

English Proficiency Test Analysis and a Development of Test Bank Software and Computer-aided Examinations

Sripen Srestasathien¹

บทคัดย่อ

การวิจัยนี้เป็นการวิจัยเชิงพัฒนามุ่งหาแนวทางเพื่อวิเคราะห์ข้อสอบ สร้างข้อสอบมาตรฐานวัดความรู้ภาษาอังกฤษสำหรับบุคคลทั่วไป สร้างซอฟต์แวร์ฐานข้อมูลคลังข้อสอบ สร้างสนามสอบภาษาอังกฤษระบบคอมพิวเตอร์ และสร้างระบบประมวลผลการสอบแบบอัตโนมัติ กลุ่มตัวอย่างที่ใช้ในการทดลองประกอบด้วยผู้สมัครสอบคัดเลือกเข้าศึกษาต่อระดับบัณฑิตศึกษา มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ที่สอบวิชาภาษาอังกฤษทั่วไป ตั้งแต่ปีการศึกษา 2545-2547 จำนวน 11,134 คน นักศึกษาระดับปริญญาตรีชั้นปีที่ 4 วิชาเอกภาษาอังกฤษ จาก 4 มหาวิทยาลัย จำนวน 340 คน และนักศึกษาระดับบัณฑิตศึกษา มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ 20 คน ที่สมัครเข้ารับการทดสอบวัดความรู้ภาษาอังกฤษด้วยระบบคอมพิวเตอร์ที่ผู้วิจัยพัฒนาขึ้น เครื่องมือที่ใช้ในการวิจัยคือข้อสอบภาษาอังกฤษทั่วไป ซึ่งมีลักษณะใกล้เคียงข้อสอบ TOEFL จำนวน 11 ชุด ชุดละ 100 ข้อ หลังจากทำการทดสอบแล้วผู้วิจัยนำคะแนนและตัวเลือกคำตอบแต่ละข้อมาวิเคราะห์หาค่าความยากง่ายและค่าอำนาจจำแนก คัดเลือกข้อสอบที่ได้มาตรฐานไว้ ปรับปรุงข้อสอบที่มีข้อบกพร่องเล็กน้อย แล้วจัดเก็บในซอฟต์แวร์ฐานข้อมูลคลังข้อสอบที่ผู้วิจัยสร้างขึ้นด้วยโปรแกรม MySQL, OOP และ PHP ผู้วิจัยได้พัฒนาสนามสอบด้วยคอมพิวเตอร์ระบบ LAN ประกอบด้วยเครื่องแม่ข่าย 1 เครื่องและเครื่องลูกข่าย 20 เครื่อง

และจัดทำโปรแกรมประมวลผลการสอบเพื่อพิมพ์ผลการสอบให้ผู้เข้าสอบทราบทันทีหลังการสอบ ผู้วิจัยได้ทดลองใช้สนามสอบกับกลุ่มนักศึกษาที่สมัครใจทดลองสอบ ผลการวิจัยพบว่าข้อสอบที่วิเคราะห์ส่วนใหญ่อยู่ในเกณฑ์มาตรฐาน สนามสอบใช้งานได้ ซอฟต์แวร์คลังข้อสอบมีความจุข้อสอบ 200,000 ข้อ สามารถสุ่มข้อสอบให้ผู้เข้าสอบได้ตามที่ออกแบบไว้ เก็บข้อสอบเก่าและข้อสอบที่พัฒนาขึ้นใหม่ที่อยู่ในเกณฑ์มาตรฐานไว้ 650 ข้อ ผู้บริหารระบบสามารถพิมพ์ผลการสอบได้ทันที

คำสำคัญ: ข้อสอบมาตรฐาน การสอบวัดความรู้ภาษาอังกฤษ ซอฟต์แวร์คลังข้อสอบ การจัดการสอบด้วยระบบคอมพิวเตอร์

