HU-General English Test: A Preliminary Study

พัชรี เชบ-บึนเนอร์

Patcharee Scheb-Buenner

ดร. อาจารย์ประจำหลักสูตรภาษาอังกฤษธุรกิจ วิทยาลัยนานาชาติดิษยะศริน
มหาวิทยาลัยหาดใหญ่

Background

Tests are powerful tools in education. Tests play a fundamental role in evaluation and assessment. Tests are used in a wide range from a small scale (quiz, mid term exam) up to large scales such as standardized tests (TOEFL, IELST, etc.) Specifically language proficiency tests play a major role in university admissions. Each language teacher needs tests and tests have a variety of forms and different purposes. Hummel (1997) explains the uses of standardized test as follows: 1) selection and placement of students into various programs; 2) to diagnose specific strengths and weaknesses associated with learning, performance in school, emotional problems, etc. When a test is used as an aid to identifying/diagnosing a problem, it will most likely also be useful in identifying necessary remediation for the deficit; and 3) these tests are most commonly used, in education, for evaluation purposes to determine how students are progressing compared to others (in the school, district, state, region, nationally), and to measure the effectiveness of the instruction and curriculum of the school.

Test can be used for all admissions i.e. undergraduates and graduates levels. Secondly this test can be a benchmark or one of requirements for graduation of international students to start the program or an exit exam as a bench marker before students graduate from university to assure prospective employers that the students have sufficiently a good command in English. Test takers can apply their test scores when needed such as to continue study in a particular institute. Prappal (2008) illiterates that most Thai teachers lack knowledge about language testing, assessment and evaluation, even though there have been training. This issue should be stressed because tests are very crucial to language teaching and learning. Standardized test means any empirically developed examination with established reliability and validity as determined by repeated evaluation of the method and results (Mosby’s Medical Dictionary, 2009). Standardized tests are instruments that measure and predict ability/aptitude and achievement.

Therefore a number of educational institutions in Thailand have produced their in-house standardised tests to cope with the great demands in English proficiency requirement. There have been a great number of well-known standardized tests in Thailand developed by public universities, such as CU-TEP, TU-GET, PSU-GET.

The aims of the tests generally are to measure students' English proficiency levels, especially a placement exam for entering educational institutions. Hatyai University (HU), as a fast growing private university in the southern Thailand, aims to produce a standardised test for described purposes. To produce a standardised test, a research study needs to be conducted to gain quality characteristics of exam.

There are a number of studies showing the predictive power of standardized tests. CU-TEP is one of well-known standardizes tests; the test is administrated to all first and fourth year students. One aspect of applications of standardized test is for prediction or relationship indication between English proficiency and academic achievement. There are many studies concluding such a relationship. Light et al (1987) revealed the significant correlation between TOEFL and grade point average and academic majors. Enedina García-Vázquez, et al (1997) also revealed significant correlations between reading and writing in Spanish and achievement scores, as well as grade point average. The strongest relations were found between Written Language and academic success. Abedi (2002) studied English Language Learning students (ELLs) and non-ELLs. The former performed lower than the latter group in reading, math, and science. Butler and Castellon-Wellington (2005) also found the concurrent performance As expected, the students with limited English proficiency (LEP) in the sample performed less well than the non-LEP students. For both 3rd grade and 11th grade, the English Only (EO) students outperformed Fluent English Proficient (FEP) students and LEP students on all the Stanford 9 subtests, with the FEP students outperforming the LEP students. Prappal (2008) summarized the three factors influencing the increase in standardized tests in Thailand which are increases in numbers of international programs, high demand in English proficiency, internationalization of educational institutions, and economical development. I added that the increasing numbers of students continuing their studies. Assumption College Thonburi (ACT) use IELST and CU-TEP as exit exams for Grade 12 and Grade 11 respectively (Sanongguthai, 2011). The results of this study guide ACT teachers to adjust teaching strategies. Moreover the students show that they are prepared themselves to study in international programs. They expect the teachers to provide the appropriate language input needed for IELST.

Research Objectives

This preliminary study aims to develop test specifications based on the standard criteria and to investigate the quality of the test which are validity, reliability, acceptable difficulty index and discrimination power.

