# EFL 學生於紙本與電腦閱讀測驗表現之比較

# A Comparability Study of EFL Students' Performances between a Paper-Delivery and a Computer-Delivery Reading Exam


**by**

**翁佩禪　Pei-Chan Weng**


**THESIS**

Presented to the Faculty of the

Department of Foreign Languages and Literature of

Tunghai University

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF ARTS in

Teaching English As A Foreign Language


**TUNGHAI UNIVERSITY**

**June 2009**

**中 華 民 國 九 十 八 年 六 月**

## TABLE OF CONTENTS

## ACKNOWLEDGEMENTS

Writing this thesis was a long and difficult process. I would like to express my sincere gratitude to those who supported me through this process..

First and foremost, I would like to express my deepest appreciation to my advisor, Dr. James Sims. He was always patient and willing to offer his time to arrange countless meetings. In addition, his sense of humor and easygoing attitude made the process of doing this study easier and enjoyable. I really appreciate his support and patient guidance.

Next, I would like to express my gratitude to Dr. Chia-hui Chiu of the Department of Foreign Languages and Literature in Tunghai University and Dr. Yuh-fang Chang of the Foreign Languages and Literatures Department at Chung Hsing University for being my thesis committee members. Their insightful suggestions helped me refine my study and broadened my professional knowledge. I also want to thank the professors who offered their students to be the participants in this study. Without their help, this thesis would never have been completed.

Special thanks go to all the professors who taught me at Tunghai, and all my classmates, especially Bertha, Tina, Flora, Mandy, Janice, Grace, Catherine, Daniel, Juby, and Dusting. I am grateful that we could exchange ideas and share life experiences. Also, I want to thank my best friend, Heng-wei. Thanks for his

encouragement and support.

Finally, I want to thank my lovely family, my parents (Zi-geng and Xiu-mei), and my two brothers (Ren-wei and Yi-min). Their encouragement helped me overcome all the difficulties. I am grateful to be one member of the family. Nothing is more precious than having them on my side.

## ABSTRACT

Technology has had an important impact on education. In the area of language assessment, tests that are delivered through the traditional paper-and-pencil platform had been the main stream for decades. However, due to the rapid development of science and technology, in recent years, computers have become more and more prevalent and important in educational settings. More teachers and institutions use computer delivered exams instead of traditional paper-and-pencil exams. However, few studies have investigated the differences in measurements of reading comprehension between the platforms of paper and computer in Taiwan.

This study aimed to investigate university freshmen's performance on paper-delivery exams and computer-delivery exams of reading. Additionally, gender was examined to see whether this variable affected the scores of reading exams on the different test-delivery-mediums. Moreover, this study also investigated completion time for taking paper-delivery reading exams and computer-delivery reading exams.

The data for this study were collected from administering two different exams (Reading Exam A and Reading Exam B) which shared similar nature. Both of the two reading exams were delivered via paper and computer to 210 university freshmen. After collecting the data, independent-sample t-tests were utilized to analyze the mean scores on the reading exams for both paper-delivery platform and the

computer-delivery platform between groups, while paired-sample t-tests were utilized to analyze the mean scores on the reading exams for both paper-delivery platform and the computer-delivery platform within groups. In addition, scores in terms of participants' gender and completion time of the paper-delivery and computer-delivery platforms were analyzed. Paired-sample t-tests were used to determine if there were any significant differences between paper-delivery reading exam and computer-delivery reading exam for both females and males. As for completion time, participants' completion time was calculated by subtracting the starting time for the ending time.

The results showed that there was no significant difference between the two test platforms on reading. In terms of gender, it was found that there was no significant difference by gender. In addition, university freshmen completed the computer-delivered reading exams faster than the paper-delivered reading exams.

Keywords: Test-Delivery-Medium; Computer-Delivery Reading Test;

Paper-Delivery Reading Test

## 摘要

在語言測驗的領域中，傳統的紙本測驗一向是主流。由於科技的迅速發展，在近幾年中，電腦於教育上的應用已經愈來愈普及與重要。愈來愈多的老師以及相關機構，紛紛採用電腦測驗來取代傳統的紙本測驗。然而，在台灣，紙本與電腦閱讀測驗比較之相關研究仍然不多。

本研究主要目的，是針對大一學生的紙本閱讀測驗與電腦閱讀測驗之表現作分析。此外，不同性別的學生，於紙本閱讀測驗與電腦閱讀測驗之表現亦會進行分析。最後，學生於紙本閱讀測驗與電腦閱讀測驗之完成時間，亦會納入比較。

本研究的對象是 210 位東海大學的大一學生，依學生的新生入學考試成績，將學生分為兩組。針對此兩組學生，分別進行紙本閱讀測驗與電腦閱讀測驗。根據蒐集到的資料，本研究採用電腦統計軟體 SPSS 15.0 for Windows 來分析。再以 T 檢定的統計方法，來比較紙本閱讀測驗與電腦閱讀測驗的平均數。

本研究主要發現如下：第一、大一學生於紙本閱讀測驗與電腦閱讀測驗的平均數並無顯著差異。第二、不同性別的學生於紙本閱讀測驗與電腦閱讀測驗的平均數並無顯著的不同。第三、相較於傳統的紙本閱讀測驗，學生完成電腦閱讀測驗的時間比較快。



關鍵字：電腦閱讀測驗；紙本閱讀測驗

## LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

## Background of the Study

With the rapid development of science and technology, computers have become increasingly prevalent and important in education in recent years. A variety of computer software systems have been developed to help assess the performance of students at various levels (He & Tymms, 2005). According to research (Buchanan, 2000; Tymms, 2001; Tsai & Chou, 2002; Russell *et al*., 2003; Wang *et al*., 2004; Conole & Warburton, 2005), there has been an increase in the application of computer-based tests (CBT) in education organizations. Computer-based tests have many advantages over traditional paper-and-pencil based testing and assessment (Alderson, 2000; Brown, 2007; He & Tymms, 2005). These generally include unbiased and accurate scoring with immediate feedback, increased efficiency, convenient individualized administration, and improved test security.

Of the computerized testing procedures currently in use, computer-adaptive testing (CAT) has attracted particular attention in recent years. A number of testing programs and licensing agencies in the United States have been switching from

paper-and-pencil testing to CAT for the sake of efficiency and effectiveness (Dunkel, 1997). For example, the American College Testing Program offers a computer-adaptive placement test (named COMPASS) given to students entering college to assess their preparedness to do college work in reading and math.

Studies have indicated that it is becoming common to utilize computerized language tests; however, still many researchers claim that test administers should be concerned about utilizing computer technology in the environment of second/foreign language testing (Chalhoub-Deville and Deville, 2001; McNamera, 2000). One of the issues which has raised concern regarding the use of computerized language test is the effect of test-delivery-medium on the performance of test takers and its subsequent threats on the construct of validity of this kind of test. For example, Alderson (2000) claimed that reading texts on screen is not the same as reading texts on paper. In addition, some comparability studies showed that the time to complete computerized tests is different from the time to complete paper tests (Alexander, Bartlett, Truell, & Ouwenga, 2001; van de vijver & Harsveld, 1994).

These issues raise concerns about the platforms to measure reading. In addition, there is a need to determine if computer-delivery and paper-delivery exams test the same construct. As a result, many researchers (Chalhoub-Deville and Deville, 2001; Sawaki, 2001) are concerned about the test-delivery-medium effect and suggest test

administers to conduct further related studies. Therefore, the investigation of test-delivery-medium on the performance of test takers is needed.

The use of computers in language assessment has increased over the last decade and lots of comparability studies have contradictory results. In addition, few studies of this kind have been conducted in Taiwan. As a result, this study aims at addressing the issue of the comparability between computer-delivery exam and paper-delivery exam on reading comprehension of university freshmen in Taiwan.

## Statement of the Problems

In recent years, more and more tests have been delivered by computers. For example, a computer-based version of the TOFEL was introduced in 1998. Due to the prevalence of using computers, it is important to know whether reading text on a computer screen is different from reading text from printed pages (Alderson, 2000).

In addition, when readers read, they may utilize reading skills such as summarizing, inferencing, and scanning to facilitate their reading. However, is a reader's reading performance the same when using these reading skills on a computer screen and printed pages? For example, when a reader utilizes the reading skill of scanning to get the main idea of a passage from printed pages, it may be different from reading texts from a computer screen because note-taking, underlining, or

circling are not possible on a computer screen. In other words, test takers can write on a paper exam to highlight important information, but are unable to do this on a computer screen.

Even though several test comparability studies are available in the literature, not all studies obtain the same results. Furthermore, few comparability studies between paper and computer were conducted in Taiwan's EFL environment. Therefore, more investigation and analysis on language testing via computers are needed.

## Purpose of the Study

The purpose of this study is to investigate and compare paper-delivery and computer-delivery exams of reading. The aim is to determine whether there are any statistical differences in terms of test takers' performance on the two different test-delivery-mediums. Additionally, gender is examined to see whether the variable affects the scores of reading exams on different test-delivery-mediums. Moreover, this study also investigates completion time for taking paper-delivery reading exam and computer-delivery reading exam.

## Research Questions

This study attempts to answer the following questions:

1. Are there any statistical differences in EFL university freshmen's reading performance in terms of the total scores of computer-delivery reading exam and paper-delivery reading exam?

2. Are different groups of test-takers, according to their gender, affected differently by the test-delivery-mediums?

3. Is the completion time for computer-delivery reading exam and paper-delivery reading exam the same?

## The Significance of the Study

This study aimed to contribute to the area of language testing by comparing the test-delivery mediums. Since the use of computers has become more prevalent in language assessment, it is worthwhile to understand the differences between measurements of reading scores on computers and printed paper. By conducting this study, the researcher hopes that teachers and students have a better understanding of the test administered via computers and paper. Furthermore, the researcher hopes that the findings serve as a reference for teachers who are considering adapting tests to a computer platform.

## Definition of Terms

To ensure a consistent use and understanding of the terms used throughout this study, some key terms are defined as follow:

Reading Exam: In this study "reading exam" refers to the most common and traditional multiple-choice technique which is to read a passage and answer some questions (Brown 2007). Based on short passages, questions cover the comprehension of: main idea (topic), inference (implied detail), detail (scanning for a specifically stated detail), and vocabulary in context.

Paper-delivery Reading Exam (PDRE): The PDRE in this study refers to the traditional paper-and-pencil exam. The exam is administered via printed pages and test-takers need to write down their answers on answer sheets.

Computer-delivery Reading Exam (CDRE): A computer-delivery reading exam is somewhat similar to a traditional paper-and-pencil exam but utilizes a computer-mediated exam format. In this study, the CDRE is non-adaptive and the response format is the same as the PDRE. That is, test-takers are required to write down their answers on answer sheets.

# CHAPTER 2

# REVIEW OF THE LITERATURE

## The Nature of Reading

In order to develop reading assessment instruments, it is important to understand the nature of reading. Briefly speaking, reading refers to the ability to understand a written text. However, there are various particular variables and characteristics during the process of reading.

Many scholars provided different definitions of reading to explain the nature of reading. Grabe (1999) defined fluent reading as a rapid, interactive, and purposeful process that requires efficiency, strategies, sufficient knowledge of language, the world, and the reading topic.

According to Alderson (2000), reading is a process during which a reader interacts with a written text. In the process of reading, readers undergo a series of things. They look at written texts and think about what they are reading. For example, they think about the meaning of written texts, the relation between written texts and the things they know. Naturally, it is also possible that readers are completely unconscious of how they are reading, and of what is happening in the text.

To understand how a reader comprehends a written text, a distinction among levels of understanding of a text is essential. Some theories distinguish 3 levels of understanding of a text, which include the literal understanding of text, the inferred meaning of text, the understanding of the main implications of text (Alderson, 2000). According to Alderson (2000), the levels of comprehension are ordered. Readers first learn how to understand the literal meaning of texts, then to infer meanings from text. Later on, readers may approach texts critically and to evaluate texts.