Abstract

This study was of a research and development type whose purposes were to analyze General English Proficiency Tests (GEPTs) used in post-graduate entrance examinations, to construct standardized English Proficiency Test (EPT) items, to develop item bank software and a computer-based test (CBT) system, as well as data processing system software. The subjects of the study were 11,134 post-graduate entrance exam applicants taking GEPTs in academic years 2002 to 2004 for obtaining admission to study

¹ Associate Professor, Department of Languages, Faculty of Applied Arts, King Mongkut's University of Technology North Bangkok, Tel. 08-1644-2676, E-mail: sripen95@gmail.com

in the Graduate School, King Mongkut's University of Technology North Bangkok (KMUTNB), three hundred forty of the students majoring in English at four universities, and 20 current KMUTNB post-graduate students. The two latter groups joined the project on a voluntary basis. Eleven GEPTs each consisting of 100 items similar to those of the TOEFL were used in the study. To standardize the test items, the answers to each item were analyzed to find their difficulty level and discrimination. The test items were then sorted out and improved. Appropriate items were stored in the test bank software constructed by the researcher using MySQL freeware, OOP, and PHP. A computer-based examination room was developed using the LAN system. It consisted of a server and 20 client computer sets. Twenty KMUTNB graduate students tried the test in the constructed CBT room. It was found that most analyzed test items complied with standard criteria. Six hundred and fifty items were saved in the constructed item bank software, which could store up to 200,000 items and could randomize items for use as designed. The constructed computer-based exam worked appropriately. The scores of the CBT test could be processed by the developed software and the results could be printed out by the test administrator immediately after the test was finished.

Keywords: Standardized Test, English Proficiency Test (EPT), Test Item Bank Software, Computer-based Test (CBT)

1. Introduction

English is a major language internationally. It currently plays an important role in education, research, knowledge transfer, careers, as well as everyday communication. In applying for a seat in universities

in both English or none-English speaking countries, applicants are often required to take an English proficiency examination or submit English proficiency scores obtained from some specific institutes such as TOEFL, IELTS, GRE, etc., for the university admission committee's consideration [1]. A number of companies currently require their job applicants to submit TOEIC scores. To obtain such scores, examinees have to spend a lot of money applying for the exam, resulting in great monetary loss each year to examination institutes abroad, as many people have to take tests repetitively before they can get the required scores. Realizing such a loss, the researcher would, therefore, like to construct a standardized test to be used at KMUTNB, and at a cheaper cost of the testing service, for those who want to evaluate their English language skills before they take their required test.

In addition, one of the main jobs of the Department of Languages, Faculty of Applied Arts, KMUTNB, where the researcher is working, is to provide an English proficiency examination for the university entrance examination for both undergraduate and graduate studies levels. At least two graduate entrance exams are arranged each year, and a new test must be written for each exam. Writing test items is quite a tedious task for English teachers. No analysis of those test items has been made to standardize the test items. If this had been done, appropriate items would have been kept for future use, and the teachers' burden would have been lessened. This inspired the researcher to analyze those used test items, standardize them, and provide more new items to keep them in a test item bank [2] to be constructed. She also aimed to initiate a reliable test center in order to provide a potential English proficiency test service for the university personnel, students, and outsiders at a reasonable cost. This would save test

writers' time and energy in providing test items for every exam, and it would also save money for the test-takers caused by taking the costly TOEFL, IELTS, or TOEIC repetitively. Further, the scores obtained from the center can be used for applying to graduate study seats at KMUTNB and elsewhere. The center would serve as a place for people to try testing and checking their English ability as to whether it is appropriate to take the TOEFL, IELTS or TOEIC test or not. If not, they would know from these test results what aspect needs improvement.

2. Materials and Methods

Seven sets of previously-used General English Proficiency Tests of the KMUTNB post-graduate entrance examination, and four sets of newly-written tests, were used in the study. Generally, the KMUTNB English proficiency test consists of 5 main parts; namely, conversation, paraphrasing, error detection, grammar, and reading, making a total of 100 items. Only the parts that were similar to the TOEFL test were analyzed and standardized for this study. They included grammar, error detection, and reading parts. Conversation was, however, analyzed and adapted for the listening test in the study. The level of difficulty and discrimination of the test items were calculated according to the standard formula proposed by Jamornman [3], Davidson [4], Huges [5], and Bachman [6].