Literature Review

Althouse (2001) explains the test development to meet the quality of reliability and validity. First conduct task analysis, this means the systematic method of analysis is required to obtain knowledge and skills essential for students. Next develop a test blueprint or specification . This step is to ensure that the test forms are consistent. Important information is test purpose, description of target students, total number of exams, number of items per objective , content out line, exam formats. After that test items are developed. Subsequently the test items will be reviewed and validated in terms of correctness, accuracy, relevance, and clarity which this step needs a reviewer or expert to review and validate the test. Next the test will be administered with candidates. The candidates or students will sit the tests for an appropriate time. In the next step the tests collected will be analysed item difficulty and item discrimination and reliability. The low quality test items will be flagged or cut off. After obtaining the test with good quality items, more items should be added or develops a parallel set of tests to have a multiple form of test. Then passing scores should be set. Lastly ongoing maintenance should be conducted.

Theoretical framework

Widdowson (1978, 1983), Savignon (1972, 1983) and Canale and Swain (1981) broaden view of language ability to communicative competence. Their works have been highly influential to language testing. Language use are viewed as a creation of discourse, or situated negotiation of meaning, language ability as multicomponent and dynamic. Language should not be regarded an isolated trait and its discoursal and sociolinguistic aspects of language use should be taken into account. Canale and Swain (1981) assert theory of communicative competence. They defined the term in association with four components. Grammatical competence includes knowledge in words and rules. Sociolinguistic competence refers to appropriateness of language uses in social context to fulfill communicative functions. It deals with the use of appropriate grammatical forms for different communicative functions in different sociolinguistic contexts. Strategic competence means appropriate use of communication strategies, such as the use of reference sources, grammatical and lexical paraphrase, requests for repetition, clarification, slower speech, or problems in addressing strangers when unsure of their social status or in finding the right cohesion devices. Discourse competence is defined in terms of cohesion and coherence in different types of text.

This study aims to build on different conceptual frameworks. First regarding the communicative competence theory, NIEST has designed a framework of English proficiency for measuring students when taking a national entrance exam. Wiriyajitra (2002) illuminates the standards of English proficiency which was proposed by NIESR has two goals; each goal consists of a variety of standards which express proficiency as follows:

Goal 1: To use English to communicate in social settings both inside and outside the university. Standard 1: Students will use spoken and written English for personal statement, and for enjoyment and enrichment. Standard 2: Students will use spoken and written English to participate appropriately in social interaction.

Standard 3: Students will recognize and understand cultural differences.

Standard 4: Students will use appropriate learning strategies to extend their communicative competence.

Goal 2: To use English to help achieve personal and academic goals and to promote life-long learning: Standard 1: Students will use English to access and process information and to construct knowledge in both spoken and written forms.

Standard 2: Students will use English to participate in academic contexts.

Standard 3: Students will use appropriate learning strategies to acquire, construct, and apply academic knowledge and to develop critical thinking skills.

Regarding NIEST'S goals and standards, it appears embracing and corresponding to communicative competency as previously described. Table below demonstrates the relationship between each standard and the competence.

Goal and Standard	**Grammatical competence**	**Sociolinguistic competence**	**Strategic competence**	**Discourse competence**
1.1 Spoken and written English for personal statement, and for enjoyment and enrichment.	√		√	√
1.2 Spoken and written English to participate appropriately in social interaction.			√	√
1.3 and understand cultural differences		√		
1.4 appropriate learning strategies to extend their communicative competence.			√	
2.1 use English to access and process information and to construct knowledge in both spoken and written forms.	√	√		
2.2 use English to participate in academic contexts.	√	√		
2.3 use appropriate learning strategies to acquire, construct, and apply academic knowledge and to develop critical thinking skills.	√		√	

Research Methodology

This paper is a preliminary report which aims to describe the quality of the first draft of HU GET. 106 test items were designed based on the framework previously discussed. The test was proofread by an American lecturer and checked its comprehensibility. The listening test part was produced by DRIC foreign staff, DRIC foreign students and an Australian. The sound quality was checked. Three students which I taught were asked to read the test if there is any point, particularly test directions, on the test they might not understand or confused them. They agreed that they could understand the test.

The test first was more than 130 items, some irrelevant items or passages were taken out. The test consisted of 106 items in a multiple choice format with four distractors. This format was applied because its advantages over other test format in terms of directness, comprehensiveness, easiness of scoring and administration (Ebel, 1979).