Kintsch and Yarbrough (1982) differentiated among levels of reading comprehension. They claimed that comprehending the words is easier than comprehend the meaning of a sentence, while comprehending the organization of a text is more complicated than comprehending the meaning of a sentence. Understanding the words and sentences of a text is related to microprocesses which involve local and phrase-by-phrase understanding; whereas understanding the organization of a text is related to macroprocesses which involve global understanding.

In order to have a better understanding of the reading theories, next section briefly reviews the historical perspectives on reading.

## Past and Present Perspectives on Reading

When reading method was innovated in the 1920s, reading comprehension construct was systematically involved in the educational settings. At that time, vocabulary played a very important role that could improve the reading comprehension ability (Zimmerman, 1997).

During the late 1960s, with the popularity of the audio-lingual method, reading became a kind of teaching instrument on oral language teaching (Grabe, 1991). In the process of reading, readers' roles were passive that they should find out the meaning of a written text from the smallest units. That is, written texts were decoded by readers piece by piece. This view gave rise to a perspective of reading called bottom-up process.

According to Alderson (2000), the term bottom-up processing has been used to describe the decoding of the letters, words, and other language features in the text. In the model of the bottom-up processing, the reader process each individual letter encountered.

Brown (2007) also defined bottom-up processing as the process that readers must first recognize a multiplicity of linguistic signals (letters, morphemes, syllables, words, phrases, grammatical cues, and discourse markers) and use their linguistic

data-processing mechanisms to impose some sort of order on these signals.

According to Alderson (2000), bottom-up approaches are "serial models", where the reader starts with the printed word, recognizes graphic stimuli, decodes them to sound, recognizes words and decodes meanings.

However, Goodman (1969) questioned the "serial models" by using the "miscue analysis" which analyzed the mistakes made by readers while reading a written text aloud. The results of the miscue analysis showed that reading is not only simply a decoding process. Another alternative explanation to the reading process called top-down process was developed. The top-down process was based on the schema theory. The rational of the schema theory is that text themselves do not carry meanings. When processing a written text, readers' existed knowledge about the world will have a strongly effect to readers' reading ability and efficiency.

In recent literature on reading, the term top-down processing has been used to describe the application of prior knowledge to working on the meaning of a text (Alderson, 2000). The model emphasizes the reader rather than the text is at the heart of the reading process. Top-down processing also emphasizes the reconstruction of meaning rather than the decoding of form.

According to Goodman (1969), reading is a "psycholinguistic guessing game."

In the process of playing the game, readers use their existent knowledge to guess or predict texts' meaning. From this perspective, reading can be seen as a kind of dialogue between the reader and the text, or even between the reader and the author (Widdowson, 1984).

Later on, due to some deficiencies of the top-down approach, another alternative model, called interactive model, was used to describe the reading process. Grabe (1991) claimed that "interactive" represents the interaction between a reader and a written text.

There are still many other models to reading. Due to the focus of the study is not on reading theories, the introduction and explanation to other reading models are not mentioned here. However, most models to reading assumed that reading is an interactive process

## Variables that Affect the Nature of Reading

This section describes the variables that influence the nature of reading. Research shows those variables have impacts on either reading process or product and hence need to be taken into account during test design. According to Alderson (2000), the variables that affect the nature of reading can be divided into two subcategories: 1) reader variables, and 2) text variables. Following is an introduction to Alderson's

reader variables and text variables.

***Reader Variables***

Readers' background knowledge, subject/topic knowledge, cultural knowledge, and linguistic knowledge are some of the variables that influence the process of reading.

*Background knowledge* is related to readers' schemata. Readers integrate the new information from written texts into their existing schemata when they process texts. Therefore, what readers have known affects what they understand when reading texts. In terms of *Subject/topic knowledge*, it is obvious that if a reader knows absolutely nothing about a topic of a text, he/she will feel it is difficult to process the text. As for *cultural knowledge*, it refers to the way that other people's world work. Such worlds may be difficult to control or predict because of personal history and experiences unique to one person. Therefore, if a reader is not familiar with the cultural knowledge of a text, it is difficult to understand a text. In terms of *linguistic knowledge*, it involves phonological, orthographic, morphological, syntactic, and semantic information. In addition, it also includes discourse-level knowledge such as text organization and cohesion.

A reader's ability and purpose to reading are also the important variables that affect reading. Because every reader has different personalities and backgrounds, the

reader variables that influence reading are various and numerous. However, those reader variables provide test designers with an implication that they have to take those variables in to consideration and be aware that variables may well influence test scores or measures of reading.

***Text Variables***

The other variable that affects reading is text variables. According to Alderson (2000), text variables range from aspects of text content, to text types or genres, text organization, sentence structure, lexis, text typography, layout, the relationship between verbal and non-verbal text, and the medium in which the text is presented.

*Text content* affects how readers process text. For example, it is generally assumed that texts that describe real objects, events or activities are easier to understand than abstract texts. This claim provides test developers a direction for choosing appropriate text contents when designing reading tests.

*Text readability* also affects reading process. In educational context, it is important for test designers to identify what features make text readable. The commonly used readability formula is the Flesch that has been used from 1948 to now. In the Flesch Reading Ease test, higher scores indicate material that is easier to read; on the other hand, lower numbers indicate material that is more difficult to read.

Test designers have to consider the readability of written texts used for testing reading

comprehension. Each text for reading test should be appropriate in difficulty for the test takers being tested.

Another possible variable that has influence on reading is *typographical features*. The features of print, fonts, and layout might be crucial in causing reading ease or difficulty. For test designers, it is important to make sure that written texts are suitably presented for test takers.

Finally, especially relevant to the late 20$^{th}$ century, the effect of text variables on reading is the *medium* in which the text is presented. Nowadays, because more and more tests are being delivered by computer, it is important to know whether reading text on a screen is different from reading a printed text. This study aims to investigate this issue.

According to Alderson (2000), many studies are related to user interface characteristics which involve the aspects of the display on screen such as text fonts, color, and line spacing. In addition, Alderson (2000) suggested that more research is needed to explain how people process information presented via screens or other mediums. If there exits differences, more research have to engage in how they differ from each other. This was also the main focus of this study and was discussed later.

## Reading Purposes

Reading is an activity with a purpose. In the process of reading, readers may read a lot of materials for different intentions. A reader read novels or fictions for pleasure. A reader scans newspapers and magazines to find the sections he/she is interested in. A reader focuses on every unit of a text if he/she takes a reading test. Different reading purposes might cause the variations in the process of reading. In addition, different reasons for reading a text will affect the way and the skills readers read it.

Rivers and Temperly (1978) advocated seven main purposes for reading which involve reading to gain information out of curiosity or other reasons; to obtain knowledge on how some tasks are performed or some equipment work; to play a game; to correspond with others; to get information about the place and time of an event or the availability of something; to know what has happened or is happening; to enjoy or get pleasure.

In addition, Grabe and Stoller (2001) suggested four purposes for reading. These four purposes are to get major information, to learn new information, to integrate information and criticize texts, to have general comprehension. There are still other classifications of reading purposes advocated by other scholars and there exist some similarities and differences among these classifications of purposes.

In the process of reading, a good reader may set a purpose for reading or

determine what he/she would like to get out of a piece of writing. In order to get an idea of the purpose, good readers would try to ask themselves questions that begin with where, when, who, what, why, and how.

According to Alderson (2000), the only purpose test takers have in taking a reading test is to answer test designers' questions. Therefore, if test takers do not have to access questions, reading may become purposeless.

In addition, the purpose for reading also determines the appropriate approach to reading comprehension. In terms of the appropriate approach, it was briefly discussed in the next section "reading skills."

## Reading Skills

Scholars who believed that reading is a skill-based process defined reading skills into various classifications. Davis (1968) advocated eight reading skills which involve recalling word meanings; drawing inferences about the meaning of a word in context; finding answers to questions answered explicitly or in paraphrase; weaving together ideas in the context; drawing inferences from the context; recognizing a writer's purpose, attitude, tone and mood; identifying a writer's technique; following the structure of a passage.

Brown (2007) claimed that the skills and strategies for accomplishing reading are a crucial consideration in the testing of reading ability. Brown (2007) also divided reading skills into micro- and macro-skills to present the spectrum of possibilities in reading comprehension tests.

In this study, reading strategies such as skimming (understanding main idea), scanning (locating the specific information), and inferencing (understanding the information that is not explicitly mentioned in the text) were mainly tested. In addition, guessing the meaning of the unknown words and summarizing were also adapted. Those reading strategies are important and frequently used by readers. In the following, the five reading skills are briefly discussed.

***Skimming***

Skimming refers to the process of rapid coverage of a reading text to determine its main idea. This skill gives readers a sense of the topic and purpose of a text, the organization of the text, the point of view of the writer, and its ease or difficulty.

Skimming is essential when the reader wants to get a general idea of whether the topic is of his/her interest, whether s/he wants to read it at all, whether s/he is already familiar with the content of the selected text, whether s/he wants to find out how, whether the text provides new information or whether it is a general introduction to the topic.

### *Scanning*

Scanning refers to a strategy used by all readers to find relevant information in a text. Scanning through the text means that readers keep eyes on a text carefully without reading each line and keep the pace of reading up to find the information that they are looking for. For example, if the text is an essay, readers need to locate the setting for a narrative or story. Brown (2007) claimed that since the main purpose of scanning is to quickly identify important information, timing may also be calculated into a scoring procedure if the test focuses on the skill of scanning.

### *Inferencing*

A text is usually a combination of sentences which contains at least a meaning. Sometimes these meanings are directly stated in the text and the reader can understand them easily. However, some other times, the writer does not directly state the meaning and leave it to the reader to figure out. In these cases the reader needs to analyze the pieces of information of a text and logically come to a conclusion and understand the indirect meaning.

### *Guessing*

Guessing means that readers read a passage which involves new words and phrases. In order to understand the meaning form a passage, readers need to guess by

the context.

It is believed that skillful readers can understand the meaning of the unknown words through contextual clues. This ability is called guessing the meaning from the context. The contextual clues could be synonyms, antonyms, textual cohesive devices, collocations or even sentences or paragraphs around the word.

***Summarizing***

To summarize is to put in one's own words a shortened version of written or spoken material, stating the main points and leaving out everything that is not essential.

According to Alderson (2000), summarizing refers to a reading skill in which the reader should understand the text completely especially the main idea and try to convey the whole meaning in shorter form while deleting the unimportant details. Summarizing allows readers to monitor comprehension of material.

When readers read, they may utilize those reading skills above to facilitate their reading. As mentioned in the previous chapter, is a reader's reading performance the same when using these reading skills on a computer screen and printed pages? For example, according to Choi, Kim, & Boo (2003), students found it difficult to read passages without being able to use their pens for note-taking. In addition, students

found it difficult to concentrate when reading passages on computer screen due to eye fatigue

## Techniques for Testing Reading

For test designers, numerous test techniques can be utilized. Some test techniques are common only for the sake of convenience and efficiency. In this section, a number of different techniques for the assessment of reading listed by Alderson (2000) are briefly introduced.

Following are some of the more commonly used techniques in testing reading comprehension abilities represented by Alderson (2000).

1. Discrete-point and integrative techniques: In discrete-point approaches, the intention is to test one thing at a time; that is, test designers may wish to isolate one aspect of reading ability, or one aspect of language. On the other hand, in integrative approaches, the aim is to gain a much more general idea of how well test-takers read; that is, to evaluate the general understanding.
2. Cloze test: Cloze tests are typically constructed by deleting from selected texts every n-th word (n usually being a number somewhere between 5 and 12) and simply requiring the test-taker to restore the word that has been deleted. It was claimed that this type of test is ideal for a more global understanding of

test-takers' reading ability.