Before standardizing the older test items and providing new test items, the characteristics of and guidelines for constructing worldwide standardized tests, such as the TOEFL, IELTS, TOEIC, and Michigan Test, were investigated.

The subjects of this study included 11,134 post-graduate entrance exam applicants taking GEPTs from academic years 2002 to 2004 in order

to obtain admission to study in the Graduate School at KMUTNB. In addition to these applicants, 340 English majors from four universities, namely, Mahasarakham University, Mahasarakham Rajabhat University, Udorn Rajabhat University, and Ubon Rajabhat University, and 20 KMUTNB post-graduate students participated in this study. The two latter groups joined the project and took the adapted and newly-made test on a voluntary basis.

Then a database system, a test item bank for storing and sorting test items for future use and for providing an English proficiency test, a computer-based testing system, a test center, and a webpage were developed using MySQL freeware, Object-oriented Programming (OOP), and Hypertext Preprocessor PHP. MySQL is an open source database used for developing and managing a large database system, while OOP is a programming paradigm using “objects”—data structures consisting of data fields and methods together with their interactions. It was used to design the applications and the computer program in this study. As for the PHP, it is a general-purpose, server-side scripting language used for Web development to produce dynamic Web pages. In addition, a test score processing system that could announce and print test-takers' scores soon after the test was designed and constructed [7], [8].

3. Results and Discussion

With respect to the analysis of international standardized tests, only information on the TOEFL CBT will be presented in this article, as the construction of items in this study mainly followed it. A computer-based TOEFL has been continually improved by the TOEFL BOARD. It consisted of four major parts: listening comprehension, structure, reading comprehension, and writing. The listening part was made up of three kinds of

test items: a short conversation, a longer conversation, and a talk or lecture. In the short conversations, a man and a woman did the talk and a third person asked questions. The content of the talk was about US college student life, i.e. studying, friends, social functions, homework, doing something, accommodations, etc., while most longer conversations were about teaching and learning, and the talk and lecture part was related to academic matters and providing some information [9]. The TOEFL grammar test was divided into two parts: sentence completion and error detection. The sentence completion part asked test-takers to choose the grammatically correct answer to fill in the given sentences, while the error detection part required them to choose the part of sentence that was grammatically incorrect [9], [10]. Five to eight passages of general and academic knowledge with a length of 150-300 words each were given to the test-takers in the reading comprehension part. The total number of questions of this part was either 44 or 60, and the time allowed was 70 or 90 minutes [9], [10]. The last part concerned writing ability and it required test-takers to write an essay in 30 minutes in response to a single topic that appeared on the screen. Either typing the essay on the computer or writing it on paper was allowed [9], [10].

The test items in the short conversation listening and structure parts were randomized according to the test-takers' performance. They were divided into 3 levels of difficulty: simple, moderate, and difficult. The scores awarded varied according to the difficulty level; 1 point for the simple level, 2 for the moderate level, and 3 for the difficult level. The upper- or lower-level item given would be determined by the previous item score. If a test-taker got a moderate level item as the first question and could answer it correctly, he or she would be given a question at a more difficult

level, and vice versa. Test-takers had to answer every question—going back to previous questions was not possible. Regarding the reading part, examinees could go back and forth to questions connected with the same reading passage. The scores awarded for this part were the same for every item [9]-[11].

In this study, the test items in the GEPT were analyzed to find their level of item difficulty (P) and discrimination (r). According to Jamornman [3], Davidson [4], Hughes [5], and Saiyos and Saiyos [12], to calculate the item difficulty level of the test, where 1 point was given to the correct answer and 0 to the wrong one, and when the test was taken by a large number of examinees, 25%, 27%, or 33% of the higher-score and the lower-score groups had to be split from the whole group for calculation. For this study, 27% of each group was taken. The formula used for calculating the item difficulty level was:

$$P = \frac{H-L}{N} \quad [3]$$

where, H was the number of subjects in the high-score group, L was the number of those in the lower-score group, and N was the total number of the two groups in each test. The P-value ranged from 0-1. Any items with a P-value between 0.33-0.66 was considered moderately difficult, while those with a P-value lower than 0.33 were difficult and those whose P-value was higher than 0.66 were easy [3], [12], [13]. Nonetheless, the items with a P-value of 0.2-0.8 were considered appropriate for being retained in the test bank [4], [5].

