be made, particularly for future studies.

1. The findings regarding the TLU domain for the academic English skills needed for EFL university students can be applied to further studies relevant to the design of training courses or other alternative assessments.

2. Task analysis is recommended as the primary step when designing a course or a test. In this way, the contents can meet the needs of students.

3. The test tailored from the TLU domain would be an effective placement test for assessing the academic English proficiency of EFL university students.

4. According to the present rapid changes in technology, technology should be considered for the development of language testing so that test authenticity can be increased.

5. Since the focus of this study was to develop a test for three main skills—listening, reading, and writing for academic English—it is recommended that tests for the speaking skill be the focus of future research.

## ACKNOWLEDGEMENTS

This research was funded by the Science and Technology Research Institute, King Mongkut's University of Technology North Bangkok, Thailand.

# REFERENCES

Ansastasi, A. & Urbina, S. (1997). Psychological Testing. Upper Saddle Rever, NJ: Prentice Hall.

Athiworakun, C. & Wudthayagorn, J. (2018). Mapping Srinakharinwirot University – Standardized English Test (SWU-SET) onto the Common European Framework of Reference (CEFR). Suranaree J. Soc. Sci. (12) 2: 69 – 84.

Bachman, L. F. (2004). Statistical Analyses for Language Assessment. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667350

Bachman, L.F. & Palmer, A.S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press. https://doi.org/10.1177/026553229601300201

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

Bamford, J. (2006). International students and their experiences of UK higher education. Paper presented at the Society for Research into Higher Education Annual Conference, Brighton, UK.

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In Bachman, L. F. and Cohen, A. D. (eds.), Interfaces between second language acquisition and language testing research. Cambridge: Cambridge University Press, 112-140. https://doi.org/10.1017/CBO9781139524711.007

Buck, G. (1991). The testing of listening comprehension: An introspective study. Language Testing 8: 67—91. https://doi.org/10.1177/026553229100800105

Cambridge. (2020). International language standards. Retrieved from https://www.cambridgeenglish.org/exams-and-tests/cefr/

Common European Framework of Reference for Languages-CEFR. (2018). Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Craighead, W.E. & Nemeroff, C.B. (2004). The Concise Corsini Encyclopedia of Psychology and Behavioral Sciences (3rd Ed.).

Darr, Charles. (2005). A hitchhiker's guide to validity and reliability. SET Magazine. 55-60. https://doi.org/10.18296/set.0639

Dornyei, Z. (2011). Research Methods in Applied Linguistics. New York: Oxford University Press.

Drummond, R.J. (1996). Appraisal Procedures for Counseling and Helping Professions- (4th ed.). Upper Saddle Rever, NJ.: Prentice Hall.

Ebel R, L. (1979). Essentials of education measurement. New Jersey: Prentice Hall.

Glen, S. (2015). Statistics How To: Concurrent Validity Definition and Examples Retrieved from https://www.statisticshowto.datasciencecentral.com/concurrent-validity/

Han, E. (2007). Academic Discussion Tasks: A Study of EFL Students' Perspectives EFL Journal 9,1: 8-12.

Hawkey, R. (2004). A modular approach to testing English language skills: the development of the Certificates in English Language Skills (CELS) examinations. New York: Cambridge University Press.

Hinkle, D. E, William, W. and Stephen G. J. (1998). Applied Statistics for the Behavior Sciences. 4th ed. New York: Houghton Mifflin.

Hiranburana, K. et al. (2018). Framework of Reference for English Language Education in Thailand – (FRELE-TH) Based on the CEFR: Revisited in the English Educational Reform. Pasaa Paritat 33, 51 – 91.

Kenyon, D.M. & Romhild, A. (2013). Standard Setting in Language Testing. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781118411360.wbcla145 https://doi.org/10.1002/9781118411360.wbcla145

Knoch, U. & Frost, K. (2017). Setting Empirical Standards on the Diagnostic English Language Needs Assessment (DELA). The University of Melbourne: Language Testing Research Centre.

Lin, W.L. & Yao, G. (2019). Concurrent Validity. Springer Nature: Switzerland.

McIntire, S. A., & Miller, L. A. (2005). Foundations of psychological testing (2nd ed.). Thousand Oaks: Sage Publishing Co.

McNamara, T., & Roever, C. (2006). Language Testing: The Social Dimension. UK: John Wiley & Sons.

Murphy, K.R. & Davidshofer C.O. (1998). Psychological Testing: Principles and Applications. Prentice Hall: USA.

O'Connell, M. (2016). What role should universities play in today's society? Retrieved from https://theconversation.com/what-role-should-universities-play-in-todays-society-63515

Office of Educational Assessment. (2020). Understanding Item Analyses. Retrieved from https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/

Read, J. (2015). Assessing English Proficiency for University Study. UK: Palgrave Macmillan. https://doi.org/10.1057/9781137315694

Sattle, J. (1992). Assessment of Children: Revised and Updated Third Edition. San Diego, CA: Jerome Sattler, Inc.

Scarcella, R. (2003) Academic English: A conceptual Framework. Linguistic Minority Research Institute Newsletter. University of California, Santa Barbara. Retrieved from: http://iteslj.org/Articles/Uribe-AcademicEnglish.html

Sornkam, B, Person, K. R., & Yordchim. S. (2018). Reviewing the Common European Framework of Reference for English Language in Thailand higher education. Paper presented at Graduate School Conference, Bangkok, Thailand.

Spolsky, B. (1989). Conditions for second language learning. Oxford: Oxford University Press.

Stevens, Stanley. (n.d.) On the Theory of Scales of Measurement. Retrieved from https://cehs01.unl.edu/aalbano/intromeasurement/mainch2.html

Wudthayagorn, J. (2018). Mapping the CU-TEP to the Common European Framework of Reference (CEFR). LEARN Journal: Language Education and Acquisition Research Network Journal (11) 2: 163 – 180.

Yuyun, I., Meyling, M., Laksana, N., & Abenedgo, D. (2018). A Study of English Proficiency Test among the First Year University Students. Journal of Language and Literature, 18(1), 1 – 8. https://doi.org/10.24071/joll.2018.180101