3. Gap-filling: it is a special type of cloze in which the test designer does not use a random procedure to select the words for deletion.

4. Multiple-choice techniques: Multiple-choice techniques are by far the commonest way for testing test-takers' reading comprehension. For test designers, the construction of multiple-choice questions is a very skilled and time-consuming business. They even have to do a pre-test to analyze item difficulties and item discriminations, and reject or modify items that have not performed well.

5. Alternative objective techniques: Alderson (2000) introduced three of the alternative objective techniques which involved matching techniques, ordering tasks, and dichotomous items. In terms of matching techniques, two sets of stimuli have to be matched against each other. In terms of ordering tasks, test-takers are given a scrambled set of words, sentences, paragraphs or texts, and have to put them into the correct order. In terms of dichotomous items, test-takers are presented with a statement which is related to a target text and have to indicate whether this is true or false.

6. Editing tests: Editing tests contain passages which errors have been introduced, and test-takers have to identify those errors.

7. Short-answer tests: In short-answer tests, test-takers are simply asked a question which requires a brief response. It is possible to interpret test-takers' responses to see whether they have really understood or not.
8. Information-transfer tests: In information-transfer tests, test-takers' task is to identify the required information and then to transfer it into a table, map or whatever.

Each of the techniques mentioned above has its own problems and limitations. For test designers, they need to choose the most appropriate techniques for the purpose of their test. In addition, it is important to understand that there is no "best technique" for testing reading. Test designers have to know that no single test technique can fulfill all the varied purposes for which they might test.

Since the main purpose of the study was to compare the performance of Taiwan EFL learners on computer-delivery and paper-delivery test on reading, a brief introduction of the history and work on computer-delivery language testing is provided.

## Computers on Language Testing

In recent years, computers have begun to offer attractive alternatives for testing in many educational settings and are currently being used to enhance the effectiveness

of language testing.

Tests that are administered at computer terminals, or on personal computers, are called computer-assisted language test. Receptive-response items including multiple-choice, true-false, and matching items are fairly easy to adapt to the computer-assisted testing medium. Even productive-response item types including fill-in and cloze can be created using authoring software like *Testmaster*. (Brown, 1997)

Various acronyms are used to describe different aspects of computer-assisted testing. In this section, two recent uses of computers in language testing are briefly introduced: 1) Computer-based testing, and 2) Computer-adaptive testing.

***Computer-Based Testing***

According to Wei (2007), the term, "computer-based testing," simply refers to testing administered at computer terminals or on personal computers. This kind of test can be either adaptive or non-adaptive. For the adaptive one, it is called "Computer-adaptive tests (CAT)." However, according to Brown (2007), computer-based testing is a general term which refers to the use of computer technology in a limited way in testing - probably just presenting items to test takers. That is, in computer-based testing, a fixed set of items are presented to test takers and

those items are not selected according to test takers' previous right-wrong response patterns.

Computer-based testing has been widely used by some organizations. Educational Testing Service (ETS) is already offering the GRE and PRAXIS as computer-based tests in 180 countries. In 1998, a computer-based version of the TOFEL examination was also adapted.

***Computer-Adaptive Testing***

Computer-adaptive tests are a subtype of computer-assisted language tests. However, the computer-adaptive language test has three extra characteristics: 1) the test items are selected and fitted to the individual students involved, 2) the test is ended when the student's ability level is located, and, as a consequence, and 3) computer-adaptive tests are usually relatively short in terms of the number of items involved and the time needed. (Brown, 1997)

Dunkel (1997) defined the computer-adaptive testing as tailored testing. That is, when a test-taker takes the computer-adaptive testing, the test items/questions the test taker receives are "tailored" to the test-taker's ability. If a test-taker gets an item correct, the next item is more difficult. On the other hand, if a test-taker gets an item wrong, the next item is easier. The first two sections (Listening and structure) of the

TOFEL test, for example, are computer-adaptive.

### *Internet-Based Testing*

According to Wei (2007), delivering a test via the Internet has become an alternative option for educators. Internet-based testing can be non-adaptive or adaptive, and can be more effective and efficient than traditional paper-and-pencil tests. The NWTPAW (National English test in Proficiency for All on the Web) is a nationwide Internet-based test.

## Effectiveness of Computer in Language Testing

Recent years have seen a burgeoning of assessment in which the test-takers perform responses on a computer. Some computerized testing has been utilized for large-scale language tests. However, there are advantages and disadvantages in using computers in language testing. This section depicts the advantages and limitations in language testing on computers.

### Advantages of Using Computers in Language Testing

According to Brown (1997), the advantages of using computers in language testing can be further divided into two categories that involve testing considerations and human considerations.

In terms of *testing considerations*, some of the advantages of using computers in language testing are as follows:

1. Computers are much more accurate at scoring selected-response tests than human beings are.
2. Computers are more accurate at reporting scores.
3. Computers can offer immediate feedback in the form of a test scores, complete with a printout of basic testing statistics.
4. The use of different tests for each student should minimize any practice effects, studying for the test, and cheating.

In terms of the *human considerations*, some of the advantages of using computers in language testing are as follows:

1. The use of computers allows students to work at their own pace.
2. Compared to traditional paper-and-pencil tests, computer-assisted language tests generally take less time to finish and are therefore more efficient.

***Limitation of Using Computers in Language Testing***

According to Brown (1997), the limitations of using computers in language testing can also be further divided into two categories that involve physical considerations and performance considerations.

In terms of *physical considerations*, some of the limitations of using computers in

language testing are as followed:

1. Computer equipment may not always be available. Reliable sources of electricity are not universally available.
2. Screen capability is another physical consideration. The amount of material that can be presented on a computer screen is limited. Such screen size limitations could be a problem, for example, for a group of teachers who wanted to develop a reading test based on relatively long passages.
3. The graphic capabilities of many computers (especially older ones) may be limited, and even those machines that do have graphics may be slow (especially the cheaper machines).

In terms of *performance considerations*, some of the limitations of using computers in language testing are as followed:

1. Differences in the degree to which students are familiar with using computers or typewriters keyboards may give rise to discrepancies in their performances on computer-assisted or computer-adaptive tests.
2. Computer anxiety (i.e. the potential debilitating effects of computer anxiety on test performance) is another potential limitation.

To sum up, when using computers on language testing, there are both advantages and limitations. Test designers should take those elements into consideration.

## Comparability Studies of Computer-Delivery Test and Paper-Delivery Test

Prior to the mid-1980s, few comparability studies on different test mode were conducted. Twenty years ago, for many test-takers, using a computer to take a test was a new experience. Therefore, it was difficult to investigate the positive and negative effects of using a computer for testing. (Russell, 2002; Russell, et. al., 2003)

Nowadays, due to increased assess to computers, the number of comparability studies of computer and paper has increased. Comparability studies investigate the possible effects due to the use of computerized tests instead of traditional paper-and-pencil tests. The purpose of this kind comparability study makes sure that test score interpretations remain valid and that test-takers are not disadvantaged by taking a computerized test instead of a traditional paper-and -pencil test.

In terms of test score on computerized tests and traditional paper-and-pencil tests, most of the early studies revealed that the mean score of paper-and-pencil tests to be significantly higher than the mean score of computerized tests (Bunderson *et al.*, 1989). They stated that lack of familiarity with computers might be the major factor in producing lower scores on computerized tests.

In last decade, Taylor, et. al. (1998) addressed issues about equity and bias of test-takers' performances on TOEFL Computer-based Test. The study investigated

approximately 90,000 TOEFL examinees in terms of their experience with computers. The result indicated that more than 80% of the TOEFL examinees were familiar with computers. In addition, there was no evidence of adverse effects on the computer-based TOEFL performance due to lack of prior computer experience. In other words, because of the ever-growing use of computers, the issue of computer familiarity is not expected to be a serious concern about comparability between paper-and-pencil tests and computerized tests.

However, Wallace and Clariana (2000) claimed that the issue of computer familiarity still affected the results of comparability study between paper-and pencil tests and computerized tests. According to Wallace and Clariana (2000), students who were less familiar with computers would not do well with online learning and testing. Correspondingly, Watson (2001) reported that students with greater familiarity of computer benefited most from computer-aided learning.

Even though nowadays students are more familiar with using computers, Choi, Kim, & Boo (2003) claimed that not all students preferred to read passages on computer screen. Students who were used to reading passages and taking notes with their pens, found it difficult to read passages without being able to use their pens for note-taking. Students also found it difficult to concentrate when reading passages on computer screen due to eye fatigue. They complained about eye fatigue caused by

continued intense exposure to monitor screens. However, Choi, Kim, & Boo (2003) also claimed that with the development of computer and internet technology, this problem may be solved easily as people become more accustom to reading from a computer screen.

The issue of "whether there was any significant difference in reading performance of paper-and-pencil tests and computerized tests" was also conducted by many researchers, but these results were inconclusive or had contradictory results. What follows is a review of some of these studies.

Some studies (Blackhurst, 2005; Bodmann & Robinson, 2004; Delaware, 2008) showed that computers were used to administer traditional multiple choice test without any significant effect on test-takers' performance. In other word, it was claimed that there was no significant difference between computerized tests and paper-and-pencil tests.

In Blackhurst's study (2005), 785 participants took part in this study and they had to take both paper-based and computer-based versions of exams. The results revealed that the participants' scores did not differ significantly by different modes of administration (computer-based exam and paper-based exam). Likewise, in Bodmann & Robinson's study (2004), fifty-five undergraduate students participated a study to

compare different modes. Approximately half the class took the first test on the computer, while the rest of the class took the first test on paper. The procedures were switched for the second test. For the test scores, the result showed that there was no significant difference between the paper-and-pencil test and computerized test. Moreover, in Delaware's study (2008), fifty-four undergraduate students took both the on-line test and the in-class test. The result revealed that there were no significant differences in students' test scores between the online tests and the in-class tests. In other words, different test mediums did not affect students' test performance.

Other studies reported that test performance was significantly affected by different test mediums. Choi, Kim, & Boo (2003), Bunderson *et al*. (1989) reported that participants performed significantly better on paper-and-pencil tests, while Clariana & Wallace (2002) claimed that the scores on computerized tests were significantly higher than that on paper-and-pencil tests.

Bunderson *et al*. (1989) overviewed several comparability studies and concluded that the scores on paper-and-paper tests were more often significantly higher that on computerized tests. This result may be due to participants' unfamiliarity with computers. In recent literature, Choi, Kim, & Boo (2003) also claimed that students performed better on paper-and-pencil tests for the reason that they found it difficult to read passages without being able to use their pens for note-taking and to concentrate

when reading passages on computer screen.

On the other hand, in Clariana & Wallace's study (2002) of 105 business undergraduates who took both a paper-and-pencil test and a computerized test showed that the computerized test group outperformed the paper-and-pencil test group.

In addition to the comparison on test score between computers and paper, studies also investigated the completion time and test-takers' characteristics such as gender, proficiency levels, and native language. For example, Bodmann & Robinson, 2004 and Wang *et al.*, 2008 showed that the time to complete computerized tests was shorter than the time to complete paper tests.

In Bodmann & Robinson's study (2004), fifty-five undergraduate completed a computer-based assessment faster than a paper-based assessment. According to the study, the reason may be due to the flexibility of paper-and-pencil tests. For paper-and-pencil tests, students could go back to review previous items, and thus the completion time increased. In addition, Wang (2008) review of comparability studies reported that the reason why completion time was faster on computerized tests may be due to the simplicity of clicking answers with a mouse or typing an answer from a keyboard.