With respect to item discrimination (r) calculation, the purpose of the method was used to discriminate the poor from the intelligent test-takers or students [4], [5]. The formula used was

$$r = \frac{H-L}{n1} \quad [3]$$

where, H was the number of test-takers in the high-score group that answered the item correctly, L was the number of poor examinees that answered the same item correctly, and n1 was a number of any of the two groups [3], [4], and [13]. The r-value ranged from minus number to 1. Jamornman [3] interpreted the r-value as follows: 0.33 - 0.66 demonstrated that the test item could considerably discriminate the poor from the intelligent test-takers, while a value below 0.32 and above 0.66 could highly discriminate the two groups. However, the minus r-value or an inverted r-value indicated that more poor examinees could answer that item correctly than the intelligent could. Such item was recommended not to keep it.

For the calculation for the P and r values in this study, a software was developed to make it faster and easier. Seven sets of a GEPT were computed and the results are presented in the table below:

Table 1 The results of the GEPT analysis

GEPT Set No.	No. of test-takers	27% of Answer Sheet to be analyzed	No. of items with p-value 0.2-0.8	No. of items with r-value 0.33-0.66
1	1,833	495	87	37
2	1,733	468	61	6
3	1,811	489	48	3
4	2,088	564	93	22
5	1,711	462	89	30
6	227	62	92	21
7	1,731	338	76	14
Total	11,134	2,878	546	133

Table 1 shows that the number of test-takers taking all tests, except test No. 6, was quite large, so 27% of the high-score and low-score groups were taken for calculation of item difficulty level and discrimination. The total number of the examinees was 11,134, while

the total number extracted for calculating was 2,878. The total number of items with a P-value of 0.20-0.80 was 546, while the total with an r-value range of 0.33-0.66 was 133. The 546 items were investigated in depth, item by item, to find out their P and r, as shown in some example items in Table 2 below.

Table 2 Level of difficulty and discrimination values of the reading part of a test

Item No.	H/L*	P	r	Item No.	H/L	P	r
81	505/197	0.62	0.55	91	239/74	0.28	0.29
82	383/151	0.47	0.41	92	166/145	0.28	0.04
83	240/63	0.27	0.31	93	423/156	0.51	0.47
84	334/243	0.51	0.16	94	293/138	0.38	0.27
85	357/147	0.45	0.37	95	110/90	0.18	0.04
86	193/76	0.24	0.21	96	217/106	0.29	0.20
87	199/94	0.19	0.19	97	174/157	0.29	0.03
88	294/174	0.41	0.21	98	215/149	0.32	0.12
89	383/173	0.49	0.37	99	238/136	0.33	0.18
90	186/104	0.26	0.15	100	121/125	0.22	-0.01

*H was the number of test-takers in the high-score group that answered the item correctly, and L was the number of examinees in the low-score group that answered the same item correctly.

Table 2 shows the results of the reading comprehension item calculation for the P and r values in a test, beginning from item 81. It can be seen that the P-value of all items was in the standard range of 0.20-0.80, except for items 87 and 95, while the r-value of most items did not comply with the standard criteria. Each item, therefore, was improved by the researcher and checked for its appropriateness by two experts that were native speakers and familiar with test construction before it was stored in the constructed test bank.

Four newly-constructed tests were tried with 340 undergraduate English-major students from four universities, and they were analyzed and improved through the same process as that of the seven GEPTs.

The results of the analysis are presented in the table below.