The test was administered to 60 HU students. The test was analysed its quality. The participants were allowed to take the test with unlimited time in order to find an appropriate time allotment for an actual test or the next stage. Most participants finished the test within 2 hours, but some last students took up to 2.30 hours. An item analysis was carried out to determine Item difficulty or Facility (IF) and Item Discrimination (ID). The acceptable IF ranges between 0.20 and 0.80. ID index was below 0.25, the items needs removed or improved (Althouse, 2001). Kuder-Richarderson formula 20 was used to calculate the reliability of the pilot test. A 0.80-0.99 reliability was accepted based on Ebel (1979). However the statistic program software employed in this study has a larger range of reliability which is above 0.70. The researcher also has an informal interview with the student participants. The researcher taught some of the student participants so she managed to learn the students' opinions in classroom. Descriptive statistics used were mean scores and percentage. Inferential statistics used was Kuder-Richardson formular 20 to calculate reliability of this pilot test. This formular was applied because it was approved in language studies to estimate an internal consistency reliability of test that has a single form (Brown, 1985). Quality characteristics of test as mentioned will be achieved by a different means such as content validity by expert reviews, reliability by a statistic software.

Results and Discussions

Test specification

The test specification has developed based on the framework as discussed previously. The format and part of the test followed the well-accepted standard tests. Multiple choices were applied and test specification.

The test developed was analysed its relevancy based on the framework or standard discussed previously in the following.

Standard 1.1 refers to many dialogues in the listening part related to personal statement, enjoyment, and enrichment. The dialogues to check the test takers understanding on information in the conversation related to making an appointment with a doctor, a man looking for a restaurant in a mall, also the dialogue.

Standard 1.2 focuses on uses the language appropriately in social interaction. The dialogues about job interview and looking for a restaurant which it is required a speaker to use different registers of language to communicate.

Standard 1.3 refers to cultural difference, it is related to sociolinguistic competence in terms of using pronouns since English language is normally necessary to have a pronoun as a subject in a sentence. This standard can be interpreted two ways. One is the concept of use the language to learn different cultures from different countries such as Passage about Nepal which describes big holidays in Nepal. The other is to learn the culture of the language. This is seen in Dialogue 4, it shows a practice in seeing a doctor required an appointment. For example,

25. When can she see the doctor?

a. At 2 pm Tomorrow

b. At 2.30 pm Tomorrow

c. At 1 pm Today

d. At 2.30 pm Today

Standard 1.4 is that learners can use appropriate learning strategies to extend their communicative competence. The test takers need to use different strategies to answer and complete the test .

Standard 2.2 requires learners to use English to participate in academic contexts. This can be seen in Passage 5(mosquitoes), sentences in error identification part which used structures of passive.

Standard 2.3 focuses on use appropriate learning strategies to acquire, construct, and apply academic knowledge and to develop critical thinking skills. Passage 2 (Nepal) and Passage 3 (JA) have questions which require the test takers to think critically beyond the level of comprehension as seen in the following.

31. According to the Proceedings of the National Academy of Sciences, it views the happiness as____________.

a. environmental issue

b. emotional issue

c. human's common issue

d. economic issue

This question from Passage 3 (Happiness) requires the test takers to think critically and to be able to imply on what they read. Overall the test specification and content of the test were also validated by an expert.

Test Administration and Item Analysis

The findings of this preliminary test focuses on the test item analyses and discuss good quality poor quality items The descriptive data and test items were classified on indexes of difficulty and discrimination .

60 participants administered the 106 test items for 2.30 hour. At the beginning, the researcher informed the participants this trial test and research being conducting. The researcher proctored the test by herself. The participants started with the listening part; they took about 40 minutes to complete this part. The participants took about 1.40 hours to complete the rest of the test. The maximum score was 58 and the lowest was 18. Mean score was 31. The reliability was 0.73.

The listening Comprehension part

There were 29 items of this part consisting of four short news reports, four dialogues and one announcement. Each extract had two to four questions. The participants were given one minutes for two item questions and up to two minutes for four item questions before next part started. This listening part took approximately 40 minutes. The IF index of this part ranged between 0.16-0.56 which most are considered difficulty. The IF index is lower it means the test is more difficult.

