

CHAPTER III

RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research methodology used in this study. The scope covers research design, population and samples, research instruments, data collection and data analysis.

The purposes of this study are to investigate the effects of test authenticity and test delivery mediums on reading comprehension scores and also to explore test takers' attitudes towards these two variables.

3.2 Research design

3.2.1 Investigated variables

1. Independent variables: The independent variables selected in this study are 1) test authenticity and 2) test delivery mediums. The former consists of two levels: authentic and inauthentic. The latter includes computer-based tests and paper-based tests.

2. Dependent variable: The dependent variable in this study is reading comprehension score.

3.2.2 Research procedure

The procedure was divided into three phases as follows:

Phase I: The development and validation of the research instruments. The instruments used in this study were divided into two main sets: 1) Reading Comprehension Tests and 2) The Interview Schedule. Each set was described as follows:

Set 1: Reading Comprehension Tests used in this study were composed of four versions:

The paper-based tests consist of two versions:

Version 1: Authentic English Reading Comprehension
Conventional Paper-and-Pencil Test (ACON)

Version 2: Inauthentic English Reading Comprehension
Conventional Paper-and-Pencil Test (ICON)

The computer-based tests are composed of two versions:

Version 3: Authentic English Reading Comprehension Computer-
Adaptive Test (ACOM)

Version 4: Inauthentic English Reading Comprehension Computer-
Adaptive Test (ICOM)

Set 2: The Interview Schedule: In order to investigate the test takers' attitudes, qualitative data from the interview was conducted by means of retrospective method. Thus, their attitudes towards the four versions of the tests were collected. This interview was conducted with 5 test takers randomly selected from each group.

Phase II: The implementation of the study

Phase III: Analysis of the data

3.3 Population and samples

3.3.1 Population

The population of this study was approximately 1,500 first year undergraduate students at King Mongkut's Institute of Technology North Bangkok in the second semester of the academic year 2007. The students studied in the Faculties of Engineering, Applied Science, Technical Education, Industrial Technology and Management and Agro-Industry. Typically, most are male, aged between 18 – 22.

3.3.2 Samples

This experimental research employed the 2*2 factorial design. The research design used in this study was the Randomized Block Design. This design was appropriate to this study since it was constructed to reduce variance in the data. The sample was, therefore, divided into relatively homogeneous subgroups or blocks (McLinden, D. J. & Trochim, 1998) The students in each subgroup were relatively

homogeneous with respect to their English I grades they studied in the first semester. The sample was blocked into 3 groups: B-A, C-C+, and D-D+.

In terms of optimum sample size, at a 95% confidence level and ± 5 precision, the resulting sample size obtained from Yamane formula was 286 students from a population of about 1,500 (Yamane, 1973). Then students were randomly selected and used as subjects. Then, they were divided into 4 groups of 75. In total, 300 students were used in this study. The following figure shows the 4 blocked groups.

| | | | |
|--------------|--------------|--------------|--------------|
| B-A n=25 | B-A n=25 | B-A n=25 | B-A n=25 |
| C-C+ n=25 | C-C+ n=25 | C-C+ n=25 | C-C+ n=25 |
| D-D+ n=25 | D-D+ n=25 | D-D+ n=25 | D-D+ n=25 |
| Group 1 | Group 2 | Group 3 | Group 4 |

Figure 3.1: Four blocked groups

Then students were randomly assigned to each block. After that, the students' midterm and final scores from the English I course were used to test whether there was any significant difference (at the .05 level) between the groups in English Proficiency. (The English I score was used since the study was implemented in the second semester.) Lastly, each of the four groups was randomly assigned to take one version of the reading test, as shown in the following table.

Table 3.1: Four Groups of the Subjects Assigned to Take Four Forms of Reading Tests

| | |
|---|--|
| <p>Group 1: Authentic English Reading Comprehension Conventional Paper-and-Pencil Test <i>(ACON)</i></p> | <p>Group 2: Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test <i>(ICON)</i></p> |
| <p>Group 3: Authentic English Reading Comprehension Computer-Adaptive Test <i>(ACOM)</i> (consisting of the items from ACON)</p> | <p>Group 4: Inauthentic English Reading Comprehension Computer- Adaptive Test <i>(ICOM)</i> (consisting of the items from ICON)</p> |

The procedure for assigning students to the four groups is as follows:

1. Students were randomly selected and then they were divided into 4 groups based on the scores obtained from the midterm and final tests of the English I course.

2. ANOVA was used to ensure that the midterm and final scores for the 4 groups from the English I course were not significantly different.

3. The four groups were randomly assigned to take different forms of the reading test.

- Group 1: Authentic English Reading Comprehension Conventional Paper-and-Pencil Test *(ACON)*

- Group 2: Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test *(ICON)*

- Group 3: Authentic English Reading Comprehension Computer- Adaptive Test *(ACOM)*

- Group 4: Inauthentic English Reading Comprehension Computer- Adaptive Test *(ICOM)*

4. A week after the test administration, the retrospective interview was conducted in order to collect information about student attitudes toward test authenticity and test delivery mediums. Five test takers were randomly selected from each group for the interview. Each student was recorded about 10–15 minutes. All the reports were

transcribed. The data were coded and recoded by the researcher and another colleague who was trained for the task.

3.4 Stages of research

3.4.1 Phase 1: The development and validation of the research instruments.

The two main sets of research instruments; reading comprehension tests and the retrospective method were developed and validated through the following procedures.

Set 1: Reading Comprehension Tests

In terms of the development of a reading comprehension test, three stages which included the design, operationalization and administration stages suggested by Bachman and Palmer (1996) were conducted. The following are the details of each stage:

Stage 1: Design

The product of the design stage is a design statement, which is a document that includes the following components:

1. a description of the purpose(s) of the test,
2. a description of the TLU domain and task type,
3. a description of the test takers for whom the test is intended,
4. a definition of the construct(s) to be measured,
5. a plan for evaluating the qualities of usefulness, and
6. an inventory of required and available resources and a plan for their allocation and management.

Stage 2: Operationalization

Operationalization involves developing test task specifications for the types of test tasks to be included in the test, and a blueprint that describes how test tasks will be organized to form actual tests. Operationalization also involves developing and writing the actual test tasks, writing instructions, and specifying the procedures for scoring the test.

Stage 3: Test administration

The test administration stage of the test development involves giving the test to a group of individuals, collecting information, and analyzing this information, for two purposes:

1. assessing the usefulness of the test, and
2. making inferences or decisions for which the test is intended.

Therefore, the reading comprehension tests in this study were developed through the processes mentioned above.

The paper-based tests

Version 1: Authentic English Reading Comprehension Conventional Paper-and-Pencil Test (ACON)

This test was created through the three stages as follows:

Stage 1: Design

(1) The purpose of the ACON was set. This test aimed to assess the reading proficiency of first year students at King Mongkut's Institute of Technology North Bangkok.

(2) The target language use (TLU) domain was defined by distributing the open-ended questionnaire (See Appendix 1) to 10 English instructors and 20 students at King Mongkut's Institute of Technology North Bangkok in the first semester of the academic year 2006. This questionnaire aimed to obtain information about English reading in real life. The open-ended questions asked about their reading purposes, reading settings, reading topics, and reading materials in real life. The findings obtained from this questionnaire are described below:

Reading purposes in real life: doing research or academic reports, finding jobs, applying for a job, operating machines or equipment, dealing with other people or companies, entertaining, broadening outlooks, updating news, preparing for tests, and getting particular information (ie. tourist attractions)

Reading settings: library, computer room, classroom, around campus, workplace or office, factory, home or residence, on the streets, movie theatre, tourist attractions, banks, hospitals, restaurants, book stores, public transportation (ie. bus, sky train), and airport

Topics: politics, economics, entertainment, sports, technology (invention i.e. computer, robots), health, jobs, education, scholarships, science (scientific theories), tourism, instructional manuals, history of people or places, and business

Reading materials: textbooks, novels, short stories, newspapers, journals, magazines, academic reports, research papers, technical reports, brochures, leaflets, posters, manuals, bulletin boards, billboards, labels, business correspondence, e-mail, memorandum, websites, the Internet and television

Moreover, according to Tanthanis's study (2002), the three topics with the most frequently chosen by male first year students as their most interesting reading topics

were sports, computers and technology, and science. The three most interesting topics for female students were language and communication, nature and environment, and health. Hence, the two topics not included in the initial survey questionnaire (language and communication, and nature and environment) were added to the topic list.

Then, a revised questionnaire with closed-type items (See Appendix 2) was constructed based on the findings obtained from the open-ended version. 15 reading topics were included in this questionnaire. The respondents were asked to identify the reading purposes, materials and settings provided for each topic. The purpose of this second questionnaire was, therefore to define the target language use (TLU) domain, “What are their reading purposes?”, “Where do they read in their real life?”, “What do they read ?” and “What materials do they read ?”

To obtain the needed information mentioned above, the first questionnaire with the open-ended questions was trialed on 5 groups of people: 10 English instructors, 10 instructors of other disciplines, 10 supervisors, 15 currently enrolled students and 10 graduates with employment. The data obtained and the theoretical model of language ability was used to establish the test construct.

The findings obtained from the revised questionnaire were as follows:

All 5 groups selected the following three reading topics most frequently: Jobs, Technology and Entertainment. Therefore, these three reading topics were selected as the reading topics for constructing reading proficiency tests in this study (See Appendix 3).

The three reading topics were further analyzed in terms of their reading purposes, reading materials and reading settings.

The reading topic “Entertainment” was associated with 4 reading purposes included in the TLU domain of this study: relaxing, reading for pleasure, updating news and getting particular information. Regarding reading materials, journals or magazines were included in the TLU domain. In terms of reading settings, home, library, and movie theatre were included in the TLU domain (See Appendix 4).

The reading topic “Job” was associated with 4 reading purposes included in the TLU domain of this study: finding jobs, applying for a job, updating news, and dealing with other people or companies. Regarding reading materials, newspapers and the Internet were included in the TLU domain. In terms of reading settings, home, computer room and workplace were included in the TLU domain (See Appendix 5).