In terms of gender, some studies showed that test takers' gender was not related

to performance difference on computerized tests and paper-and-pencil tests. Blackhurst (2005) claimed that different groups of participants (for example, gender and first language) were not affected by the format of tests. Clariana & Wallace's study (2002) found gender, competitiveness, and computer familiarity were unrelated to test mode effects. Likewise, Delaware (2008) investigated the impact of assessment methods on students' performance on exams. Gender was included as one of the independent variables. The findings revealed that students' gender did not affect the test scores on paper-and-pencil tests and computerized tests.

In conclusion, each medium (paper and computer) has its own advantages and limitations. A review of the literature revealed contradictory results. Some studies showed that there were significant differences between the performance of reading scores based on different test mediums (paper and computer). However, other studies showed that there were no significant differences. Furthermore, few studies have been conducted for EFL university students in Asia. Therefore, this study aims to clarify the issue.

# CHAPTER 3

# METHOD

This part presents the research methodology. The first section describes the characteristics of the subjects involved in this study. The second section introduces the two instruments (Reading Exam A and Reading Exam B), the two different delivery platforms (Paper-delivery and Computer-delivery), and the test specifications. The third section gives a detailed description of the data collection procedures undertaken in this study. Finally, the last section of this chapter reports the statistical software and procedures used to analyze the data.

## Participants

The participants for this study were freshmen in their second semester of Freshman English for Non-Majors (FENM) at Tunghai University for the 2008 academic school year. They were from the College of Management, Science, and Social Science.

210 non-major freshmen (143 females and 67 males) participated in this study. To ensure there was no significant difference in the reading ability between groups of the onset of the study, the participants were divided into two groups, named Group

One and Group Two respectively based on the results of Tunghai English Placement Exam (TEPE). The TEPE is an English proficiency test (Sims, 2006) which includes three sections: grammar (20%), reading (40%), and listening (40%).

In order to determine if there were differences in the two groups' reading ability, at the onset of the study, SPSS 15 was used to do the EXPLORE to check assumptions and obtain confidence intervals (CIs) testing mean differences. As the study focuses on reading, only participants' reading scores from the TEPE were utilized to do the EXPLORE.

The result of EXPLORE Analysis is presented in Table 3.1. The skewness and kurtosis for the TEPE scores on the reading section for both groups were well within the range of -2 and +2, which suggested that the distributions of the TEPE scores on the reading section for both groups were reasonably normal. The researcher also found that the means for Group One and Group Two were 30.47 and 31.60, respectively. The lower and upper bounds of the $CI_{.99}$ are 29.45 and 31.49 and 31.60 and 30.80 for Group One and Group Two, respectively. According to Bachman (2005), since both group means fell within each other's $CI_{.99}$, we can be 99 percent confident that the two groups' performance on the reading section of Tunghai English Placement Exam (TEPE) was not different. To sum up, there appeared to be no significant differences in the reading ability of the two groups based on the result of

Tunghai English Placement Exam (TEPE).

***Table 3.1 The result of EXPLORE Analysis Descriptive***

| Group | | | Statistic | Std. Error |
|---|---|---|---|---|
| Group One | Mean | | 30.47 | .391 |
| | CI.99 for Mean | Lower Bound | 29.45 | |
| | | Upper Bound | 31.49 | |
| | Skewness | | -.872 | .211 |
| | Kurtosis | | .491 | .419 |
| Group Two | Mean | | 31.60 | .303 |
| | CI.99 for Mean | Lower Bound | 30.80 | |
| | | Upper Bound | 32.39 | |
| | Skewness | | .042 | .209 |
| | Kurtosis | | -.240 | .416 |

## Instrument

The instruments that were used to collect the data in this study were two different exams, Reading Exam A and Reading Exam B (see Appendix A). Both of the two reading exams were delivered through two platforms, paper and computer. The following section presents the procedures outlined by Brown (2007) to evaluate the reading instruments utilized in this study including test specifications, followed by the construction of test items, and the evaluation and revision of test items. Finally, the

validity and reliability of the two instruments are addressed.

### *The Test Specifications and Construction of the Reading Exams*

There were two different reading exams with similar nature and construction. Each exam was composed of four passages of 150-350 words with the first article less difficult than the next. Most passages were adopted from the reading section of Tunghai Freshman English Non Major Exams (FENM). As Sims (2006) stated, the passages have three characteristics; "1) a clear, straightforward, factual introduction, simple in style, and a very clear, explicit thesis statement at the end of the introductory paragraph; 2) a body with unified, coherent paragraphs headed by clear topic sentences; and 3) a clear conclusion in the last paragraph."

Regarding the test format, a multiple-choice format was selected. The multiple choice format is most frequently used in educational testing. As Bailey (1998) stated, multiple-choice tests are fast, easy, economical to score, and can be scored objectively. According to Jacobsen (2008), there are several advantages to multiple choice tests. First, it allows more adequate sampling of content. Second, it tends to more effectively structure the problem to be addressed. Third, items can be more efficiently and reliably scored than supply items. Fourth, different response alternatives can provide diagnostic feedback (item analysis). Fifth, items can be constructed to address various levels of cognitive complexity.

Because the format of multiple-choice has the advantages above, this format was adapted in the study. There were totally 24 items within each test. Each question had three or four plausible choices, but only one correct answer.

In accordance with Tunghai FENM testing guidelines (Sims 2006), test committees composed of five to seven experienced Freshman English teachers developed most the reading exams. Most of the passages had been used on midterm or final exams in previous years. The reading passages and questions were selected and modified based on the results of test analysis from these previous administrations.

Regarding the construction of test items, four guidelines were utilized for designing the multiple-choice items: 1) each item measured a specific objective; 2) both the question and distracters are stated simply and directly; 3) the intended answer is the only correct answer; and 4) items are accepted, discarded or revised based on item difficulty, item discrimination, and distracter analysis.

According to Hughes (2003) and Brown (2007), reading exams should ask both macro questions, questions that require some integration or generalization from specific sentences, and micro questions, questions that focus directly on specific sentences or parts of sentences. Therefore, as suggested by Sims (2006), each reading exam had macro questions about the following: main idea of article, main idea of paragraphs, and inferencing; and micro questions about: general /details and

vocabulary in context. In addition, the layout of different types of questions within the two different reading exams was the same. For example, Question One in the Reading Exam A was a micro question about vocabulary in context, and so was Question One in the Reading Exam B. In other words, all the questions on Reading Exam A corresponded to the questions on Reading Exam B.

To sum up, Reading Exam A and Reading Exam B were constructed to measure the same reading skills using the same question types. Detailed information on the test specifications of the two different reading exams is listed in Table 3.2, along with their corresponding items.

The Exams were designed to measure five specific reading skills or items. These were 1) Main idea of passage; 2) Main idea of paragraph; 3) Inference; 4) Vocabulary in context; 5) General comprehension / Details. Each reading exam had equal numbers of corresponding items. As shown in Table 3.2, each reading exam involved three main idea of passage questions, five main idea of paragraph questions, two inference questions, six vocabulary in context questions, and eight general comprehension / details questions.

***Table 3.2 The Test Specifications of Reading Test A and B***

| Test | Test Specification | Item Number |
|---|---|---|
| Reading Exam A | (1) Main idea of passage | 10, 13, 19 |
| | (2) Main idea of paragraph | 4, 5, 6, 14, 20 |
| | (3) Reading for inference | 11, 12 |
| | (4) Vocabulary in context | 1, 2, 3, 15, 16, 21 |
| | (5) Reading for details | 7, 8, 9, 17, 18, 22, 23, 24 |
| Reading Exam B | (1) Main idea of passage | 10, 13, 19 |
| | (2) Main idea of paragraph | 4, 5, 6, 14, 20 |
| | (3) Reading for inference | 11, 12 |
| | (4) Vocabulary in context | 1, 2, 3, 15, 16, 21 |
| | (5) Reading for details | 7, 8, 9, 17, 18, 22, 23, 24 |

***The Evaluation and Revision of Test Items***

According to Bachman (2004), item analysis can provide feedback to test designers to improve the usefulness of a test, enabling the test designers to 1) control the characteristics of the total score distribution such as the item difficulty and item discrimination; 2) increase the reliability of a test; and 3) make a diagnosis whether items function appropriately.

Based on Bachman's (2004) statement, this study utilized item analysis to determine the item difficulty and discrimination to evaluate the test items on the Reading Exam A and Reading Exam B.

For the item analysis of the two different reading exams, the researcher did a pilot study to collect students' scores to estimate item difficulty and item discrimination. The pilot study was administered for 213 non-English major freshmen

(different from the participants of the study) in March, 2008. Students used for the pilot test took the two different reading exams. After students' scores were calculated, the researcher utilized this quantitative data to conduct item analyses.

In terms of the difficulty level of the two exams, the correct percentage of 22 items on the Reading Exam A was above 50% and the correct percentage of 21 items on the Reading Exam B was above 50%; the difficulty level of two items was below 50% on the Reading Exam A and the difficulty level of three items was below 50% on the Reading Exam B (See Appendix B).

With regard to the item discrimination of the two different reading exams, the researcher compared the upper and lower 27% of students' performances. Through analyzing each item's discrimination, the researcher found that all items could discriminate between the upper and lower 27% of the participants (See Appendix B).

In the two reading exams, two items were problematic and needed to be revised. For Question 7 on the Reading Exam A, 10% students chose A, 23% students chose B, 38% students chose C, 29% students chose D. The correct answer was A, but most students chose C. This indicated that Question 7 needed to be revised and retested. This was because compared to the item difficulty of Question 7 on Reading Exam B (67%), the item difficulty of this item on Reading Exam A (10%) was too low (See Appendix B).

The original item of Question 7 on the Reading Exam A was:

7. According to the passage, which of the following is a reason arranged marriages work?

   (A) Parents make a logical choice.

   (B) Parents know who their children love.

   (C) It takes a lot of thinking.

   (D) It will not make young people unhappy.

   Correct answer: (A)

The researcher revised option (A) and the new item was "Parents can make a logical choice." After revising this item, the result of its item difficulty became 82% (See Appendix B). This revision ensured the item difficulty for Question 7 on Exam A was equivalent to the item difficulty of the correspond Question 7 on Exam B.

In addition, for Question 11 on the Reading Exam B, 80% students chose A, 17% students chose B, 3% students chose C. The correct answer was B, but most students chose A. This indicated that Question 11 needed to be revised and retested. This was because compared to the item difficulty of Question 11 on Reading Exam A (85%), the item difficulty of this item on Reading Exam B (17%) was too low (See Appendix B).

The original item of Question 11 on the Reading Exam B was:

11. Based on the passage, which of the following opening lines is likely to be most successful?

   (A) Haven't I seen you in a beauty contest somewhere?

(B) Hi, my name is Jeff.

(C) You look like a girlfriend I dumped.

Correct answer: (B)

The researcher revised option (A) and the new one was "These flowers aren't as beautiful as you." After revising this item, the result of its item difficulty became 79% (see Appendix B). This revision ensured the item difficulty for Question 11 on Exam B was equivalent to the item difficulty of the correspond Question 11 on Exam A.

***The Readability of the Reading Exams***

The researcher also analyzed whether the two reading exams shared similar degree of readability. First, the readability was calculated to determine 1) the number of words, paragraphs, sentences per passage; 2) the average of words, paragraphs, and sentences per passage; and 3) the Flesch Reading Ease and Flesch-Kincaid Grade Level. These were calculated by using the Readability Statistics under spell check in World for Windows.

Readability statistics was used to make sure that both the two reading exams were at a similar reading level. Some modifications of the texts were carried out to make sure the readability statistics of Exam A and Exam B similar. The readability Statistics for the two reading exams are presented in Table 3.3.