Table 1: IF Index of Listening Comprehension Part

Items	Evaluation of IF
7, 18, 27 (3)	Highly difficult (below 0.20)
1,2, 10, 11, 12, 13, 15, 16, 19, 20, 22, 24 (12)	Rather difficult (0.40-0.59)
6, 8, 9, 14, 17, 21, 25, 26, 28, 29 (10)	Satisfactorily difficult (0.20-0.39)
4, 3, 5, 23 (4)	Easy (below 0.60)

There are 3 items which were considered very high (below 0.20). It is approximately 10% of the test in the listening part. The high difficult items were from dialogues and announcement and one question asking about a big number. The moderate difficulty items are 10 items or 34%. The questions from the long conversations or dialogues are found rather difficult. The easy ones (14%) are mostly from short dialogues. This listening part has very a small number of highly difficult items. It refers that the part is not very difficult to that participants. Merely 25% of the participants reported that the listening part is the most difficult.

The Error Identification Part

Grammatical knowledge is evaluated in this part. This part is commonly used in well-known standardized tests (CU-TEP, TU-GET). There are 39 items in this part which item analysis reveals as shown on Table 2

Table 2: IF Index of Error Identification Part

Items	Evaluation of IF
31, 38, 42, 49, 60 (5)	Highly difficult (below 0.20)
30, 33, 35, 37, 39, 40, 41, 43, 44, 45, 47, 48, 52, 53, 55, 57, 59, 61-66, 68 (24)	Rather difficult (0.40-0.59)
32, 36, 46, 50, 51, 54, 56, 58, 67 (9)	Satisfactorily difficult (0.20-0.39)
34 (20)	Easy (below 0.60)

There are 29 (24+5) out of 39 items or 74% which high indexes of difficulty. The items considered relatively difficult are connectors, part of speech (verb, subject form), subject-verb agreement and plural forms, articles, passive forms, adverb-adjective forms. Regarding a high percentage of difficult items, this part is considered difficult corresponding to the interview data collected from the participants. Approximately 75% of them agreed that this part was the most difficult to them.

The Reading Comprehension part

This part consisted of 7 passages with 38 items. Two were a cloze test. Table revels the IF indexes of this part.

Table 3: IF Index of Reading Comprehension Part

Items	Evaluation of IF
74, 75, 77, 87, 91, 93, 97, 99, 103 106 (10)	Highly difficult (below 0.20)
70, 71, 72, 73, 77, 78, 79, 80, 81, 83, 85, 86, 89, 90, 93, 94, 95, 96, 98, 101, 104, 105 (22)	Rather difficult (0.40-0.59)
76, 83, 88, 100, 102 (5)	Satisfactorily difficult (0.20-0.39)
69 (1)	Easy (below 0.60)

This part consists of 32 (10+22) items or 80%. It means that this part is difficult. However the participants did not mention that this part is difficult to them. They merely reported that there were many passages and long ones.

Discrimination Power

Table 4: ID Index of All Items

Items	Evaluation of ID
3, 4,8,11, 14, 15, 17, 21, 28, 29, 32, 34, 46, 47, 51, 55, 56, 58, 59, 67, 69, 70, 73, 76, 82, 93, 102 (27)	Highly discriminated (more than 0.40)
5, 6, 9, 10, 19, 23, 25, 35, 36, 44, 48, 61, 65, 70, 84, 85, 89, 90, 94, 104, 105 (21)	Moderately discriminated (0.30-0.39)
20, 22, 37, 54, 64, 72, 79, 83, 97 (9)	Satisfactorily discriminated (0.20-0.29)
1, 2, 7, 12, 13,16, 18, 24, 26, 27, 30, 31, 33, 38, 39, 40, 41, 42, 43, 45, 49, 50, 52, 53, 57, 60, 62, 63, 66, 68, 74, 75, 77, 78, 80, 81, 86, 87, 88, 91, 92, 95, 96, 98, 99, 100, 101, 103, 106 (49)	Poorly discriminated (below 0.20)

Due to the limited space available here, the findings of discrimination power do not separate based on the parts. This test has 46% of poor discrimination power. It means that 56% of the test can discriminate between low and high proficiency students.

Items with good quality

Having analysed the pilot version, it was found that it needs a certain amount of adjustment and improvement for the next step of this test development. In brief, there are 53 items which consist of the items with high discrimination power and with satisfactorily and moderate difficulty indexes. The low discrimination power and very high difficulty level were considered poor quality items. There are several reasons to explain the quality of this test. This is a pilot version which was constructed at the first time; it has not been well developed. Therefore there are poor quality items. The test specification does not provide details and well-balanced of contents (grammar or vocabulary). The student participants may not take the test seriously since they were aware that the test results would not affect them.