The reading topic “Technology” was associated with 4 reading purposes included in the TLU domain of this study: broadening their outlook, updating news, doing research or academic reports and operating machines or systems. Concerning reading materials, journals, magazines and the Internet were included in the TLU domain. In terms of reading settings, library and workplace were included in TLU domain (See Appendix 6).

(3) The target group taking this test was first year students who enrolled in English I course in the first semester of academic year 2007. These students were from the Faculties of Engineering, Applied Science, Technical Education, Industrial Technology and Management and Agro-Industry at King Mongkut’s Institute of Technology North Bangkok. Most of them were male, aged between 18 – 22.

(4) The findings obtained from (2) were used to define the target language use (TLU) domain. Therefore, the TLU domains (See Appendix 7) were used to construct the ACON.

In terms of the theoretical model of language ability used to establish the test construct, the three levels of comprehension or sophistication of thinking suggested by Mohamad (1999) were used. These three levels are presented in hierarchy from the least to the most sophisticated level of reading.

Level one: Literal - what is actually stated. The following elements are measured:

- Facts and details
- Rote learning and memorization (vocabulary)
- Surface understanding only (word meanings in context)

Common questions used to elicit this type of thinking are who, what, when, and where questions.

Level two: Interpretive or inferential - what is implied or meant, rather than what is actually stated. Level two includes the following skills:

- Summarizing the main idea
- Tapping into prior knowledge / experience
- Attaching new learning to old information
- Making logical leaps and educated guesses (predicting outcomes)
- Drawing inferences (reading between the lines to determine what is

meant by what is stated / explaining the author’s purpose)

Tests in this category are subjective, and the types of questions asked are thought-provoking questions phrased with words like why, what if, and how.

Level three: Critical - taking what was said (literal) and then what was meant by what was said (interpretive) and then extending (apply) the concepts or ideas beyond the situation. The skills at this level are:

- Analyzing
- Synthesizing
- Applying

These three levels of comprehension are, therefore, used as the constructs of the reading tests.

(5) A plan for assessing the qualities of usefulness was included at the initial consideration of the appropriate balance among the six qualities of usefulness: reliability, construct validity, interactiveness, impact, practicality and authenticity (Bachman and Palmer, 1996). However, in terms of the quality of test authenticity, the authentic reading tests were expected to possess a very high degree of this quality while inauthentic reading tests were expected to have a very low degree. The checklist for evaluating usefulness developed by Bachman and Palmer (1996) was used in this study.

In terms of evaluating the test authenticity, after the ACON was created, three native instructors were asked to evaluate the authenticity of this test. In Bachman and Palmer's framework, the relationship between the TLU tasks and the test tasks is most important. Brown (2004) adds some other ways to evaluate the extent to which a test was authentic by asking the following questions:

1. Is the language in the test as natural as possible?
2. Are items as contextualized as possible rather than isolated?
3. Are situations interesting, enjoyable, and/ or humorous?
4. Is some thematic organization provided, such as through a story line or episode?
5. Do tasks represent, or closely approximate, real-world tasks?

Question 5 is the key question, according to the Bachman and Palmer's definition.

The evaluation of test authenticity form (see Appendix 8) was, therefore, established by using these suggestions.

(6) In terms of an inventory of required and available resources and a plan for the allocation and management, the following resources for test development were planned:

Test developer and test writer were the researcher's roles. Therefore, the researcher developed the test from the very beginning to the end, from the table of specifications to administration, try-out, and use, and to archiving.

The equipment and test materials used for the ACON was answer sheets for the item types in short answer, gap-filling, and information transfer formats.

In Stage 2: Operationalization

(1) The test specification was established by combining three levels of comprehension and TLU domains for the ACON.

(See the test specification of the Authentic English Reading Comprehension Test and the Inauthentic English Reading Comprehension Test in Appendix 9.)

(2) The reading texts included in the test were selected according to the following steps:

- The reading texts were selected according to the reading topics mentioned in the test specifications (jobs, technology and entertainment). Alderson (2000) suggested that texts chosen in reading tests are usually between 150–350 words in length and the length of the selected texts in this study also fall within this range.

- Each 150–350 word-length text was evaluated to determine the readability of each text. The Fry Readability Program (www.educational-psychologist.co.uk/fry_readability_program.htm) was used. The program can provide a rough guide and a useful indication as to whether the content of the text is at the right level for the intended students or the test-takers. The value obtained from the program is a rough measure of the reading age of the readers and the difficulty of the text. Since the population of this study was aged between 18 and 22 years, the readability level used was 18 years or above.

However, readability index provides only a rough guide because it cannot take account of the conceptual density of the material (O' Donnell and Wood, 2004), the structural and rhetorical features of the text (Clapham, 1996), and a reader's knowledge and interest (Clapham, 1996, Nuttall, 1996). Moreover, the rough guide obtained is generally used for native readers of English language. The rough measure may not be appropriate for nonnative students. However, Clapham (1996) suggests that in addition to

evaluating readability of a text, reading experts, teachers and testers (who are familiar with the text and with the backgrounds of the students) or test takers who are expected to read the texts should take part in selecting appropriate texts.

- Then the three selected reading texts which were the same length (150-350 words) and at the same readability level (18 years or above) for each topic (Entertainment, Jobs and Technology) were considered according to the test task characteristics mentioned in the test specification. The following criteria for selecting appropriate reading passages suggested by Day (1994) were also considered.

1. **Lexical Knowledge:** The number of unknown words is acceptable in a reading passage.

2. **Background Knowledge:** The passage is on a topic that is known or familiar to the students.

3. **Syntactic Appropriateness:** A passage contains grammatical constructions that the students tend to know. These constructions are not new or not too difficult for them to recognize or understand.

4. **Organization:** A passage is well organized.

5. **Discourse Phenomena:** This includes the arrangement of topics and comments in a reading passage, and considerations of cohesiveness and coherence. The students who read the passage tend to be able to handle the presentation of ideas and arguments in the passage. The cohesion markers and transition devices are within the linguistic competence of the students, and they can follow the line of reasoning utilized by the writer of the passage.

Therefore, both the test task characteristics and the criteria suggested by Day (1994) were used to establish the passage selection form (See Appendix 10).

- Three English instructors were asked to select the passage for each test task by using the passage selecting form. The passages with the highest scores were included in the ACON.

(3) The researcher wrote the test items according to the test specification and prepared both answer key and scoring scheme.

(4) After the draft ACON was finished, the test was edited by the researcher's adviser and other experts.

In Stage 3: Test administration

ACON was piloted four times in the academic years 2005 and 2006. Finally, all the items with particular b-parameter or item difficulty level required to be put in the fixed position of the computer-adaptive flowchart were obtained. The following are the procedures for developing ACON in each pilot.

1. The first pilot: ACON was tried out with 255 first-year students at King Mongkut's Institute of Technology North Bangkok in the first semester of academic year 2005. The test was analyzed in order to find out the test reliability and the quality of the items.

For the item analysis, the three-parameter IRT model was applied. The criteria used for selecting the appropriate items for proficiency tests suggested by Sukamolson (1999: 160) are: the a-parameter (discrimination) should fall in the range of 0.5 – 2.5, the b-parameter (difficulty) should be in the range of -2.5 - + 2.5 and the c-parameter (guessing) should fall in the range of 0 – 0.40. However, with regard to the acceptable range of the b-parameter, Hambleton and Swaminathan (1985) suggest the range of -3.0 - +3.0 in the fixed-branching or pyramidal model of computer-adaptive tests (CAT). This study, which aims at creating a CAT by using this pyramidal model, therefore, uses this range of the b-parameter.

For the first pilot, the test was analyzed by using XCALIBRE for Windows version 1.10 (<http://www.assess.com/Software/xcalibre.htm>). The reliability of the test or KR-21 (XCALIBRE reports KR-21 as test reliability) obtained was 0.887. The range of a, b and c parameters were 0.52 – 1.23, -0.64 – 2.41, 0.11 – 0.15 respectively. The program deleted fifteen items during the calculating process because they were mostly omitted by the students.

Since the fixed-branching or pyramidal CAT involved the placement of the test items in a branching tree in advance, depending on the response to each item, a more difficult item will be presented next for a correct response and an easier item will be provided for an incorrect response. Not only is the test adaptive, but the sequence of items is also content-based. The following figure illustrates the ideal flowchart of the CAT conducted in this study.