***Table 3.3 Readability of the Two Reading Exams***

| | A-1 | B-1 | A-2 | B-2 | A-3 | B-3 | A-4 | B-4 |
|---|---|---|---|---|---|---|---|---|
| Calculation | | | | | | | | |
| Words | 272 | 351 | 161 | 147 | 293 | 290 | 250 | 221 |
| Paragraphs | 4 | 6 | 1 | 1 | 5 | 4 | 3 | 2 |
| Sentences | 20 | 22 | 13 | 12 | 13 | 15 | 15 | 11 |
| Average | | | | | | | | |
| Sentences per paragraph | 5.0 | 3.6 | 13.0 | 12.0 | 2.6 | 3.7 | 5.0 | 5.5 |
| Words per paragraph | 13.6 | 15.9 | 12.3 | 12.2 | 22.5 | 19.3 | 16.6 | 20.0 |
| Characters per word | 4.8 | 4.2 | 4.0 | 4.2 | 4.4 | 4.6 | 4.6 | 4.7 |
| Readability | | | | | | | | |
| Passive | 0% | 18% | 7% | 0% | 23% | 6% | 13% | 36% |
| Flesch Reading Ease | 54.0 | 65.2 | 71.8 | 70.6 | 58.0 | 53.0 | 48.8 | 54.9 |
| Flesch-Kincaid Grade Level | 8.9 | 7.9 | 6.3 | 6.4 | 10.7 | 10.6 | 10.5 | 10.1 |

Note. A-1 refers to Exam A Paragraph 1; B-1 refers to Exam B Paragraph 1.

The Flesch Reading Ease and Flesch-Kincaid Grade Level for each correspond passages revealed that each passage had similar readability and grade level. This indicated that both exams were of equivalent reading difficulty or level.

***The Validity and Reliability of the Reading Exams***

To evaluate validity and reliability of the instruments, the researcher conducted analyses in order to determine or argue whether the two reading exams are appropriate and reliable instruments.

As suggested by Alderson, Clapham, and Wall (2002), Cohen (1994), Hughes

(2003), and Tuckman (1998), an exam is determined to be valid based on a comparison of test specifications and test content. Following Hughes' recommendations that were maintained above, these comparisons were made by four Freshman English teachers who were trained in language teaching and testing but were not directly involved in the production of the exam. These teachers concluded that the exams were a valid measure of reading for Taiwanese students.

As to the reliability of the two different reading exams, the researcher split the test items into odd an even number groups to estimate the two tests' reliability coefficient. After the calculation, the reliability coefficient of the two different reading tests was found to be .82 and .87, respectively. Therefore, the two reading exams can be considered as reliable instruments.

***The Computer-Delivery Platform and the Paper-Delivery Platform***

After creating two different reading exams with similar nature, the researcher delivered the two reading exams through two different mediums, paper and computer.

Regarding the paper-delivery platform, it was the printout of Microsoft Word and the font was 12 with single space; for the computer-delivery platform, all the formats were the same as the paper-delivery platform expect it was displayed in Adobe Acrobat 7.0 Document on the computer screen. Both the computer-delivery platform and the paper-delivery platform required writing down the answers on the answer

sheets.

## Data Collection Procedure

As is stated above, the participants were divided into two groups. Each group had to separately take one reading exam through the computer-delivery platform and another reading exam through the paper-delivery platform. For the participants in Group One, they took the Reading Exam A via the computer-delivery platform in computer rooms ST023 and ST019 in the Science and Technology Building on March 19, 2008. Then they took Reading Exam B via the paper-delivery platform in their normal classrooms on March 26, 2008. For the participants in Group Two, they took the Reading Exam A via the paper-delivery platform in their normal classrooms on March 19, 2008. Then they took Reading Exam B via the computer-delivery platform in the same computer rooms on March 26, 2008. Table 3.4 outlines the data collection procedure of the study.

***Table 3.4 The Data Collection Procedure***

| | Week 1 | Week 2 |
|---|---|---|
| Group 1 | Exam A on Computer | Exam B on Paper |
| Group 2 | Exam A on Paper | Exam B on Computer |

Before each exam was administered, the participants were informed of the purpose of the exam by the researcher. Then the participants were instructed how to fill out the answer key. For both the paper-delivery reading exams and computer-delivery reading exams, students were instructed to write down their answers on an answer sheet and select only one answer for each question. In addition, the participants were instructed to write down their "starting time" and "ending time." A large clock was positioned in the front of each classroom for students to use to write down their times. After announcements were given, the participants started to take the test. There was not time limit for both the paper-delivery reading exams and computer-delivery reading exams.

The announcement and instruction for the paper-delivery reading exams and the computer-delivery reading exams were almost the same, except for the computer-delivery reading exams students had to go to the following website (http://teaching.thu.edu.tw/g941210/) to view the exams. Once all the students had logged onto the website they began the computer-delivery reading exams.

As for the paper-delivery reading exams, the participants were given paper copies to read. These paper copies and answer sheets handed out by the teachers. Once again, a large clock was positioned in front of the classroom for students to use.

After participants completed each reading exam, they had to write down the

basic information which includes college, class, student number, gender, and completion time.

## Data Analysis Procedures

The participants' scores on Reading Exam A and Reading Exam B were analyzed by using the SPSS 15.0 for Windows.

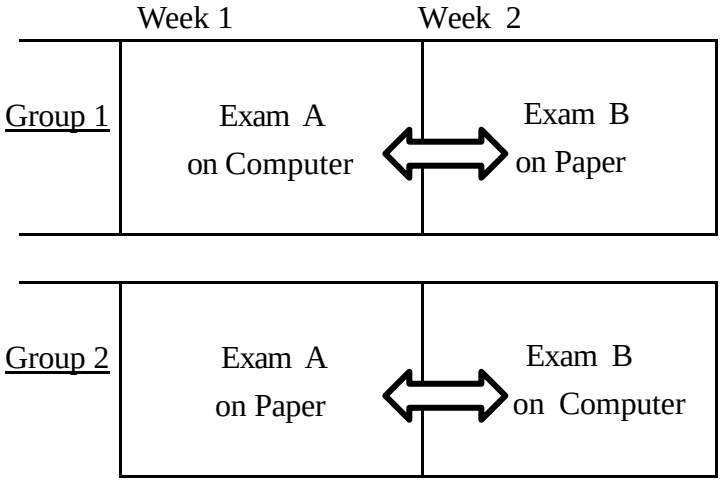
First, independent-sample t-tests were utilized to analyze the mean scores on the reading exams for both paper-delivery platform and the computer-delivery platform between groups. In other words, an independent-sample t-test was done between Group One on computer-delivery reading exam A and Group Two on paper-delivery reading exam A for week one. Another independent-sample t-test was done between Group One on paper-delivery reading exam B and Group Two on computer-delivery reading exam B for week two. Table 3.5 presents the data analysis procedure between groups.

***Table 3.5 The Data Analysis Procedure (Between Groups)***

| | Week 1 | Week 2 |
|---|---|---|
| Group 1 | Exam A on Computer | Exam B on Paper |
| Group 2 | Exam A on Paper | Exam B on Computer |

Then, paired-sample t-tests were utilized to analyze the mean scores on the reading exams for both paper-delivery platform and the computer-delivery platform within groups. In other words, for Group One, a paired-sample t-test was done between computer-delivery reading exam A (week one) and paper-delivery reading exam B (week two). As for Group Two, another paired-sample t-test was done between paper-delivery reading exam A (week one) and computer-delivery reading exam B (week two). Table 3.6 presents the data analysis procedure within groups.

***Table 3.6 The Data Analysis Procedure (Within Groups)***




The probability level of significance for the t-tests was set at .05. By analyzing the t-tests, the researcher analyzed whether there was any statistical score difference between reading exams through paper-delivery platform and the computer-delivery platform.

SPSS 15.0 was also used to analyze scores in terms of participants' gender. First, a paired-sample t-test was used to determine if there were any significant differences between paper-delivery reading exam and computer-delivery reading exam for females. Likewise, another paired-sample t-test was used to determine if there were any significant differences between paper-delivery reading exam and computer-delivery reading exam for males.

In addition, completion time of the paper-delivery and computer-delivery platforms was calculated. According to each participant's information about starting time and ending time, the researcher calculated participants' completion time by subtracting the starting time for the ending time. The completion time was calculated in minute intervals.

# CHAPTER 4

# RESULTS AND DISSCUSSION

This chapter presents a summary of the results of the collected data and a discussion of the findings. First the mean scores and standard deviation for the paper-delivery exams are presented, followed by the results of the computer-delivery exams. Next the results of a comparison of the computer-delivery exams and the paper-delivery exams between the two groups are presented. Then a comparison of the computer-delivery exams and the paper-delivery exams within the same group is presented. After that, the issue of order affect is addressed by a cross comparison between computer-delivery and paper-delivery exams. Finally the results of the analysis of gender and completion time are presented.

## Mean Scores and Standard Deviation

As presented in chapter three, it is important to note that measurements of reading to both groups appeared to be equivalent of the onset of the study. For the participants in Group One, they took the Reading Exam A via the computer-delivery platform in computer rooms ST023 and ST019 in the Science and Technology Building on March 19, 2008. Then they took Reading Exam B via the paper-delivery platform in their normal classrooms on March 26, 2008. For the participants in Group

Two, they took the Reading Exam A via the paper-delivery platform in their normal classrooms on March 19, 2008. Then they took Reading Exam B via the computer-delivery platform in the same computer rooms on March 26, 2008. Table 3.4 outlines the data collection procedure of the study.

For Group One (week one), the mean score and standard deviation of computer-delivery reading exam A were 20.72 and 2.516, respectively. For Group Two (week one), the mean score and standard deviation of paper-delivery reading exam A were 20.70 and 2.604, respectively. For Group One (week two), the mean score and standard deviation of paper-delivery reading exam B were 20.27 and 2.338, respectively. For Group Two (week two), the mean score and standard deviation of computer-delivery reading exam B were 20.81 and 2.071, respectively. Detailed information about the mean scores and standard deviation is reported in Table 4.1.

***Table 4.1 Mean Scores and Standard Deviation***

| | Group One | | | Group Two | | |
|---|---|---|---|---|---|---|
| Mediums | *N* | *M* | *SD* | *N* | *M* | *SD* |
| Computer | 105 | 20.72 | 2.516 | 105 | 20.81 | 2.071 |
| Paper | 105 | 20.27 | 2.338 | 105 | 20.70 | 2.604 |

**Independent-Sample T-Tests of Students' Scores on Reading Exams**

To compare the two different mediums, paper and computer, independent t-tests were conducted to compare the t-ratio obtained from both the paper-delivery and computer-delivery versions of each Reading Exams between groups. First, an independent t-test compared Exam A on computer from Group One and Exam A on paper from Group Two (See Table 3.5). This was done to see if there were any significant differences between the paper-delivery and computer-delivery versions for Exam A. The result of the independent t-test is presented in Table 4.2. The independent t-test indicated the calculated t-ratio (0.081) at 208 degree of freedom did not exceed the t-critical value (1.960). Therefore, we can be 95 per cent confident that there was no significant difference between the paper-delivery and computer-delivery versions of Reading Exam A.

***Table 4.2 Independent T-test between the paper-delivery and computer-delivery versions of Reading Exam A***

| | N | df | t | Sig. |
|---|---|---|---|---|
| Exam A | 210 | 208 | .081 | .936 |

Likewise, the same comparison was conducted for Exam B. In other words, another independent t-test compared Exam B on paper from Group One and Exam B on computer from Group Two (See Table 3.5). This was done to see if there were any

significant differences between the paper-delivery and computer-delivery versions for Exam B. The result of the independent t-test is presented in Table 4.3. The independent t-test indicated the calculated t-ratio (1.781) at 208 degree of freedom did not exceed the t-critical value (1.960). Therefore, we can be 95 per cent confident that there was also no significant difference between the paper-delivery and computer-delivery versions of Reading Exam B.

***Table 4.3 Independent T-test between the paper-delivery and computer-delivery versions of Reading Exam B***

| | N | df | t | Sig. |
|---|---|---|---|---|
| Exam B | 210 | 208 | 1.781 | .076 |

To sum up, there appeared to be no significant differences between the paper-delivery and computer-delivery versions for either Exam A or Exam B.