Table 3 The results of the analysis of four newly-constructed English proficiency tests

Test No.	No. of test-takers	27% of Answer Sheet to be analyzed	No. of items with p-value 0.2-0.8	No. of items with r-value 0.33-0.66
1	143	39	91	32
2	79	22	81	28
3	66	18	79	26
4	52	15	86	34
Total	340	94	337	120

The analysis of these four tests was undertaken using the same process as those old tests, using the self-constructed software for computation of the P and r-values. Actually, the number of the test-takers in each test was not large. All of the answer sheets could have been analyzed, but the researcher decided to use the software to calculate their P and r-values for the sake of speed and accuracy. As a result, 27% of the test-takers were analyzed. As shown in Table 3, the total number of the test-takers from the four different universities was 340, and the total number of answer sheets to be analyzed was 94. Three hundred and thirty-seven items had a P-value ranging from 0.20-0.80, and 120 items were in the criteria of the r-value. According to the analysis, although most items were at a difficult level, they lay in the acceptable range of the P and r values. The items whose r-value was not in the required range were improved. Hence, approximately 300 items from the newly-constructed tests could be put into the test bank. Every item was checked by two experts for its correctness and appropriateness before the tests were administered and after the improvement. In conclusion, for the test item construction and standardization, there were about 650 items; 100 listening test items, 300 structure items, and 250 reading comprehension

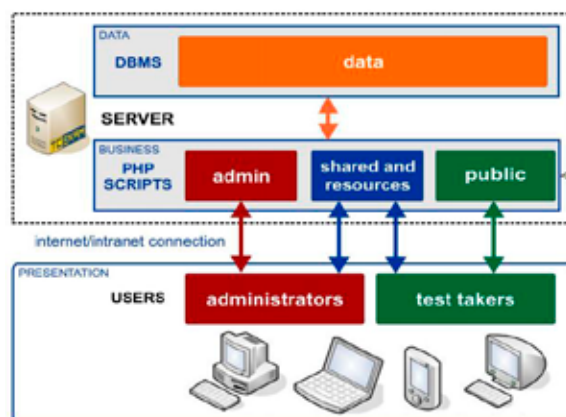


Figure 1 Design of the KMUTNB English Language Computer-based Test (CBT KMUTNB-ELT).

items were considered of standard and could be kept for future use in the test bank.

Regarding the development of the test bank and computer-based test system software, the construction of the test bank, item randomization programming, score processing software, and the webpage were designed and constructed using MySQL freeware, OOP, and PHP [2], [3], [7], [8], [14]. An expert in computer programming was hired to help with the software writing and system planning. The design of the system is displayed in Figure 1 above.

The system consisted of two main parts: the administrator's and the user's parts. All of the data, including test-items, the test-takers' necessary personal data form, a randomization system, a score processing system, etc., were kept in the server and taken care of by the administrator—the researcher herself. At the administrator site, she was able to manage all of the data related to the test, such as updating information, adding, deleting, or editing questions, processing scores, etc. See figure 2. The constructed test bank could store up to 200,000 items.

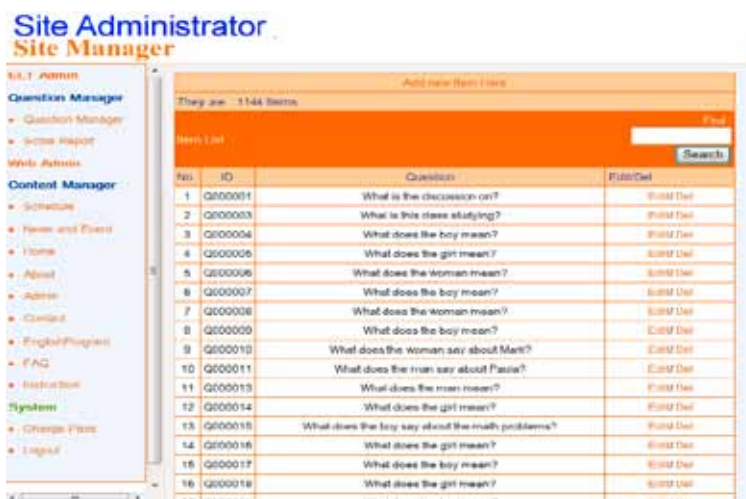


Figure 2 Administrator page for managing information, and test bank and test organization.