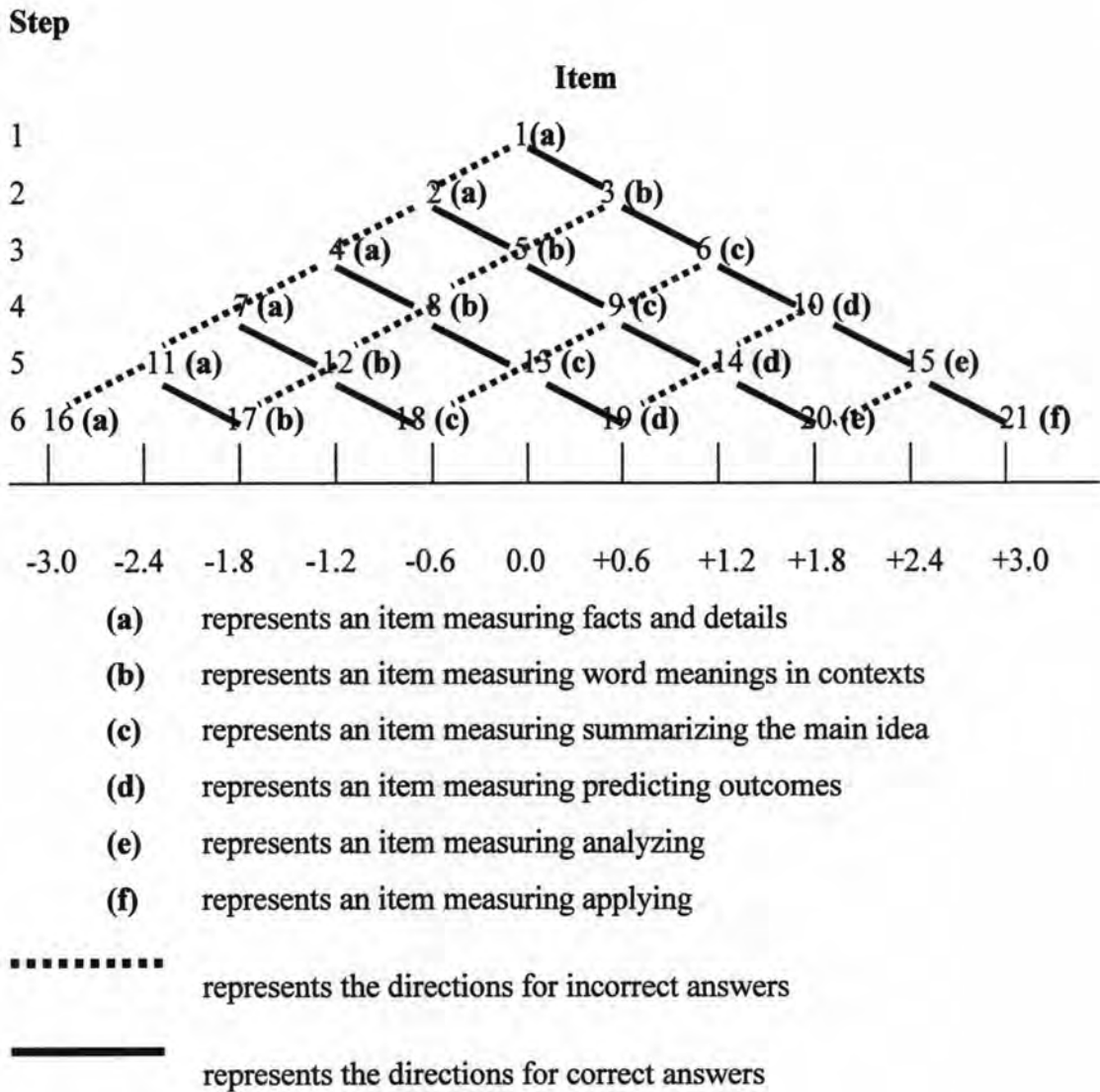


Figure 3.2: The flowchart of a content-based CAT

The difficulty levels of items obtained from this pilot were not exactly the same as the difficulty levels expected. However, the concept of the pyramidal CAT is to have more difficult items on the rightward direction and easier items on the leftward direction. The items were put in the flowchart by considering their difficulty levels and the contents. It was found that some items could be put in the flowchart while the other items could not (See Appendix 11).

The items that could be put in the particular positions were kept, while the others were adjusted or changed. After the ACON was developed for a second time, it was validated by the same three experts and the degree of test authenticity was evaluated again.

2. The second pilot: 145 first-year students at King Mongkut's Institute of Technology North Bangkok in the second semester of academic year 2005 were asked to

take the ACON. The reliability of the test or KR-21 obtained was 0.845. The range of the a, b and c parameters were 0.63 – 1.16, -1.65 – 3.00, 0.15 – 0.23 respectively. Three items were deleted due to being mostly omitted.

The items then were put in the CAT flowchart. More items from this pilot could be put in the flowchart (See Appendix 12).

The items which could not be placed in the flowchart were adjusted and changed again. Then the whole test was validated and evaluated according to the degree of test authenticity before having the third pilot.

3 The third pilot: ACON was piloted with 226 first-year students at King Mongkut's Institute of Technology North Bangkok in the first semester of academic year 2006. The reliability of the test or KR-21 was 0.813. The range of the a, b and c parameters were 0.87 – 1.88, -1.01 – 3.00, 0.15 – 0.27 respectively.

The majority of the items could be placed in the flowchart. Nevertheless, there were still some items that could not be placed. The test was adjusted again by the same procedures (See Appendix 13).

Since ACON has to possess a very high degree of test authenticity, the experts were asked to check if the language used in the test was natural in every pilot. The major changes after this third pilot, thus, were made to the language structures and vocabulary in order to have the language in the test as natural as possible.

The major changes made could be illustrated as follows:

Task 1: Reading entertainment news

Item 2

Before changing: What song was likely to have made Morris made his decision to work with Sek?

After changing: Alfon, what song was likely to have made Morris reach his decision to work with Sek?

Item 3

Before changing: Could you give me some examples of the bands that could have been very popular because of Carr?

After changing: Carr seems to be behind several famous bands, do you know some of them?



Task 2: Finding a job**Item 23**

Before changing: Which company tends to be the oldest company among these ads?

After changing: See Tom, which company is possibly the oldest one according to these ads?

Item 35

Before changing: This ad for Premas is a bit too wordy for my liking. What exactly do they do?

After changing: The Premas ad is a bit too wordy for my liking. What exactly do they do?

Item 41

Before changing: Look at this ad from Premas Group. Does it say how many businesses the company covers?

After changing: Does the Premas ad say how many types of business the company covers?

Task 3: Reading an article for preparing a report**Item 47**

Before changing: How far were travelers in the past likely to make an error on their long journeys?

After changing: Janos, how much of an error were travelers in the past likely to make on their long journeys?

Item 48

Before changing: What kind of vehicle did travels in ancient time use to travel by sea?

After changing: Sita, what kind of transport did travelers in ancient times use to travel by sea?

Item 55

Before changing: What was the main device Columbus used to tell the direction?

After changing: Janos, what was the main device Columbus used to find the ship's direction?

Then, three native instructors were asked to check the authenticity of the test by using a test authenticity form.

The test authenticity form is a questionnaire with four-point response scale. The responses ranged from strongly disagree (1) to strongly agree (4). In order to interpret the results, the range of criteria was set as follows:

| | |
|-----------|-------------------|
| 3.50-4.00 | strongly agree |
| 2.50-3.49 | agree |
| 1.50-2.49 | disagree |
| 1.00-1.49 | strongly disagree |

The results obtained from three native instructors are illustrated in the table.

Table 3.2: The Degree of Test Authenticity in ACON

| Authenticity of test tasks | \bar{X} | S.D. |
|--|-----------|------|
| The language in each test task is natural. | 3.67 | 0.58 |
| The items are contextualized. | 3.67 | 0.58 |
| The situations in the test are interesting. | 3.33 | 0.58 |
| The theme in the test tasks can occur in the real situation. | 3.67 | 0.58 |
| Tasks are related to real-world tasks. | 3.33 | 0.58 |
| Total | 3.53 | 0.52 |

According to the results above, it can be concluded that the native English instructors strongly agree that the language in each test task is natural, the items are contextualized and the theme in the test can occur in the real situation. Generally, they agree that the situations in the test are interesting and the tasks are related to real-world tasks. Moreover, the high average score can be interpreted to mean that they strongly agree that the test is authentic.

4. The fourth pilot: 675 first-year students at King Mongkut's Institute of Technology North Bangkok took the ACON in the second semester of academic year 2006. The reliability of the test or KR-21 was 0.875. The range of the a, b and c parameters were 1.50 – 2.50, -1.34 – 3.00, 0.18 – 0.30 respectively.

Since this test was analyzed according to Item Response Theory by using the 3-parameter model, a factor analysis was conducted in order to test the unidimensionality assumption. The results showed that there were 6 separate factors underlying the items measured in the test. These factors are:

1. facts and details
2. word meanings in contexts
3. summarizing the main idea
4. predicting outcomes
5. analyzing
6. applying

Generally speaking, it can be interpreted that each item tests one dimension (See Appendix 15).

All the items could be placed on the flowchart according to their difficulty levels and their measured contents. The test items are reordered and numbered according to their difficulty levels, from the easiest to the most difficult items (See Appendix 14).

Since all the items in ACON are able to be placed in the CAT flowchart according to their difficulty levels and the contents measured, ACON is ready to be transformed into the Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test (ICON) and the Authentic English Reading Comprehension Computer-Adaptive Test (ACOM). (See a complete version of ACON in Appendix 16).

Version 2: Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test (ICON)

1. Since this test is inauthentic, the definition of the target language use (TLU) domain was not considered. The item type used in this test is the multiple-choice format. All the items in ACON were transformed from short answer to multiple-choice.

2. Since piloting ICON might affect the level of difficulty of each item, a native speaker expert helped to judge the validity of the test by checking whether each pair of the test items from ACON and ICON measure the same thing (See ICON in Appendix 17).

3. After ICON had been already created, the authenticity of the test was evaluated again. The same three English instructors were asked to use the same evaluation of test authenticity form and criteria for interpreting the results.

The below range was set to interpret the results.

| | |
|-----------|----------------|
| 3.50-4.00 | strongly agree |
| 2.50-3.49 | agree |
| 1.50-2.49 | disagree |

1.00-1.49 strongly disagree

The findings are presented in the following table.

Table 3.3: The Degree of Test Authenticity in ICON

| Authenticity of test tasks | \bar{X} | S.D. |
|--|-----------|------|
| The language in each test task is natural. | 2.33 | 0.58 |
| The items are contextualized. | 1.00 | 0 |
| The situations in the test are interesting. | 1.00 | 0 |
| The theme in the test tasks can occur in the real situation. | 1.33 | 0.58 |
| Tasks are related to real-world tasks. | 1.00 | 0 |
| Total | 1.33 | 0.62 |

It is evident that the native English instructors do not agree that the items are contextualized; the situations in the test are interesting; the theme in the test tasks can occur in the real situation; and the tasks are related to real-world tasks. Moreover, they disagree that the language in each test task is natural though they rated this one more highly than other issues. Overall, the instructors strongly disagree that this test is authentic.

According to low ratings for authenticity, this test can be considered to be inauthentic.

The computer-based tests

Both ACON and ICON were transformed to ACOM and ICOM respectively and uploaded onto the website, www.ecatonline.com/ecat.