**Paired-Sample T-Test of Students' Scores on Reading Exams**

In order to determine if there was any significant difference between paper-delivery and computer-delivery reading exams within groups, paired-sample t-tests were conducted to compare the t-ratio obtained from the paper-delivery and computer-delivery versions of each Reading Exam.

First, as shown in Table 3.6, for Group One, a paired-sample t-test was

conducted between computer-delivery reading exam A (week one) and paper-delivery reading exam B (week two). The result of the paired-sample t-test is presented in Table 4.4. The paired-sample t-test indicated the calculated t-ratio (1.533) at 104 degree of freedom did not exceed the t-critical value (1.980). Therefore, we can be 95 per cent confident that for Group One, there was no significant difference between the paper-delivery reading exam B and computer-delivery reading exam A.

***Table 4.4 Paired T-test between the paper-delivery and computer-delivery versions of Reading Exams (Group One)***

| | N | df | t | Sig. |
|---|---|---|---|---|
| Group One | 105 | 104 | 1.533 | .128 |

Likewise, the same comparison was conducted for Group Two. In other words, for Group Two, another paired-sample t-test was conducted between paper-delivery reading exam A (week one) and computer-delivery reading exam B (week two). The result of the paired-sample t-test is presented in Table 4.5. The paired t-test indicated the calculated t-ratio (0.314) at 104 degree of freedom did not exceed the t-critical value (1.980). Therefore, we can be 95 per cent confident that there was also no significant difference between the paper-delivery reading exam A and computer-delivery reading exam B.

***Table 4.5 Paired T-test between the paper-delivery and computer-delivery versions of Reading Exams (Group Two)***

| | N | df | t | Sig. |
|---|---|---|---|---|
| Group Two | 105 | 104 | -.314 | .754 |

To sum up, there appeared to be no significant differences between the paper-delivery and computer-delivery platforms within each group.

The results are consistent with the findings of studies (Blackhurst, 2005; Bodmann & Robinson, 2004; Delaware, 2008) which reported that computers were used to administer traditional multiple choice test without any significant effect on test-takers' performance. In other word, there appeared to be no differences between computer-delivery exams and paper-delivery exams.

## The Effect of Gender on Reading Exams

As shown in Table 4.6, females' mean score and standard deviation on computer-delivery reading exam were 20.93 and 2.25, respectively. As for the paper-delivery reading exam, females' mean score and standard deviation were 20.64 and 2.21, respectively. In terms of the males, the mean score and standard deviation on computer-delivery reading exam were 20.37 and 2.373, respectively. As for the paper-delivery reading exam, males' mean score and standard deviation were 20.18 and 2.954, respectively.

***Table 4.6 Mean Scores and Standard Deviation (Gender)***

| | Females | | | Males | | |
|---|---|---|---|---|---|---|
| Mediums | *N* | *M* | *SD* | *N* | *M* | *SD* |
| Computer | 143 | 20.93 | 2.25 | 67 | 20.37 | 2.373 |
| Paper | 143 | 20.64 | 2.21 | 67 | 20.18 | 2.954 |

In addition, paired-sample t-tests were conducted to see whether there was any significant difference between paper-delivery reading exam and computer-delivery reading exam by genders. First, a paired-sample t-test was used to determine if there was any significant difference between paper-delivery reading exam and computer-delivery reading exam for females. As shown in table 4.7, in terms of females, the paired t-test indicated the calculated t-ratio (1.312) at 142 degree of freedom did not exceed the t-critical value (1.96). Therefore, we can be 95 per cent confident that there was no significant difference between the paper-delivery and computer-delivery reading exams for females.

Likewise, another paired-sample t-test was used to determine if there was any significant difference between paper-delivery reading exam and computer-delivery reading exam for males. As reported in table 4.7, the paired t-test for males indicated the calculated t-ratio (.486) at 66 degree of freedom did not exceed the t-critical value (2.00). Therefore, we can be 95 per cent confident that there was also no significant

difference between the paper-delivery and computer-delivery reading exams for males.

***Table 4.7 The Result of Paired-Sample T-Test within Genders***

| Gender | *df* | *t* | *Sig.* |
|---|---|---|---|
| Females | 142 | 1.312 | .192 |
| Males | 66 | .486 | .628 |

These results are consistent with the findings of other studies (Blackhurst, 2005; Clariana & Wallace, 2002; Delaware, 2008) which stated that test takers' gender was not related to performance difference on computerized tests and paper-and-pencil tests. In other words, the effects of test-delivery-medium for gender were minimal and would have no meaningful impact on results.

## Completion Time on Reading Exams

In order to investigate the time effect on the paper-delivery and computer-delivery reading exams, participants' time for completing reading exams were calculated. Of the 210 participants, 69 participants completed the paper-delivery reading test faster than computer-delivery reading test; 103 participants completed the computer-delivery reading test faster than paper-delivery reading test; 38 participants

had no difference on paper-delivery and computer-delivery reading tests. Detailed information about the comparison of completion time on paper and computer is presented in table 4.8.

***Table 4.8 Result of Completion Time***

| Mediums | *M* | *SD* |
|---|---|---|
| Computer | 23.45 | 5.45 |
| Paper | 25.56 | 5.52 |

The result showed that the time to complete a computer-delivery reading exam was shorter than the time to complete a paper-delivery reading exam. This finding was consistent with some comparability studies which claimed that the time to complete computer-delivery tests was shorter than the time to complete paper-delivery tests (Bodmann & Robinson's study, 2004; Wang, 2008).

A possible explanation of why students took more time completing the paper-delivery exam might be that students could go back to review notes or underling parts on a paper-delivery exam to check answers. However, for a computer-delivery exam, since students could not write notes or underline key points of the text on the computer screen they may not have spent time reviewing to check their answers.

# CHAPTER 5

# CONCLUSION

This chapter presents the summary and the major findings of the study. In addition, the implications for English teaching and learning are discussed. Finally, limitations of the study and suggestions for further research are included.

## Summary of the Study and the Major Findings

This comparability study was conducted to investigate and compare paper-delivery and computer-delivery exams of reading. The purpose of this study was three-fold: 1) to analyze EFL university freshmen's reading performance in terms of the total scores of computer-delivery reading exams and paper-delivery reading exams, 2) to examine the effect of gender on the test-delivery-mediums, and 3) to compare the completion time between the computer-delivery reading exam and the paper-delivery reading exam.

The data were collected from Reading Exam A and Reading Exam B through both paper-delivery and computer-delivery platforms. Participants were divided into two equivalent groups and each group had to separately take one reading exam through the computer-delivery platform and another reading exam through the

paper-delivery platform one week later. Independent-sample t-tests were utilized to analyze the mean scores on the reading exams for both paper-delivery platform and the computer-delivery platform between groups, while paired-sample t-tests were utilized to analyze the mean scores on the reading exams for both paper-delivery platform and the computer-delivery platform within groups. In addition, scores in terms of participants' gender and completion time of the paper-delivery and computer-delivery platforms were analyzed. Paired-sample t-tests were used to determine if there were any significant differences between paper-delivery reading exam and computer-delivery reading exam for both females and males. As for completion time, participants' completion time was calculated by subtracting the starting time for the ending time.

The major findings of this comparability study are summarized as follows.

According to students' scores on the Reading Exams, it was found that the mean score on computer-delivery reading exams was slightly higher than that of paper-delivery reading exams but t-tests revealed no significant difference between paper-delivery and computer-delivery platforms for either Reading Exam A or Reading Exam B.

The data from Reading Exam A and Reading Exam B through both

paper-delivery and computer-delivery platforms were also analyzed by gender. The results revealed that there was no significant difference by gender.

In terms of the completion time, the time to complete a computer-delivery reading exam was shorter than the time to complete a paper-delivery reading exam.

To conclude, this study showed that there were no significant differences between Taiwan's university freshmen's scores on computer-delivery reading exams and paper-delivery reading exams. Moreover, students completed the computer-delivery reading exam faster than the paper-delivery reading exam.

## Implications of the Study

As mentioned earlier in this study, computers have become increasingly prevalent and important in education in recent years. Many international testing services and educational programs have initiated computerized language tests in their programs. Since the scores of this kind of computerized test are being used by different testing organizations to determine test takers' ability on a particular area, the results of this study can benefit testing services that use computers to administer a language test.

According to the finding of this study, it appears no significant difference

between paper-delivery reading exams and computer-delivery reading exams. The reason could be that students today are much more familiar with computers than students in the past. Therefore, students' performance will not be significantly affected by different test-delivery-mediums.

The results of the study also encourage teachers to consider adapting tests to a computer platform. Compare to the traditional paper-delivery platform, a computer-delivery platform may also be a new and good choice for language testing.

## Limitations of the Study and Suggestions for Further Research

The findings of this study did yield some useful information on current university freshmen's reading performance by different test-delivery-mediums. However, there were some limitations of this study.

First, the focus of this study was solely on the quantitative data. In other words, this study only analyzed students' scores on the computer-delivery reading exams and paper-delivery reading exams. Therefore, further research could be conducted on qualitative aspects, such as computer familiarity, computer anxiety, the attitude toward using computers, etc.

In addition, the completion time of computer-delivery reading exams was significantly shorter than paper-delivery reading exams. The reasons behind this

phenomenon need to be explored further. For example, a questionnaire about the attitude toward using computers and paper is needed.

Finally, the participants in this study were freshmen university students about 18 years old. It is suggested that further studies be conducted with participants from several different age levels. This may give rise to different results in terms of reading performance.

## APPENDIX A

## Reading Exam A & Reading Exam B

### Reading Exam A

**Direction: Read the following passages and choose the BEST answer for each question. Please write down the answers on the answer sheet.**

**PASSAGE 1**

1. Choosing a husband or wife is one of the most important decisions in a person's life. A good marriage can mean the difference between a happy and unhappy life. In many cultures, young men and women choose their own **spouses**. In other cultures, choosing the right marriage partner is so important that parents arrange their children's marriages. Such arranged marriages have both good and bad points.

2. One good point to an arranged marriage is **financial security**. Of course, money doesn't always bring happiness, but a lack of money certainly causes problems in any relationship. A second good point of an arranged marriage is that parents may make a better choice than their children. They are older and wiser. Also, parents have been married; therefore, they know better what qualities are necessary in a spouse. Furthermore, parents may be better judges of character than young people. Young people often let **emotions** influence their judgment. To sum up, arranged marriages may be happy because parents chose with their heads, not with their hearts.

3. On the other hand, arranged marriages may have some bad points. One disadvantage is that parents may make a poor choice. As a result, the young couple may never be happy together. A second disadvantage is that a young man or woman may already have fallen in love with someone else. If the parents force him or her to marry their choice, the result may be three unhappy people.

4. In conclusion, there are advantages and disadvantages to both kinds of marriages. Arranged marriages and free-choice marriages both take work, patience, and perhaps a little bit of luck.

1. What does **spouses** mean in paragraph 1?

(A) husbands or wives (B) young men or women
(C) many cultures (D) good marriages

2. What does **financial security** mean in paragraph 2?
(A) happiness (B) money (C) marriage (D) relationship

3. What does **emotions** mean in paragraph 2?
(A) marriage (B) money (C) hearts (D) heads

4. What is the main idea of paragraph 1?
(A) Young people should choose their own marriage partner.
(B) Arranged marriages have become popular again.
(C) Happy marriages are arranged marriages.
(D) Marrying the right person is very important in life.

5. What is the main idea of paragraph 2?
(A) Because parents can find rich spouses, children should listen to their parents.
(B) Because young people are blinded by love, they should not marry someone they love.
(C) Because parents are more experienced, they often make better choices for their children.
(D) Because young people do not have money, they often follow their parents' choice.