Figure 3 The web page of KMUTNB English Language Test Center (KMUTNB-ELT).

With respect to the user or client part, the test-takers, after applying for the test, would see the web page first. Figure 3 shows the KMUTNB-ELT web page.

At the beginning the researcher intended to construct an on-line system for test-takers to register

for the test, but she was not able to do this since the system had to rely on KMUTNB's Internet network system, which was not stable. So she had to change the system to a local area network (LAN); the test-takers had to come to the test center and apply for the test



Figure 4 Page for test-taker's required personal data.



Figure 5 Explanation and demonstration of how to do the listening test.

themselves. The process, however, did not take long. When everything was settled, the test-takers would be seated at any of the 20 client computers and they would have to register on the web page and enter the required information into the page that appeared (see Fig. 4).

The required information included the test-taker's name, office, university/college, educational background, telephone number, email address, user

name, and password. Once the information was filled in, he or she had to click on the button "send." The administrator would then register the students and ask them to enter their username and password on another page to begin the test.

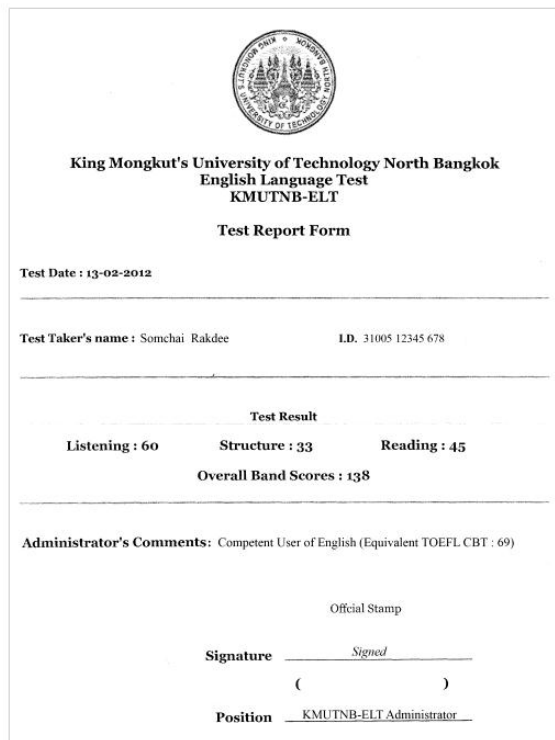
The system provided an explanation and a demonstration of how to do each part of the test. See Figure 5.

Figure 5 shows what the test-takers would see and hear on the demonstration or tutorial page of the listening test part 1. It gave instructions on and demonstrated how to do the short and longer conversation test items. The time remaining counter was also pointed out as well.

Once the test-takers understood how the system worked and were ready, they just clicked the button and a test item would pop up for them to do. They could choose to do any of the four parts first or later, but after submitting each item or part, they were not allowed to go back to the past item or part.

The constructed CBT KMUTNB-ELT was tried with 20 graduate students. Some problems were found before starting and while doing the test; more explanation was needed, and going back and forth to the questions in each reading passage was not possible.

Even though an attempt was made for the constructed system to follow the computer-based TOEFL test in every aspect, it was impossible for the present researcher to accomplish this as she wished, resulting in some differences between the TOEFL CBT and KMUTNB-ELT CBT. That is, since there was a limitation of inter-raters for the writing part, it was therefore cancelled. Two other main differences were found in two areas in this study; one in the conversation part and another in the reading part. In the short conversation section, items were randomized according to the scores awarded. The first two questions given to the test-takers were easy ones. One point each was awarded if they answered them correctly and then upper-level questions would be allocated. For the TOEFL any item—either simple, moderate or hard—was possibly given depending on the system. Another difference occurred in the reading part: six passages with five questions each were given to each test-taker in this test and after submitting the



King Mongkut's University of Technology North Bangkok
English Language Test
KMUTNB-ELT
Test Report Form