The program was mainly developed with HTML as a tool in the developing process, PHP as the programming language, My as the database and Macromedia Dreamweaver MX as an editor.

Generally speaking, ACOM and ICOM possess the same and different perspectives. The perspectives of both the tests share are as follows:

1. The program system

1.1 The system was divided into two main sections: test administrator and test taker sections.

Section A or the test administrator section allows the teacher to add, delete, and get the test takers' test scores. (See Section A in Figure 3.3.)

Section B or test taker section provides the test for each test taker. (See Section B in Figure 3.3.)

1.2 The data needed to be stored in the system are the item bank of the Authentic English Reading Comprehension Test and test takers' data (names and scores obtained from the test).

The implementation of the system is presented in the following figure.

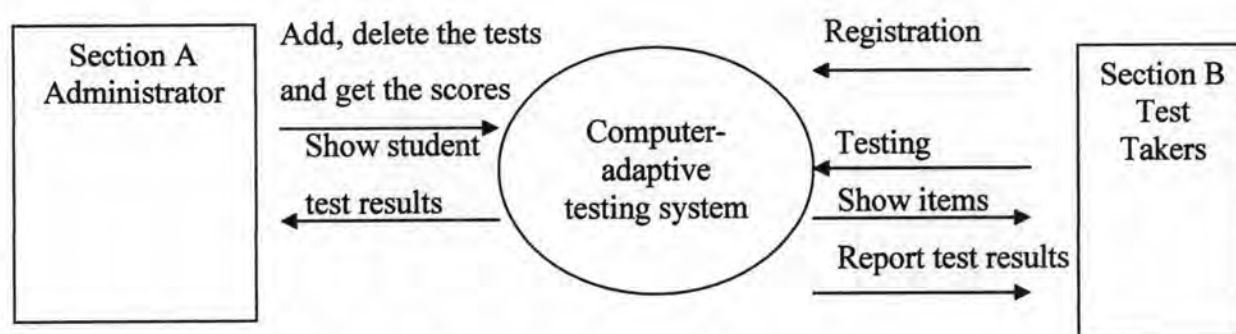


Figure 3.3: The system of the computer-adaptive test

According to figure 3.3, once the test administrator prepares the test into the computer-adaptive testing system, each test taker can register for the test by using the username and password assigned. Then the test taker can access the system. The system will provide him/her the test items and after finishing the whole test, the result will be reported. Finally, the test administrator can investigate each test taker's test result by accessing section A.

2. The program management

2.1 Ways to access the online test

The program was prepared for the three main groups of the users: the test administrator, the test takers who took the Authentic English Reading Comprehension Computer-Adaptive Test (ACOM) and the test takers who took the Inauthentic English Reading Comprehension Computer- Adaptive Test (ICOM). Therefore, the usernames and passwords were provided for the three different groups. A code was assigned to each test taker for use as username and password. The test takers who took ACOM had different usernames and passwords from those who took ICOM. For the test administrator, another username and password were assigned to access the history of all test takers, including each test taker's test result.

Accordingly, the system was set to respond to the three different groups of program users as shown in Figure 3.4.

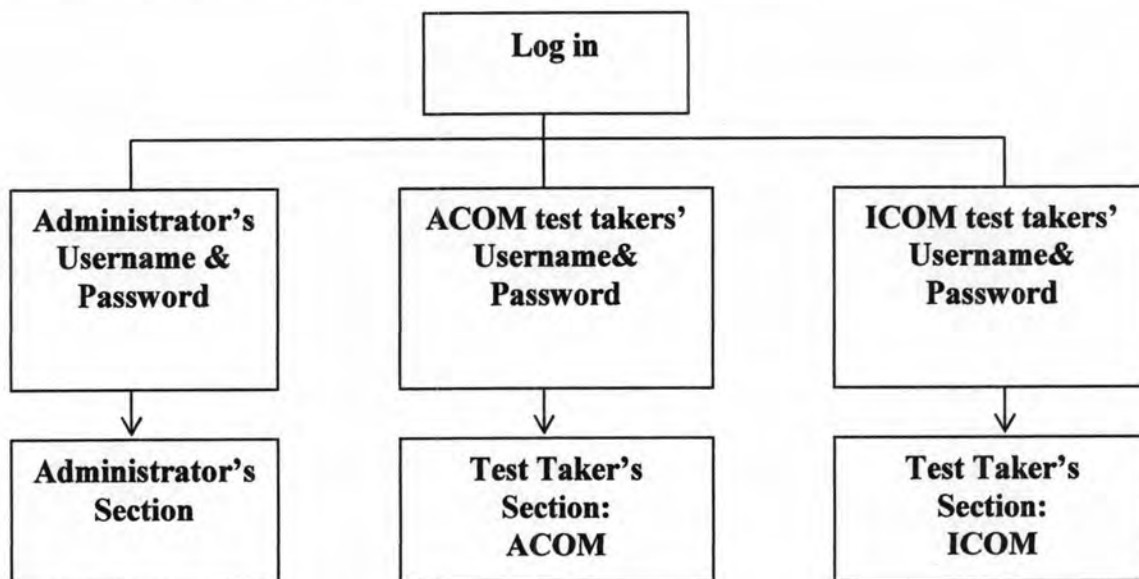


Figure 3.4: The ways to get access to the system

2.2 Test administrator section

The test results obtained from both ACOM and ICOM can be monitored by using the particular username and password provided for the test administrator.

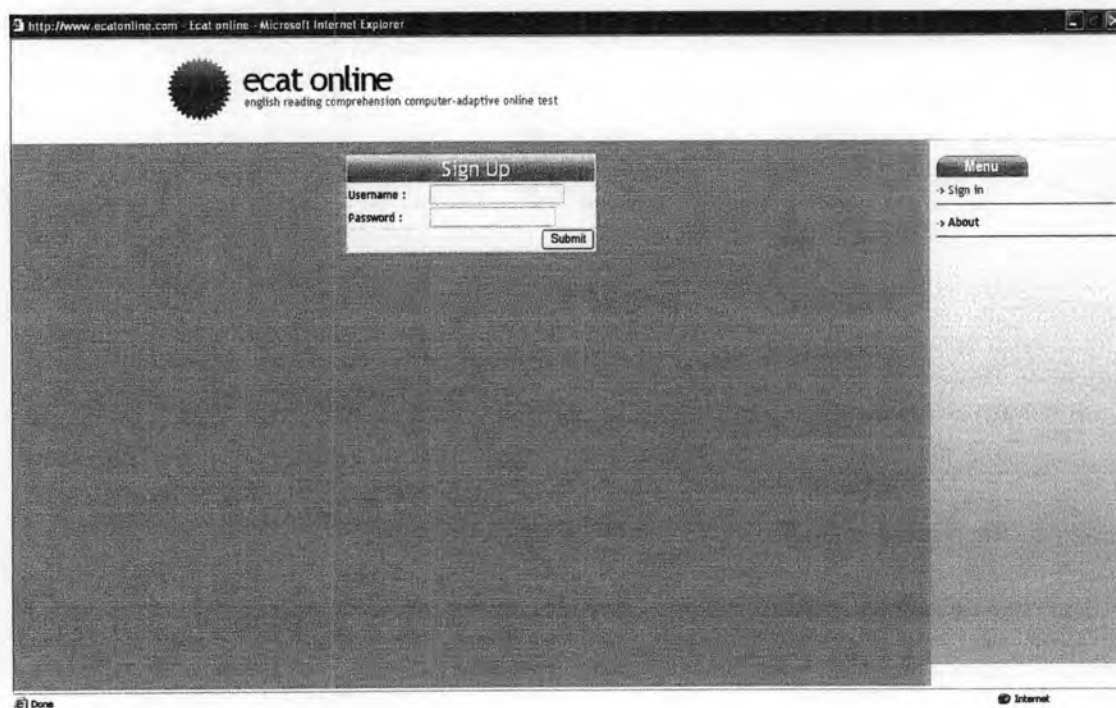
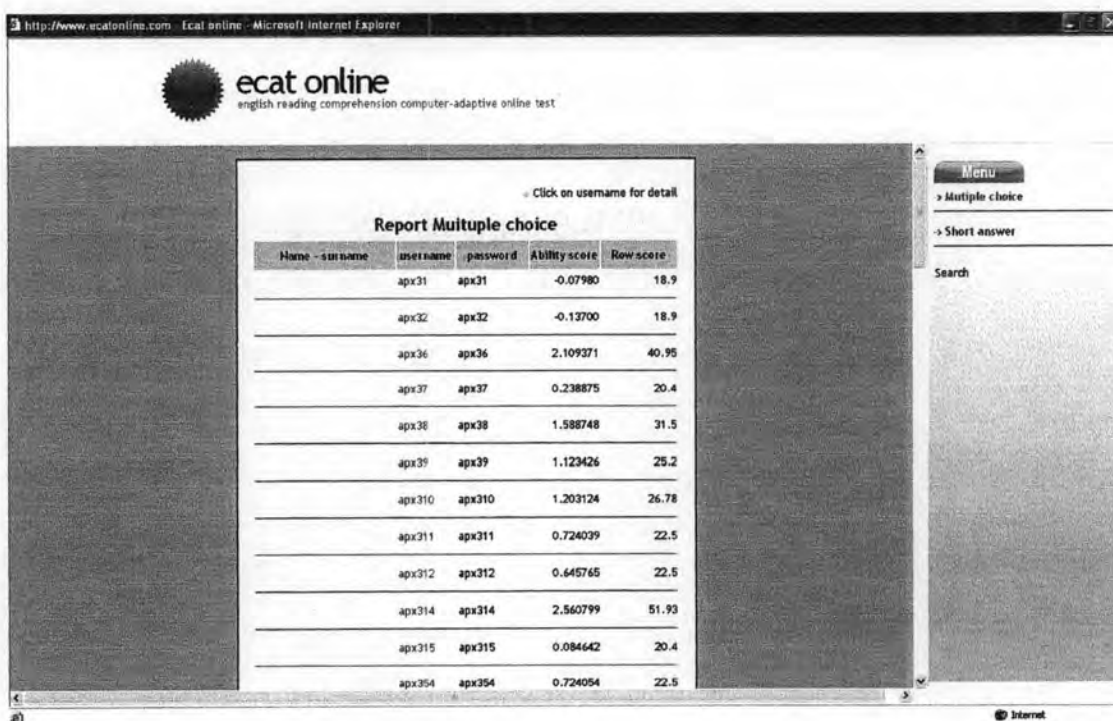


Figure 3.5: The signing in page

Figure 3.5 shows the sign in page, which is the same page for the three groups; test administrator, ACOM test takers and ICOM test takers type their assigned usernames and passwords.