6. What is the main idea of paragraph 3?
(A) Arranged marriages may cause serious unhappiness.
(B) Arranged marriages make young people unhappy.
(C) Parents should be careful about who their children marry.
(D) Young people should marry according to whom they love.

7. According to the passage, which of the following is a reason arranged marriages work?
(A) Parents make a logical choice.
(B) Parents know who their children love.
(C) It takes a lot of thinking.
(D) It will not make young people unhappy.

8. According to the passage, which of the following is a reason arranged marriages do **NOT** work?
(A) Parents make a wrong choice.
(B) Matchmakers tell lies to the young couple.
(C) Young people only use their hearts, not their heads.

(D) Parents do not help their children with money.

9. According to the passage, which of the following is true?
   (A) Arranged marriages are good for young people all over the world.
   (B) Financial security causes a lot of problems in arranged marriages.
   (C) Young people should follow their parents' orders because parents are wise.
   (D) People may be unhappy because they have chosen the wrong spouse.

10. What is the main idea of the passage?
   (A) Cultures with arranged marriages are better than those without this tradition.
   (B) There are good and bad points to both arranged and free-choice marriage.
   (C) Experienced parents are usually wiser than their inexperienced children.
   (D) Parents whose own marriage was arranged are sorry they had children.

**PASSAGE 2**

The story is told of an Amish man who was taking a long train trip. Following Amish tradition, he had left school at an early age to work in the fields. He began talking to his seatmate, who he learned was a university professor. The Amish fellow proposed a game. He said, "Since you are so well educated, how about I ask you a question, and if you can't answer it you give me a dollar. And since I have only a sixth-grade education, you ask me a question, and if I can't answer it, I give you fifty cents?" The professor agreed, and the Amish man asked his question. "If it takes an elephant three days to climb Mount Rushmore, how many potato peelings would it take to shingle a doghouse?" The professor pulled out a dollar and handed it to him. "You've got me. What's the answer?" The Amish man said, "I don't know either. Here's your fifty cents."

11. We can infer that the Amish man's question
   (A) actually had no answer.
   (B) was based on actual experience.
   (C) could have easily been answered.

12. The story implies that
   (A) the Amish man was not as intelligent as the professor.
   (B) the university professor was rude.
   (C) being educated is not the same as being smart.

**PASSAGE 3**

1. Man was trying to cure his illnesses long before anything was known about what caused them to be sick. Students of ancient history have learned about some of the methods those men used. In Iraq, for example, a list of substances, which were used 4000 years ago by the people of Babylon to make medicine, was found printed on a brick of sun-baked clay.

2. In those early times, people went to priests for help with their illnesses and thus priests soon observed that certain illnesses had certain definite causes. They easily saw the connection between too much eating and drinking at an evening feast, and a stomachache the next morning. They might put the patient on a diet, but called it a punishment for having offended a god rather than a simple step to restore good health.

3. Priests also came to recognize that some of the medicines they made from certain plants would always produce useful results. For example, seeds of the poppy plant **dulled** feelings and the patient would no longer be in intense pain.

4. Some 2500 years ago a new kind of doctor appeared, one who did his best to take medicine away from the priests. His name was Hippocrates and he was born on Cos, an island of Greece, about 460 years before the first year of our present calendar.

5. Hippocrates was the first **healer** to state that diseases resulted from natural causes, not from the actions of gods. He was the first to make his decisions on what was wrong with a patient by observing the way the patient was affected. He was also the first to distinguish between one kind of illness and another by noticing differences in the effects on the patient.

13. Which of the following is the best title for the entire passage?
    (A) Babylonian Medicine
    (B) Hippocrates
    (C) The Early Development of Medicine
    (D) Priests and Patients

14. What is the main idea of paragraph 2?
    (A) Ancient understanding of the cause of illness
    (B) Ancient punishments
    (C) The rules for good health
    (D) Ancient diet

15. What does **dulled** mean in paragraph 3?

(A) did not effect (B) increased (C) moved (D) decreased

16. What does **healer** mean in paragraph 5?
(A) A person who helps sick people (B) A person who works with religion
(C) A person who fixes cars (D) A person who eats a lot

17. Hippocrates thought that illness was brought about by:
(A) the actions of gods (B) priests
(C) other Babylonians (D) natural causes

18. Why did Hippocrates want to take medicine away from the priests?
(A) He thought priests were too expensive.
(B) He knew he had a more scientific way to treat disease.
(C) He hated priests.
(D) The priests were unable to cure him.

**PASSAGE 4**

1. One of the most powerful killers of bacteria and viruses was discovered quite by accident in the fall of 1928. At that time, in his basement laboratory in London, a bacteriologist, Dr. Alexander Fleming, was looking for a substance that would kill deadly bacteria. In order to observe their growth, he had spread on his laboratory desk some small plates containing the bacteria. One evening he accidentally failed to place a cover on one of the plates. This accident started his greatest discovery.

2. When Fleming arrived the next morning, he saw that the plate had gathered mold during the night. This did not surprise him, for the basement was damp and ventilated only by a partly opened window. But what he saw next did surprise him. Around the outside of the uncovered plate, the bacteria were still **flourishing**, while in the area close to the mold there was none. They had somehow disappeared. He transferred the mold, which he named penicillin, to a clean plate and let it multiply for two weeks. Then he began to experiment with the penicillin and found it would destroy bacteria in a test tube. Would it, he wondered, do the same to bacteria in the human body?

3. In 1929, Fleming wrote a report of his laboratory experiments, presented it at a medical meeting, and had it printed in scientific journals. But for ten years, while he continued to experiment with penicillin, his important news was largely ignored by the scientific world.

19. Which phrase best expresses the main idea of the whole passage?

(A) Bacteria and disease (B) Dr. Alexander Fleming

(C) The discovery of penicillin (D) Accidents

20. Which phrase best expresses the main idea of paragraph 1?

(A) The accident which began the discovery.

(B) Fleming's experiments.

(C) Killing bacteria.

(D) The use of penicillin.

21. What does **flourishing** mean in paragraph 2?

(A) growing (B) dying (C) fading (D) missing

22. What surprised Fleming?

(A) The window was open.

(B) There were no bacteria near the mold.

(C) There was mold on the plate.

(D) The basement was damp.

23. What was the result of Fleming's later experiment?

(A) Penicillin disappeared after two weeks.

(B) Penicillin failed to destroy bacteria in the human body.

(C) Penicillin was destroyed by bacteria.

(D) Penicillin destroyed bacteria in a test tube.

24. What was the immediate effect of Fleming's experiments?

(A) They were proved useless.

(B) They were not reported.

(C) They were largely ignored.

(D) Penicillin was immediately produced.

## Reading Exam B

**Direction: Read the following passages and choose the BEST answer for each question. Please write down the answers on the answer sheet.**

### PASSAGE 1

1. The best-known piece of clothing in India and Pakistan is the "sari," the dress most women wear. A sari is made from a long piece of cotton or silk, and comes in an endless variety of colors. The most expensive saris are those with designs of gold and silver and beautiful patterns. Less **costly** saris are made of rough cloth and have no designs.

2. Women of wealthy families have dozens of saris made out of expensive cloth decorated with attractive designs. Most women, however, have only one or two good saris and these are saved for special holidays and are worn only a few times a year.

3. The sari may be a tradition outfit for women in India and Pakistan but there is a special way to put on a sari. The sari extends to the ground like an evening dress and is five or six meters long and about one meter wide. It is first wrapped around the waist, then brought up under one arm, then passed over the opposite shoulder. The end of the cloth may also be used as a shawl to cover the head. Women wear a blouse with a sari.

4. Wrapping a sari over the body is a special art. There are dozens of different ways in which a woman can arrange the cloth to fit her mood and the occasion. Often, saris **reveal** the area a woman comes from; for example, their colors may show where she was born.

5. Many women wear another special kind of clothing. This is a two-piece outfit consisting of tight cotton pants and an upper garment which come down to the knees. A silk scarf is usually thrown back over the shoulders. Women wear this when they want more freedom of movement than the sari offers.

6. These various outfits can be quite smart, but most women in India and Pakistan usually dress much more simply. Their everyday clothes consist of inexpensive long skirts and blouses made of cotton. The colors are usually **faded** after hundreds of washings. They are not as bright as before.

1. What does **costly** mean in paragraph 1?
   (A) warm (B) expensive (C) strong (D) known

2. What does **reveal** mean in paragraph 4?
   (A) show (B) relate (C) find (D) see

3. What does **faded** mean in paragraph 6?
   (A) brighter (B) less colorful (C) unchanged (D) more beautiful

4. What is the main idea of paragraph 3?
   (A) Women wear a blouse with a sari.
   (B) Saris look like evening gowns.
   (C) Saris are all the same length.
   (D) There are certain steps in wearing a sari.

5. What is the main idea of paragraph 4?
   (A) A sari tells about a woman's feelings and background.
   (B) Women wear saris because they are happy.
   (C) Women wear special colors only on their birthday.
   (D) Putting on a sari takes a lot of time.

6. What is the main idea of paragraph 5?
   (A) Women don't always wear saris because they think saris look bad.
   (B) When women want to be active they wear a two-piece outfit.
   (C) Most women like to wear a two-piece outfit more than a sari.
   (D) The two-piece outfit is considered more beautiful than the sari.

7. According to the passage, which of the following statements is true?
   (A) Saris are popular with women throughout the world.
   (B) Most women like to dress fashionably.
   (C) Saris are worn only by wealthy women.
   (D) Women in India and Pakistan have traditional clothing.

8. According to the passage, what does a sari look like?
   (A) It is a long piece of cloth.
   (B) It is sometimes decorated with silver and gold.
   (C) It can be either colorful or plain.
   (D) All of the above.

9. According to the passage, which of the following statements is **NOT** true?
   (A) There is only one way to arrange a sari.
   (B) Part of the sari may be used to cover the head.

(C) Saris are worn during holidays.
(D) Saris come in many different colors.

10. What is the best title for the passage?
(A) How Saris Are Made
(B) India and Pakistan -- Worlds Apart
(C) Popular Female Clothing in India and Pakistan
(D) Smart people wear a sari

**PASSAGE 2**

What makes a good opening line? Sociologists have actually observed and studied how people react to typical "pickup lines." Honest and sincere opening lines were the most successful. For example, in a bar, the most preferred opening line is "Do you want to dance with me?" The least preferred is "Bet I can out drink you." In a restaurant, you'll have the most success with a sincere question like "I haven't been here before. What's good on the menu?" But stay away from "cute" remarks like "I bet the strawberry shortcake isn't as sweet as you are." In the supermarket, you might try "Can I help you to the car with that?" Avoid asking "Do you really eat that junk?" And in the laundry try "Want to go get a cup of coffee while you're waiting?" But never remark "Those are some nice underwear you've got there."

11. Based on the passage, which of the following opening lines is likely to be most successful?
(A) These flowers aren't as beautiful as you.
(B) Hi, my name is Jeff.
(C) You look like a girlfriend I dumped.

12. The implied main idea of the passage is that according to some sociologists,
(A) the cuter and more original a pickup line is, the better.
(B) simple, sincere-sounding pickup lines are most effective.
(C) pickup lines at the Laundromat are more likely to be effective than pickup lines at a bar.

**PASSAGE 3**

1. Since its beginnings, people have been interested in knowing how rock and roll music was first developed. We do not know specifically how or where **this** happened. But, we do know that rock music arose during the early 1950's from Man's need to communicate his thoughts, feelings and knowledge of new dance steps.

2. As time passed and people's ability to make electric sounds evolved, the complexity of the thoughts they were able to communicate with this form of music grew substantially. Later, the use of oral narratives and story-telling in song became widespread, enabling the rock and roll bands to pass on their **cultural values** from one music generating to another.