Conclusions and Implication

Test takers in any levels and any course or program seek for a confidence in what exam they will take, whether the contents are fully covered in order to evaluate their performance. Therefore, this pilot test is aimed to further develop as to its contents, item analysis quality, reliability and validity. Having earned these qualities, the test will definitely assure test takers the reliable and valid test results. The test consequently can be recognized or normalized in the future. Further development for this test is to review and refine the test specification, and to adjust the poor quality items and to add more items. Importantly the test will be administered with the actual population and be normalized. The exam should provide reliability. The exam should be able to distinguish between the test takers who are capable and not capable. This preliminary study shows how to develop a test to ensure validity and reliability. This study shows how to develop the test based on conceptual frameworks derived from theoretical frameworks. This test version also was piloted with participants to verify test items. This reliability of this test needs improvement. The quality of the test also needed improvement. The low quality will be improved by adjusting multiple choices. Regarding the error identification, this test does not have well-balanced grammar content so that the test specification should be revised and specify the proportion of grammar items clearly.

Some suggestions are made in order to develop a quality test. Training in testing, assessment and evaluation for language teachers should be provided. The awareness of test quality should be increased. Language teachers should analyse their test to obtain quality test items and develop a test bank so quality test items are collected and can be used in the next test. Having a test bank, it will facilitate the language teachers' teaching since they do not need to spend time on writing a new exam. When a good quality of test is constructed, it will reveal power of predictability.

References

Abedi, J. (2002) Standardized Achievement Tests and English Language Learners: Psychometrics Issues Educational Assessment, 8(3), 2002.

Althouse A. L. (2001) Test development: Ten steps to a valid and reliable certification exam.

Bachman, L. F. (1989) Assessment and Evaluation. Annual Review of Applied Linguistics, 10, 210-226.

Canale, M. and Swain, M. 1980: Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47.

Butler, F. A. and Castellon-Wellington, M. (2005) Students' Concurrent Performance on tests of English language proficiency and academic achievement. In The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation From Three Perspectives. CSE Report 663 Center for the Study of Evaluation. National Center for Research on Evaluation, Standards, and Student Testing Graduate School of Education & Information Studies University of California, Los Angeles Los Angeles, 47-78.

Ebel, R. L. (1979) Educational test and measurement; examinations; evaluation: design and construction; interpretation Eaglewood Cliffs N.J., Prentice-Hall.

Enedina Garcia-Vazquez, E., Luis A. Vazquez, Isabel C. Lopez AND Wendy Ward (1997) Language Proficiency and Academic Success: Relationships Between Proficiency in Two Langauges and Achievement Among Mexican American Students. Bilingual Research Journal: The Journal of the National Association for Bilingual Education, 21(4).

Light, R., Xu, M. and Mossop, J. (1987) English Proficiency and Academic Performance of international Students TESOL Quarterly, 21(2) 251–261.

Prapphal, K. (2008) Issues and trends in language testing and assessment in Thailand Language Testing, 25, 127-140.

Savignon, Sandra, L. (1983) Communicative Competence: Theory and Classroom Practice Texts and Contexts in Second Language Learning. The Addison-Wesley Publishing Co.

Sanongkuthai, S. (2011) Washback of IELST on the Assumption College English Program. Language Testing in Asia.

Wiriyajitra, A. www.apec.edu.tw/ retrieved on 24 July 2012.

Widdowson, H.G. (1978) Teaching language as communication. Oxford: Oxford University Press.

Appendix

Test Specification

Three parts:

1. Reading comprehension (40 items):

   - Main idea, detail information, pronoun reference, unknown vocabulary, conclusion

   - Required items 60 in the following:

   1.1 Eight passages (5 items/passage): multiple choices 40 items

   1.2 Four cloze passages (5-6 items/passage): MC

2. Writing ability (30 items)

   - Required items 60 in the following:

   2.1 Sentence completion (30 items): MC

   2.2 Error identification (30 items): MC

3. Listening skill (30 Items)

   - Main idea, detail information, pronoun reference, unknown vocabulary, conclusion

   - Required items 50 in the following:

   3.1 Six short dialogues (3-5 items/dialogue): MC

Four long dialogues/announcement/etc. (5 items/dialogue): MC