After submitting the username and password, the test administrator can select the history of both ACOM and ICOM to monitor their attendance and their scores as shown in Figures 3.6 and 3.7. Figure 3.6 presents the scores obtained from the test takers who took the multiple choice test (ICOM) while Figure 3.7 illustrates the scores obtained from those who took the short answer test (ACOM).



Click on username for detail

Report Multiple choice

| Name - surname | username | password | Ability score | Raw score |
|----------------|----------|----------|---------------|-----------|
| | apx31 | apx31 | -0.07980 | 18.9 |
| | apx32 | apx32 | -0.13700 | 18.9 |
| | apx36 | apx36 | 2.109371 | 40.95 |
| | apx37 | apx37 | 0.238875 | 20.4 |
| | apx38 | apx38 | 1.588748 | 31.5 |
| | apx39 | apx39 | 1.123426 | 25.2 |
| | apx310 | apx310 | 1.203124 | 26.78 |
| | apx311 | apx311 | 0.724039 | 22.5 |
| | apx312 | apx312 | 0.645765 | 22.5 |
| | apx314 | apx314 | 2.560799 | 51.93 |
| | apx315 | apx315 | 0.084642 | 20.4 |
| | apx354 | apx354 | 0.724054 | 22.5 |

Menu

- > Multiple choice
- > Short answer

Search

Figure 3.6: ICOM test takers' scores

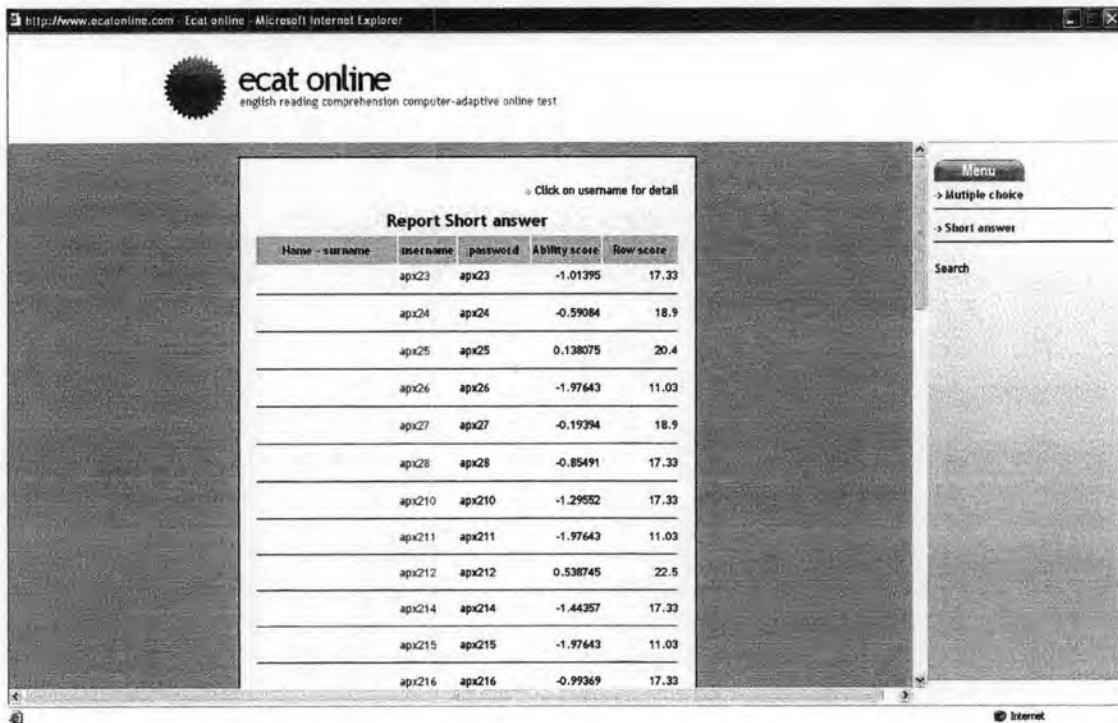


Figure 3.7: ACOM test takers' scores

However, the test taker's section allows the test takers with different usernames and passwords assigned to access different kinds of tests: ACOM and ACON. The perspectives of each test are described as follows:

Version 3: Authentic English Reading Comprehension Computer- Adaptive Test (ACOM)

ACOM was developed based on the test items from ACON. However, they were different due to the fact that ACOM is a computer-adaptive test. The test takers taking the test items possess different levels of difficulty appropriate to their reading ability. The following are important perspectives of the test.

1. The computer-adaptive test design

The computer-adaptive type used in this study is the 6-step fixed length. Test items from ACON were fixed in certain positions in the pyramidal flowcharts according to their b-parameter or difficulty levels (see Appendix 14).

2. The score calculation

Maximum likelihood formula is used to calculate the test taker's ability score (θ). Then the ability score was transformed to raw score by using the data obtained from Test Characteristic Curve (TCC) (see Appendix 19).

3. The procedures for taking the test

The test was implemented according to the following procedures:

3.1 The students access the test website: <http://ecatonline.com/ecat>. Figure 3.8 shows the ACOM homepage.

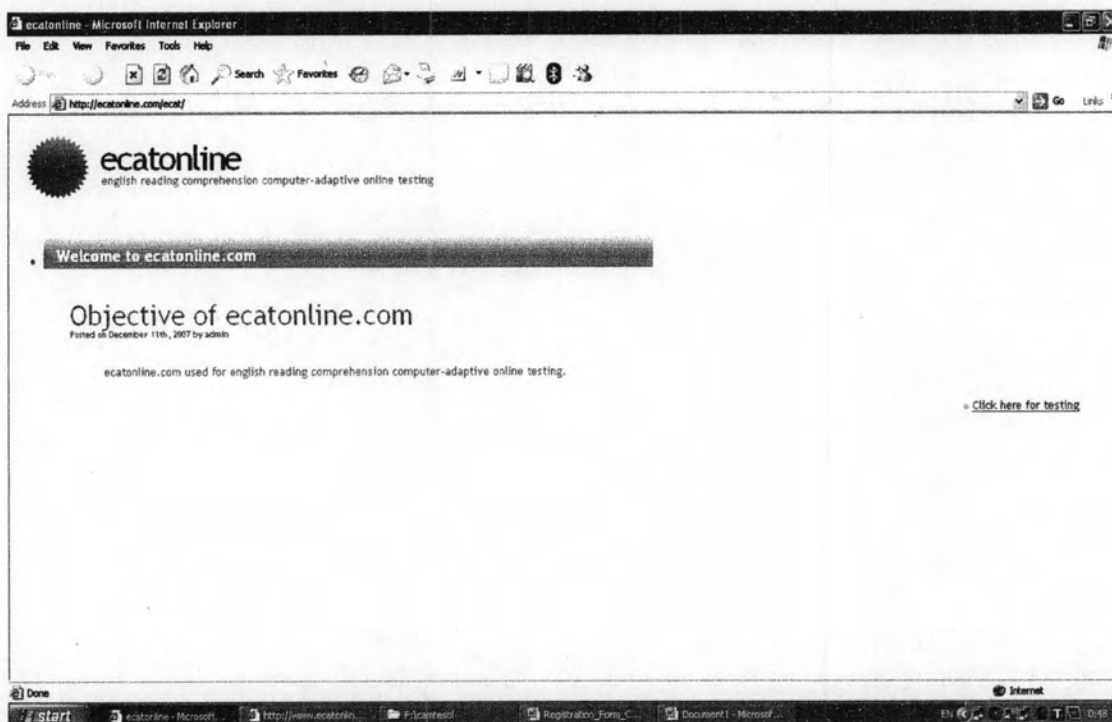


Figure 3.8: ACOM homepage

3.2 The introduction to the program will pop up, then click the “sign in” button which is under the Menu. Figure 3.9 illustrates the introduction of the program.