Test Date : 13-02-2012

Test Taker's name : Somchai Rakdee ID. 31005 12345 678

Test Result

Listening : 60	Structure : 33	Reading : 45
Overall Band Scores : 138		

Administrator's Comments: Competent User of English (Equivalent TOEFL CBT : 69)

Official Stamp

Signature _____ Signed _____
(_____)

Position _____ KMUTNB-ELT Administrator _____

Figure 6 Test-takers' CBT KMUTNB-ELT score report card.

answer to each item, they were not allowed to go back for revision, unlike the CBT TOEFL [8], [9]. The researcher found that designing and writing complicated software was a tedious and difficult task. If one missed something after finishing the entire process, it was difficult to get back and correct it.

As for the score-processing system, it worked very well. The score card could be printed out right away after the test-takers finished the test (see Fig. 6) above.

Regarding the scoring, 45 points constituted the full score of the short, randomized-conversation part and another 45 for the longer conversations and talks or lectures. The full score of the structure part with randomized items was 70, while that of the reading part

was 90, making a total of 250 points. The time allocated was two and a half hours. There was a counter informing the test-taker of the remaining time in the upper left corner. The points gained were roughly equated with those of the TOEFL. No problems were found during any part of the test, except for the reading part, which did not allow the test-takers to go back and forth to the questions for each passage. The randomization and scoring of items worked properly.

4. Conclusion

It cannot be concluded that the constructed system was proper since there were still many missing points. It required further corrections and modifications in order to follow the TOEFL computer-based test. However, this study might inspire those that are interested in this area. The utmost benefit gained from this study was the test item bank, which could store quite a large number of items for future and repetitive use.

5. Acknowledgement

The researcher was very grateful to the government for funding this research. My thanks go to Mr. Noppanai Imsamai for his patience to help me get through this long and tedious study, and to Assistant Professor Dr. Saroj Pullteap for his invaluable advice.

References

- [1] [http://www.riansingapore.com/index.php? lay=show&ac=article&Id=538638850&Ntype=6](http://www.riansingapore.com/index.php?lay=show&ac=article&Id=538638850&Ntype=6): Retrieved on March 7, 2012
- [2] S. Srestasathiern and F. Davidson, (2006). "Principles of Item and Task Bank Construction," *Journal of King Mongkut's Institute of Technology North Bangkok*, vol. 8, no. 6, December 2006 and <http://www.Thaiscience.info/Journal/article/Principles%20of%20item%20and%20task%20bank%20construction.pdf>
- [3] Utoomporn Jamornman, *Test Items: Construction and Development No. 11*, Bangkok: Funny Publishing, 1992 (in Thai).
- [4] Davidson, Fred. *Principles of Statistical Data Handling*, Thousand Oaks: SAGE, 1996.
- [5] Arthur Hughes, *Testing for Language Teachers*, Cambridge: CUP, 1989.
- [6] L.F. Bachman, *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press, 1990.
- [7] Jason W. Gilmore, *Beginning PHP and MySQL 5: From Novice to Professional*, New York: Apress, 2004.
- [8] Rusmus Lerdorf and Kevin Tatroe, *Programming PHP*, Sebastopol: O'Reilly Media, 2006.
- [9] A. Hagen, Stacy, *Test Bank for English Grammar Third Edition*, New York: Longman, 2003.
- [10] J. Rymniak, Marilyn, et al., *TOEFL CBT Second Edition*, New York: Simon&Schuster, 2002.
- [11] Rosechonporn Komolsevin, Sonthida Keyuravong, and Tinakorn-Araya Seehakom (Compilers), *TOEFL Handbook*, Bangkok: PSP, 2002.
- [12] Luan Saiyos and Angkana Saiyos, *Learning Measurement Techniques*, Bangkok: Suweeriyasarn, 2000 (in Thai).
- [13] Pawaros Butakeo, Major, (February 15, 2012). "Test Item Analysis," (in Thai). [Online]. Available: http://www.rta.mi.th/630a0u/file/item_analysis.doc
- [14] Meilir Page-Jones, *Fundamentals of Object-Oriented Design in UML*, New York: Dorset House Publishing, 2000.