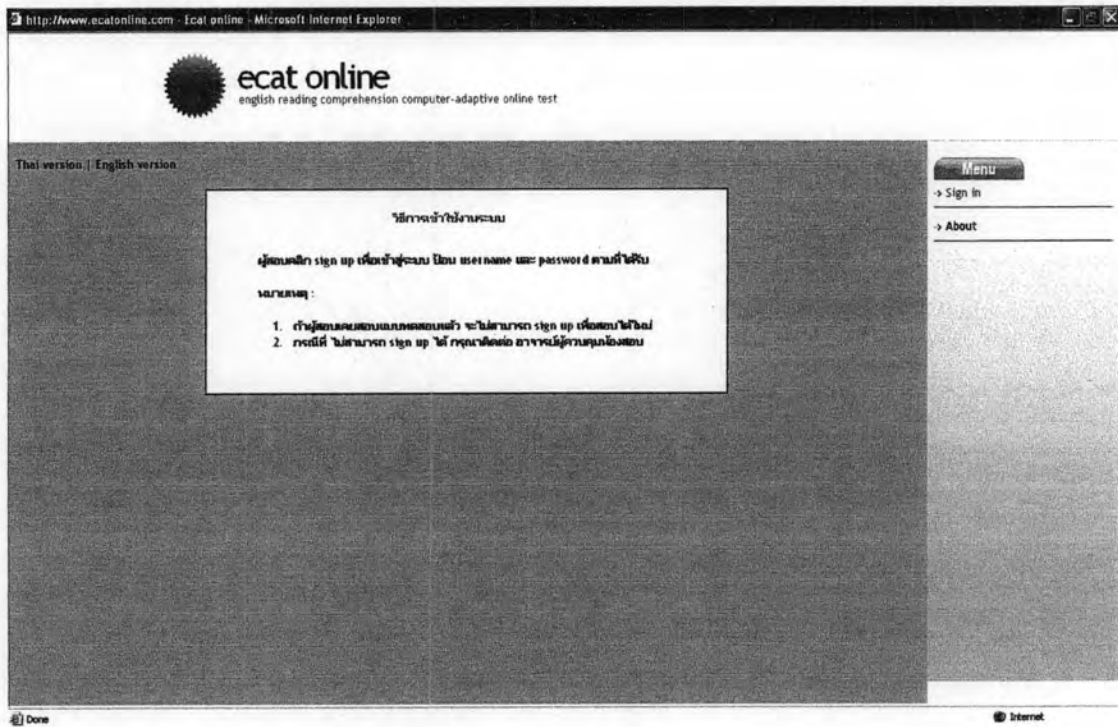


Figure 3.9: The introduction of the program

3.3 Type the username and password assigned by the test administrator and then click “submit” as shown in Figure 3.10

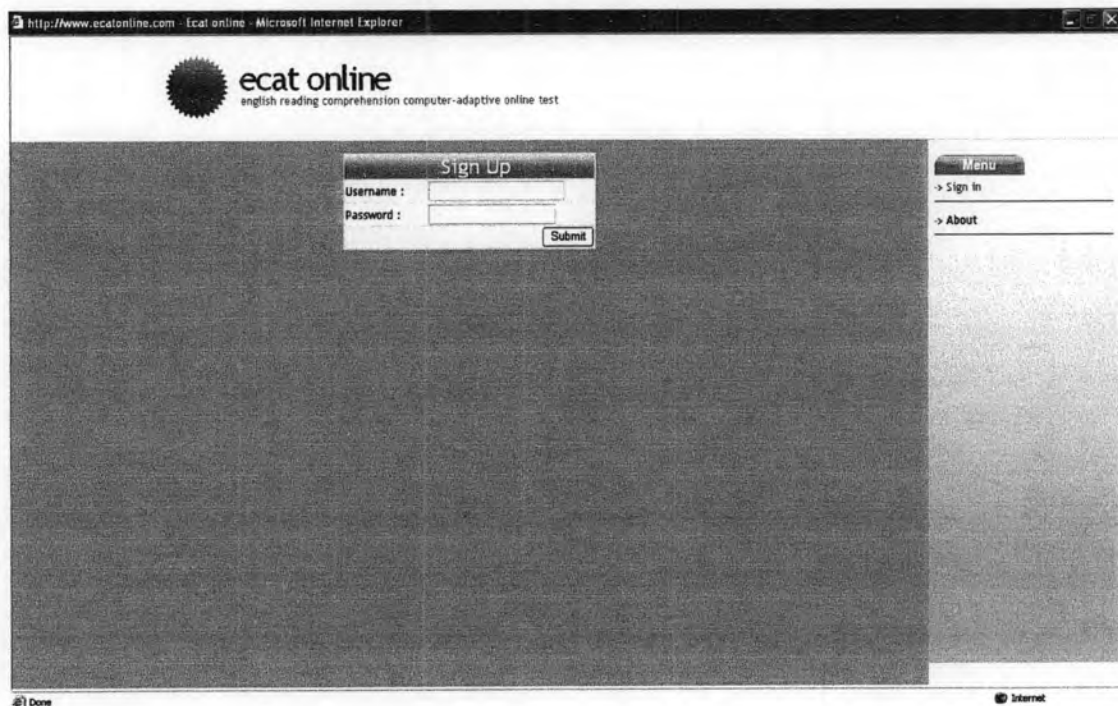


Figure 3.10: Sign in page

3.4 Before taking the test, the test instructions will be presented. Test takers can click “An example of the test” to see the sample of test items. The test can be started by selecting the reading topic. Since the test consists of three reading topics: Entertainment news; Finding a job and Technology. The students can start by selecting a topic in any order.

Figure 3.11 shows the page presenting the test instructions. All test takers will be informed before starting the test.

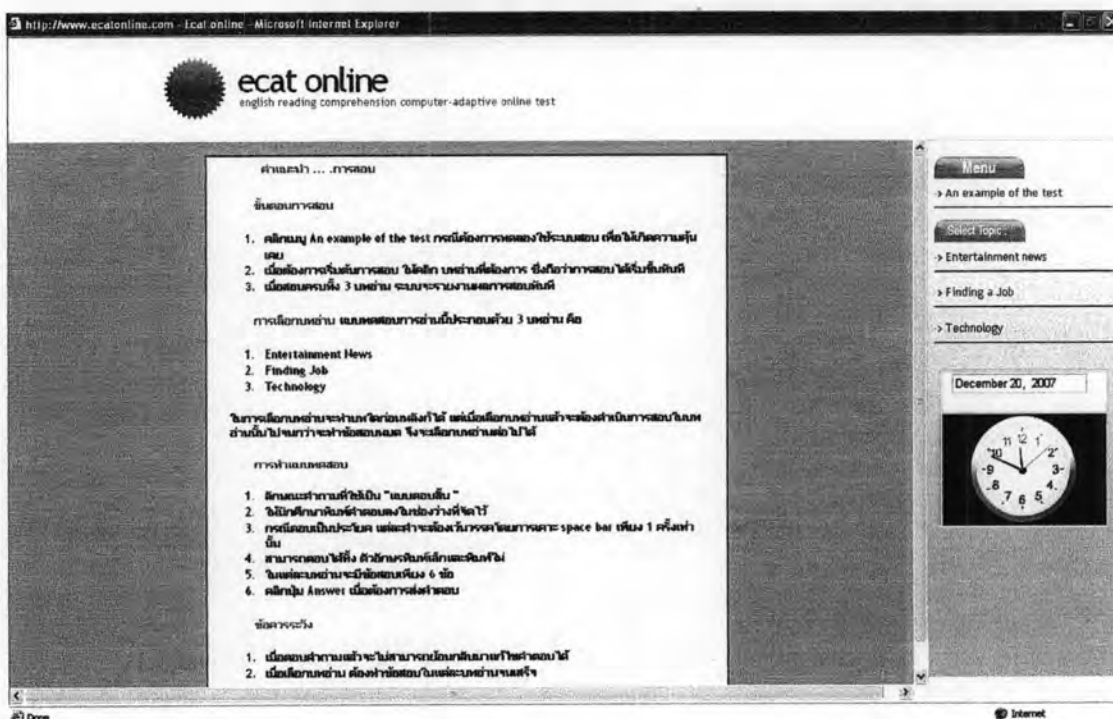


Figure 3.11: The page presenting the test instructions (ACOM)

3.5 After clicking “An sample of test item”, the screen will show an example of the test which allows test takers to familiarize themselves with the test they are going to take. The following figure illustrates an example of a test item in the ACOM test.

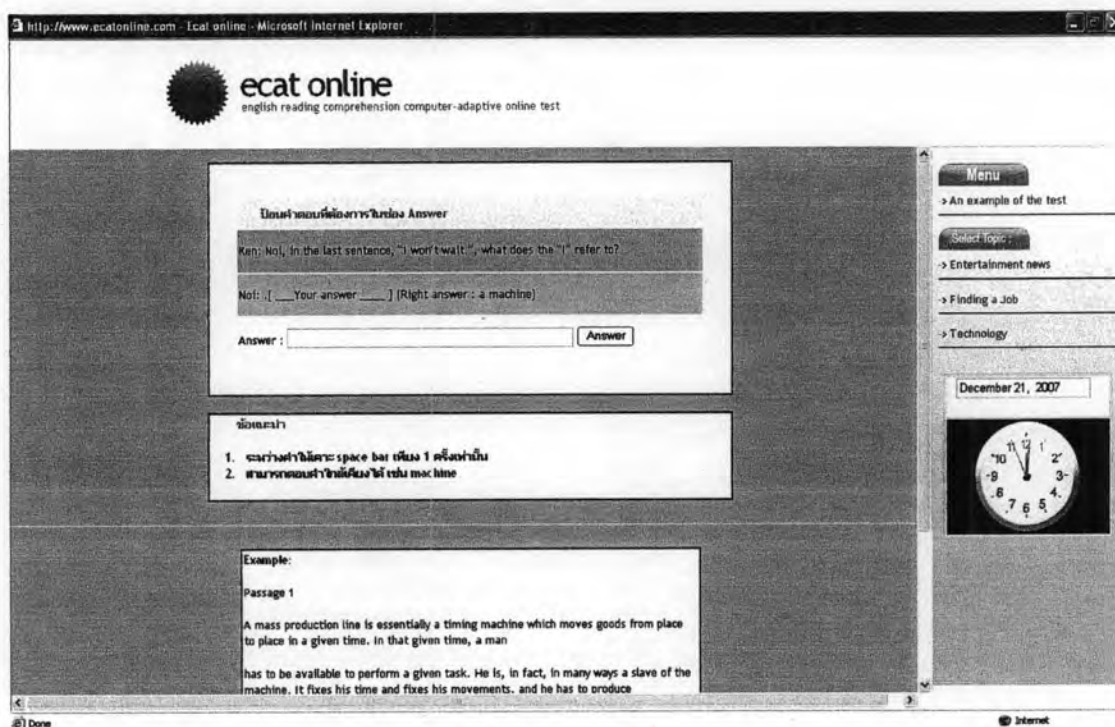


Figure 3.12: The page showing an example of the test

3.6 After selecting a reading topic, the test will start. Figure 3.13 illustrates the first item in the ACOM.

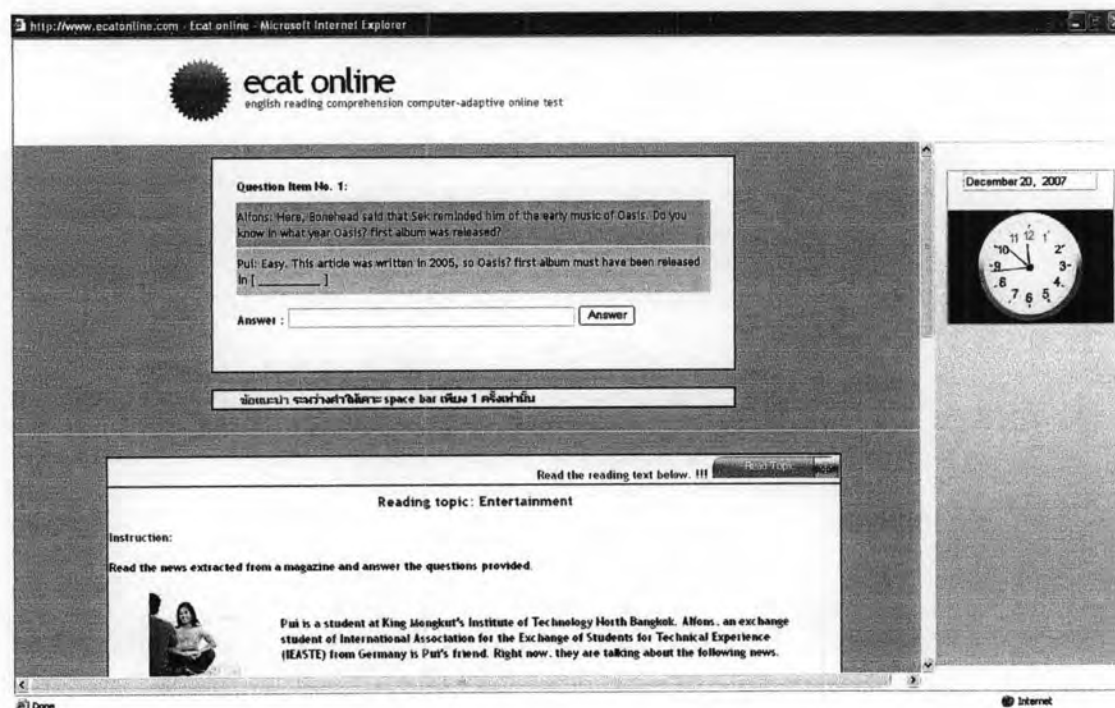


Figure 3.13: Authentic English reading test item

3.7 The test taker's test result will be immediately reported. The ability score will be automatically transformed to the raw score for each student. Since this test is domain-referenced, the reading ability levels the test takers can and cannot pass will also be illustrated. Figure 3.14 illustrates the test taker's test result.

The screenshot displays the ecat online interface. At the top, the logo and name 'ecat online' are visible, along with the subtitle 'english reading comprehension computer-adaptive online test'. The main content area is titled 'Summary Test' and includes a form for 'Name Surname' and a 'Raw Score' of 18.5/63. Below this, there are two reports. The first report is for 'Entertainment News Topic' and contains a table with 6 items. The second report is for 'Finding Job Topic' and contains a table with 2 items. Each item in the table has a 'Flowchart' (represented by circles with letters), an 'Answer' (marked with 'X' or a checkmark), and a 'Comment' (e.g., 'Literal: facts and details' or 'Inferential: summarizing the main idea').

| Item No. | Flowchart | Answer | Comment |
|----------|-------------------|--------|--|
| 1 | A1 | X | Literal: facts and details |
| 2 | A2 B1 | ✓ | Literal: facts and details |
| 3 | A3 B2 C1 | ✓ | Literal: word meaning in contexts |
| 4 | A4 B3 C2 D1 | X | Inferential: summarizing the main idea |
| 5 | A5 B4 C3 D2 E1 | X | Inferential: summarizing the main idea |
| 6 | A6 B5 C4 D3 E2 F1 | X | Inferential: summarizing the main idea |

| Item No. | Flowchart | Answer | Comment |
|----------|-----------|--------|----------------------------|
| 1 | A1 | X | Literal: facts and details |
| 2 | A2 B1 | X | Literal: facts and details |

Figure 3.14: The page presenting the test taker's test result

Version 4: Inauthentic English Reading Comprehension Computer- Adaptive Test (ICOM)

After getting access to the website with assigned usernames and passwords, the test takers will follow the following steps.

4.1 Before taking the test, the test instructions are presented. Test takers can click "An example of the test" to see the test example. The test can be started by selecting the reading topic. The students can select any part of the test.

The following figure shows the test instruction for ICOM.

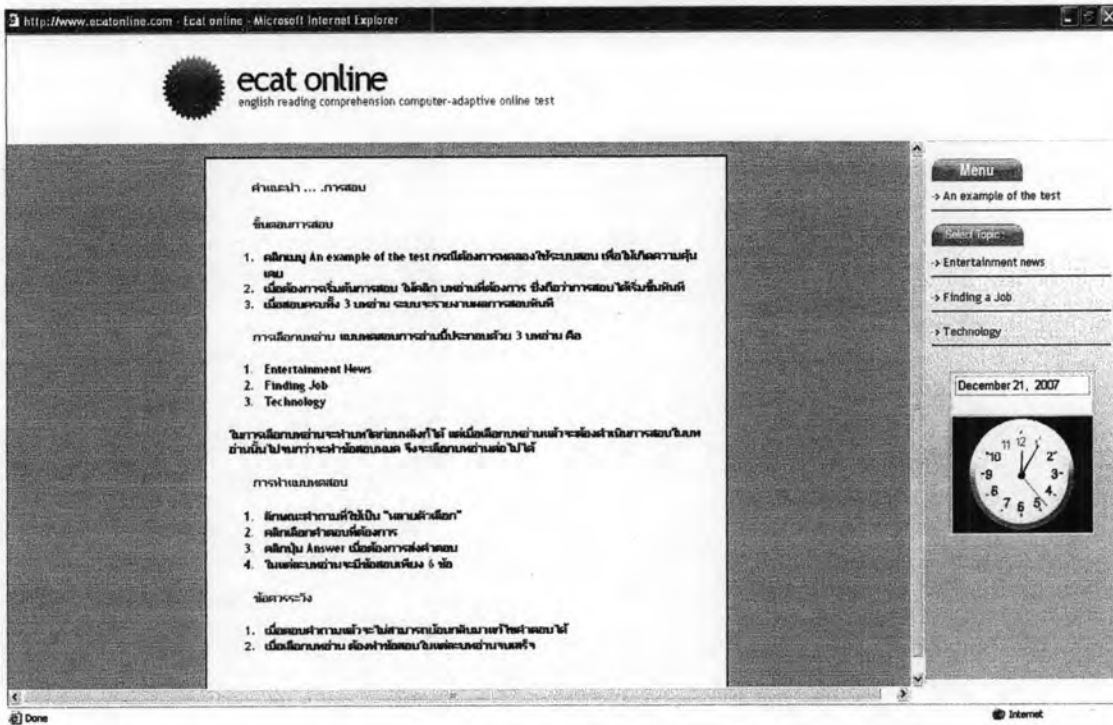


Figure 3.15: The page presenting the test instructions (ICOM)

4.2 After selecting a reading topic, the test will start as shown below.

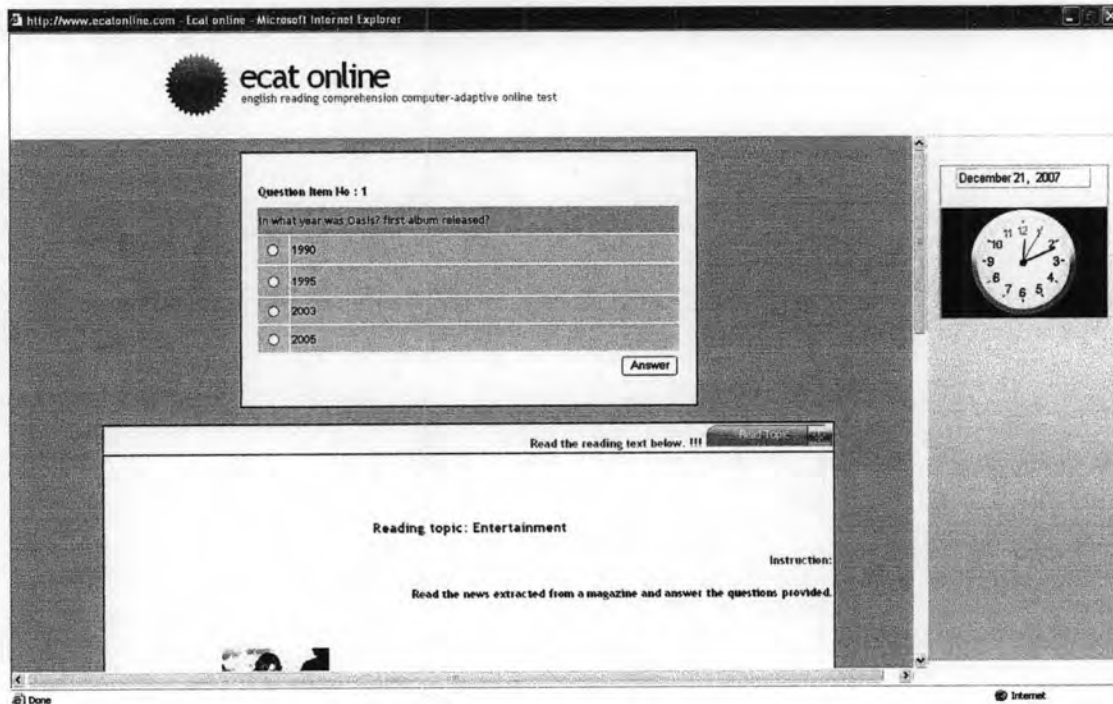


Figure 3.16: Inauthentic English reading test item

4.3 The test taker's test result will be immediately reported. The ability score will be automatically transformed to raw score for each student. Since this test is

domain-referenced, the reading ability levels the test takers can and cannot pass will also be illustrated. Figure 3.17 shows the test taker's test result in ICOM.

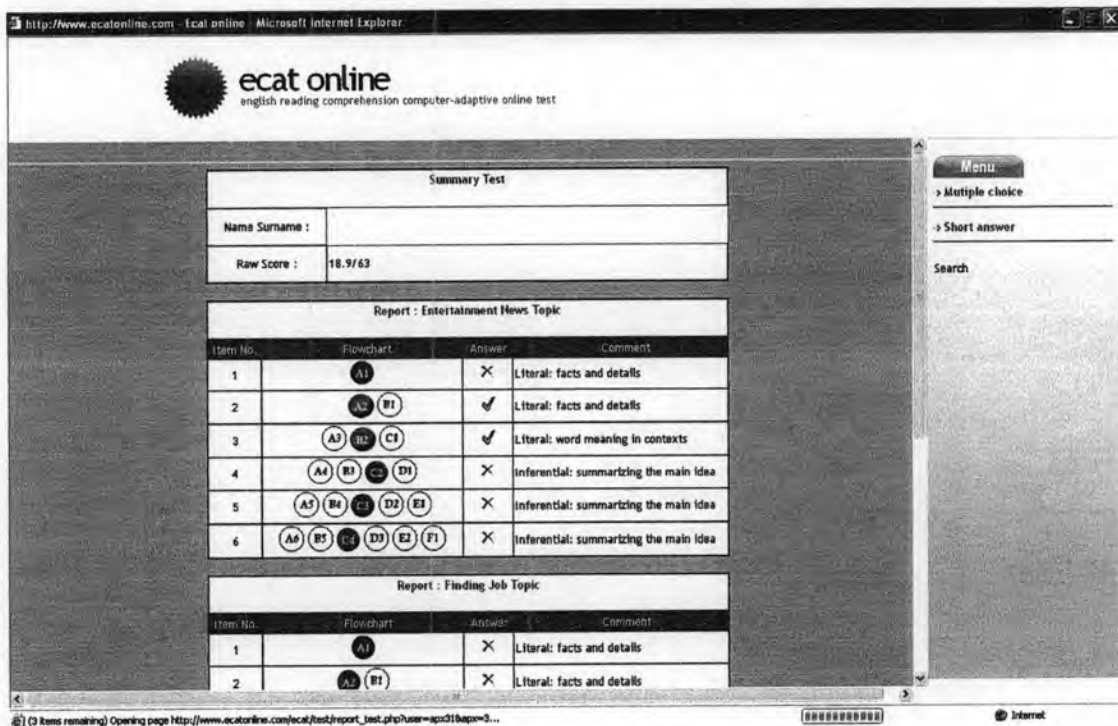


Figure 3.17: The page presenting the test taker's test result

Set 2: The Retrospective Interview

A retrospective study was conducted with 5 test takers randomly selected from each group (since there are four groups, 20 test takers participated in each group). Then they participated in retrospective interviews for an in-depth investigation on their attitudes toward the test authenticity and test delivery mediums or how they view ACON, ACOM, ICON and ICOM.

The retrospection by means of a semi-structured interview was conducted one week after the students took the tests. This interview was planned in advance according to the following steps.

1. Since the objective of this retrospection was to explore test takers' attitudes toward test authenticity and test delivery mediums, the findings from many studies related to test takers' attitudes toward the tests were used as a guideline for this in-depth investigation.

Sukamolson (1993) investigated the test takers' attitudes toward Computerized Content-based Adaptive Testing of a Domain-Referenced Test by exploring the following

issues: test appeal, procedure complexity, test investment, test motivation and general attitude to tests. Piamsai (2006) reported the findings from the open-ended part of a questionnaire that the students' attitudes towards the test could be mainly classified into four main parts: interface design, quality, application and others. Moreover, the survey conducted by Geteborg (2006) focused on student perspectives on a good language test. The respondents evaluated the test by looking at the test items, test methods, the clarity of instructions, the timing of the various sections, the relevance of the content in the light of their learning experiences or their purposes for learning the language, the relationship between how they perceive their learning abilities and their performance on the test in question, and so on.

Accordingly, the topics and issues to be covered in the semi-structured interview were specified in advance by integrating the findings from the studies as follows:

1.1 General questions asking about the quality of the test to assess reading ability:

1. *The opportunity to demonstrate strengths and weaknesses in reading ability*: investigating how adequate test takers feel in terms of the opportunity to demonstrate both their strengths and weaknesses
2. *The perception of test difficulty*: exploring how test takers view the test they took as being difficult or easy
3. *The perception of test fairness*: eliciting how they feel about the fairness of the questions or situations in the test
4. *The perception of nervousness while taking the test*: investigating how nervous they feel while taking the test
5. *The perception of the clarity of the test directions*: exploring how clear is the test directions they perceive such as knowing what they needed to do
6. *The accuracy of the test to elicit their true ability in reading*: investigating how test takers perceive that the test can accurately assess their reading ability in real life situations

From the specified issues above, the following tentative questions were provided.

1. Do you think that you have sufficient opportunity in the test to demonstrate your strengths and weaknesses in reading ability?
2. How difficult is the test?

3. How do you feel about the questions in the test? Do they affect your score? How do you feel about the situations in the test? Do they affect your score?
4. How do you feel while taking the test? Are you in control of the test situation or do you feel nervous?
5. How clear are the test directions? Do you know what to do in the test?
6. How accurately does the test elicit your true reading ability? Does your true reading ability show in your real life reading activities?

1.2 The test administration:

1. *Test administration*: exploring how they feel about the test management.
2. *Time length*: investigating preference for the length of the time used to take the test.
3. *Facilities*: exploring the opinion on all facilities provided in the exam room.

The questions asked were as follows:

1. How do you like the test administration? Do you think it is more convenient than the traditional test?
2. How sufficient is the time provided for taking the test? Are you satisfied with the length of time used in the test?
3. How effective are the facilities used for the test administration (PBT: venue, CBT: the hardware and software)? In terms of CBT, how do you like the software functions?

1.3 The test characteristics

1. *Test contents*: exploring how they are interested in the reading texts and the test items provided in the tests.
2. *Scoring*: investigating the scoring method they perceive.
3. *Interactiveness*: investigating the perception of participating as a part of the test.
4. *Authenticity*: exploring how they perceive that the tests are relevant to their real life.

The questions asked were as follows:

1. How interesting are the test contents?
2. How effective is the scoring method?

3. How does the test allow you to interact with the test tasks?

4. How authentic are the test situations provided in the test? Are they relevant to your real life activities?

1.4 Candidates' performance

1. *Familiarity*: exploring whether they are familiar with the test (item types: short answer or multiple choice, test delivery: PBT and CBT).

2. *Perseverance*: investigating how they attempt to finish the test.

3. *Attitude*: summarizing their overall attitude toward the test.

The following questions were provided.

1. Are you familiar with the test delivered by the computer?

2. How hard have you tried to do the test?

3. Generally speaking, do you have positive or negative attitudes towards this test?

2. The interview was conducted with 5 students randomly selected from each group (totally 20 students from 4 groups). The data was recorded on video.

3. Once gathered, the video recordings were transcribed.

4. Then an encoding scheme was developed for the transcribed script to facilitate analysis. The test takers' responses were encoded as positive, negative and neutral attitudes. Since different people may verbalize things in different ways, this process was also conducted by a record coder. Therefore, inter-coder reliability was computed by assessing the extent to which the two coders agreed on the codes assigned to each segment. A high level of agreement (80 per cent and above) is usually sought between coders (Green, 2004).

5. In order to reveal test takers' attitudes, a content analysis technique was employed. The approach to content analysis used in this study was by frequency counts. In this approach, the units for coding identified and coding categories (positive and negative attitudes) defined were tabulated and then carefully counted.

3.4.2 Phase II: The Implementation of the study

1. 300 students were randomly selected from the population as the subjects of this study.

2. Then they were divided into 4 groups of 75.

3. The subjects in each group were blocked into 3 subgroups according to their English I course grades: B-A, C-C+, and D-D+ (25 students per subgroup).

4. The students' midterm and final scores from the English I course were used to test whether there was any significant difference (at the .05 level) between the groups in English Proficiency.

5. Each of the four groups was randomly assigned to take one version of the reading tests.

- Group 1: Authentic English Reading Comprehension Conventional Paper-and-Pencil Test (*ACON*)

- Group 2: Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test (*ICON*)

- Group 3: Authentic English Reading Comprehension Computer- Adaptive Test (*ACOM*)

- Group 4: Inauthentic English Reading Comprehension Computer- Adaptive Test (*ICOM*)

6. All the answer sheets of the students who took the paper-based tests: ACON and ICON were marked according to the answer key provided.

7. The scores obtained from the computer-based tests: ACOM and ICOM were collected from the software.

8. A week later, the retrospective interview to collect information about student attitudes toward test authenticity and test delivery mediums were conducted. Five test takers were randomly selected from each group. Each student was recorded about 10–15 minutes, and then all the recordings were transcribed. The data were coded and recoded by the researcher and another colleague.

3.4.3 Phase III: Data Analysis

1. Maximum Likelihood Estimation (Baker, 2001) was the method used to estimate the test takers' ability score (θ) in CATs (ACOM and ICOM). Then the ability score was transformed into a true score by using Test Characteristic Curve (TCC)

2. For the PBT (ACON and ICON), the students' scores were obtained from summing the number of correct items.

3. Two-way Analysis of Variance (2*2 ANOVA) was used to identify the main effects and the interaction effects of the main study. The 2*2 factorial design was used in this study because there were two independent variables, each with two levels.

4. Partial Eta squared reported by SPSS program was used to measure the effect sizes of test authenticity and test delivery mediums (Thalheimer and Cook, 2002).

5. The mode was used as a descriptive technique to examine the test takers' attitudes. The highest frequency of the attitudes towards test authenticity and test delivery mediums were also reported.

6. The chi square (χ^2) test was applied to find out whether there was any significant difference in the proportions of samples' attitudes towards test authenticity and test delivery mediums in each group.