3. At the same time, rock and roll bands began to record their music on tapes and record albums in order to preserve their thoughts and sounds. It became customary to arrange the songs to create a much larger story or concept. At first, the concept grew out of the individual songs. Then, later on, the concept of the entire album came first and the songs were written to fit this original concept. Thus the concept album was born. The Beatles' Sergeant Pepper's is an example of this type of album.

4. Music is always changing, often as a result of contact with other cultures and beliefs. The English Punk Rock of today is very different from the English Rock and Roll of twenty years ago. However, whatever changes rock and roll music may undergo, it is still one of the most important possessions of our culture. It has made possible not only the expression and exchange of ideas and feelings, but also the transmission of various cultural attitudes and dance steps from one time to another.

13. Which of the following is the best title for the entire passage?
    (A) The Development of Rock and Roll
    (B) The Beatles and the Concept of Sergeant Pepper's
    (C) The Elements Used by Modern Rock and Roll Bands
    (D) It's Only Rock and Roll but I Like It.

14. What is the main idea of paragraph 3?
    (A) The Beatles Sergeant Pepper's is a concept album.
    (B) The development of written music took a long time.
    (C) The custom of making music is a part of daily life.
    (D) Rock music has developed from song collections to concept albums.

15. What does **cultural values** mean in paragraph 2?
    (A) dance steps (B) valuable possessions
    (C) ideas and feelings (D) electric sounds

16. What does **this** in paragraph 1 refer to?
    (A) The history of music.
    (B) People's interest in joining rock and roll bands.
    (C) The development of rock and roll.
    (D) The rise of the early 1950's man.

17. How did the development of electronic affect Man's ability to communicate through music?
    (A) Electronics enabled him to write longer songs.
    (B) Electronics enabled him to write more complicated songs.
    (C) Electronics gave the rock and roll bands more time to think.
    (D) Electronics made it difficult to understand the stories in the songs.

18. Why was storytelling important to rock and roll music?
    (A) It enabled values to be passed from one generation to the next.
    (B) People in the 1950's could not watch TV.
    (C) It encouraged people to listen to rock music.
    (D) Communication became greater and faster.

**PASSAGE 4**

1. The dolphin is a small, whale-mammal that is found in all the oceans of the world. Since the 1950's, scientists have been conducting communication experiments with the dolphin in a belief that man could someday communicate with the animal. Since vocalizing has been shown to be a consciously controlled activity of the dolphin, scientists believe that the dolphin has most of the prerequisites of speech. First, they can vocalize at will. Second, the noises they emit are many and varied; a dolphin's vocal **repertoire** includes clicks, quacks, wails, whistles, and supersonic singing noises. They are able to listen to their own voices and can make airborne sounds on request. The dolphin can adjust its pitch so that it is audible to human ears. Like children, they have the instinct and ability to mimic sounds that they hear. Finally, in what may or may not be a requirement of language, they have a brain of a size comparable to our own, and the size of the human brain has often been attributed to the development of speech.

2. It has been shown that the dolphin has twenty-seven sounds that concern feeding. If out of the twenty-seven dolphin noises one means, "herring--one kind of fish--," then scientists will have further proof that we are not the only language users on this planet.

19. Which phrase best expresses the main idea of the whole passage?
    (A) Dolphins live in the oceans of the world.
    (B) Dolphins display the ability to communicate.
    (C) Dolphins are able to make sounds, but not communicate.
    (D) Dolphins and humans are the only language users on this planet.

20. Which phrase best expresses the main idea of paragraph 1?
    (A) Dolphins are the equal of man.
    (B) Dolphins already have the ability to speak.
    (C) Dolphins demonstrate many of the skills involved in speaking.
    (D) Dolphins can make their wishes easily understood.

21. What does **repertoire** mean in paragraph 1?
    (A) range or a number of skills. (B) mouth.
    (C) requirements (D) movements.

22. Clicks, quacks, wails, whistles, and supersonic singing noises are examples of what?
    (A) sounds that all animals make.
    (B) sounds that sea creatures make.
    (C) sounds that humans and dolphins have in common.
    (D) sounds that dolphins can make.

23. According to the passage, which of the following is true?
    (A) Dolphins can not control their voices.
    (B) Dolphins can hear their own sounds.
    (C) Dolphins have tape-recorders in their brains.
    (D) Dolphins are unable to recognize different sounds.

24. Which of the following is stated in paragraph 2?
    (A) The dolphin may someday be able to speak human language.
    (B) The dolphin already knows how to say "herring" in dolphin language.
    (C) The dolphin may have some of the same linguistic capabilities as human beings.
    (D) We can understand what the dolphin tells us about fee

## APPENDIX B

### The Results of Item Analysis for Reading Exam A and B

| Item | Test | Item Difficulty | Upper 27% | Lower 27% | DS Indes |
|---|---|---|---|---|---|
| 1 | A | 77% | 83% | 67% | 17% |
| | B | 85% | 100% | 75% | 25% |
| 2 | A | 75% | 100% | 42% | 58% |
| | B | 73% | 92% | 50% | 42% |
| 3 | A | 69% | 100% | 25% | 75% |
| | B | 81% | 92% | 58% | 33% |
| 4 | A | 85% | 100% | 50% | 50% |
| | B | 56% | 75% | 33% | 42% |
| 5 | A | 89% | 100% | 67% | 33% |
| | B | 79% | 100% | 33% | 67% |
| 6 | A | 56% | 83% | 17% | 67% |
| | B | 77% | 100% | 50% | 50% |
| 7 | A | 10 %( 82%) | 8% | 0% | 8% |
| | B | 67% | 92% | 33% | 58% |
| 8 | A | 48% | 92% | 25% | 67% |
| | B | 46% | 83% | 33% | 50% |
| 9 | A | 56% | 92% | 33% | 58% |
| | B | 56% | 75% | 50% | 25% |
| 10 | A | 81% | 100% | 67% | 33% |
| | B | 58% | 92% | 8% | 83% |
| 11 | A | 85% | 100% | 47% | 53% |
| | B | 17 %( 79%) | 60% | 0% | 60% |
| 12 | A | 88% | 100% | 60% | 40% |
| | B | 69% | 87% | 0% | 87% |
| 13 | A | 71% | 100% | 27% | 73% |
| | B | 86% | 95% | 68% | 27% |
| 14 | A | 67% | 100% | 36% | 64% |
| | B | 48% | 82% | 23% | 59% |
| 15 | A | 72% | 100% | 55% | 45% |
| | B | 61% | 86% | 14% | 73% |

| | | | | | |
|---|---|---|---|---|---|
| 16 | A | 82% | 100% | 50% | 50% |
| | B | 82% | 100% | 59% | 41% |
| 17 | A | 86% | 100% | 55% | 45% |
| | B | 62% | 82% | 27% | 55% |
| 18 | A | 82% | 100% | 45% | 55% |
| | B | 77% | 100% | 50% | 50% |
| 19 | A | 67% | 100% | 13% | 83% |
| | B | 90% | 100% | 63% | 38% |
| 20 | A | 50% | 100% | 0% | 100% |
| | B | 70% | 100% | 25% | 75% |
| 21 | A | 73% | 88% | 38% | 50% |
| | B | 63% | 100% | 38% | 63% |
| 22 | A | 77% | 100% | 75% | 25% |
| | B | 83% | 100% | 50% | 50% |
| 23 | A | 73% | 100% | 38% | 63% |
| | B | 73% | 100% | 38% | 63% |
| 24 | A | 70% | 100% | 50% | 50% |
| | B | 80% | 100% | 75% | 25% |

Note. 1. DS Indes refers to the discriminability.

2. The percentage within the round bracket is the revised one.

## REFERENCES


Alderson, J. C. (2000). Assessing reading. Cambridge: CUP.

Alderson, J.C., Clapham, C. & Wall, D. (2002). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bailey, K. (1998). *Learning about language assessment*. Boston: Heinle & Heinle.

Blackhurst, A. (2005). Listening, reading and writing on computer-based and paper-based versions of IELTS. *Research Notes, 21*, 14-17.

Bodmann, S.M. & Robinson, D.H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *J. Educational Computing Research, 31*(1), 51-60.

Brown, H. D. (2007). *Teaching by Principles: an interactive approach to language pedagogy*. White Plains, NY: Pearson Education.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59.

Buchanan T. (2000). The efficacy of a world-wide web mediated formative assessment. *Journal of Computer Assissted Learning, 16*, 193-200.

Bunderson, C.V., Inouye, D.K. & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp.367-407). London: Collier Macmillan.

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning and Technology. 5* (2), 95-98. Available: http://llt.msu.edu/vol5num2/deville/default.html

Choi, I.C., Kim, K.S. & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing. 20* (3), 295-320.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology. 33* (5), 593-602.

Cohen, A.D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

Conole, G. & Warburton, B. (2005). A review of computer-assisted assessment. *ALT Journal. 13* , 17-31

Davies, F.B.(1968). Research in comprehension in reading. *Reading Research Quarterly. 3* , 499-545.

Dunkel, P. (1997). Computer-adaptive testing of listening comprehension. *Modern*

*Language Journal. 75* (1), 64-73.

Godwin-Jones, B. (2001). Emerging technologies: Language testing tools and technologies. *Language Learning and Technology. 5* (2), 8-12. Available: http://llt.msu.edu/vol5num2/godwin/default.html

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly. 5*, 9-30.

Goodman, Y. M. & Burke C. L. (1972). *Reading Miscue Inventory Manual: Procedure for Diagnosis and remediation.* New York: Macmillan.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25 (3), 372-406

Grabe, W. (1999). Developments in reading research and their implications for computer- adaptive reading assessment. In M. Chalhoub-Deville (Ed.) *Issues in computer-adaptive testing of reading proficiency.* (pp. 11-47). Cambridge: UCLES.

Grabe, W., & Stoller, F. L. (2001). Teaching and researching reading. In C. N. Candlin, & D.R. Hall (Eds.) *Applied linguistics in action series.* Great Britain, Harlow: Pearson education.

Hackett, ED. (2005). The development of a computer-based version of PET. *Research Notes, 22*, 9-13.

He, Q., & Tymms, P. (2005). A computer-assisted test design and diagnosis system for use by classroom teachers. *Journal of Computer Assisted Learning, 21*, 419-429.

Hedayati, H. (2004). The comparability of a WBT and PBT of reading comprehension. M.A. thesis.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jacobsen, M. (2008). *Multiple choice item construction.* Available: http://www.ucalgary.ca/~dmjacobs/portage/

Jones, N. (2003). The role of technology in language testing. *Research Notes, 12*, 3-4.

Jones, N., & Maycock, L. (2007). The comparability of computer-based and paper-based tests: goals, approaches, and a review of research. *Research Notes, 27*, 11-14.

Kintsch, W. & Yarbrough, J.C. (1982). Role of rhetorical structures in text comprehension. *Education Psychology, 74* (6), 828-834.

McNamara, T. (2000). *Language esting.* Oxford: OUP.

Nunan, D. (1999). *Second language teaching and learning.* Boston, MA: Heinle & Heinle.

Phakiti, A. (2006). Theoretical and pedagogical issues in ESL/EFL teaching of

strategic reading. *TESOL, 1,* 19-50.

Rivers, W.M. & Temperly, M.S. (1978).*A practicl guide to the teaching of English as a second or foreign language.* New York: OUP.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5*(2), 84-89.

Russell, M. (2002). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Available: http://epaa.asu.edu/epaa/v7n20/

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back and into the future. Technology and Assessment Study Collaborative. Available: http://escholarship.bc.edu/intasc/4

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language learning and technology. 5* (2), 38-59. Available: http://llt.msu.edu/vol5num2/sawaki/default.html

Sims, J. (2006). The creation of a valid and reliable university proficiency exam. *Tunghai Journal of Humanities, 47*, 325-344.

Taylor C., Jamieson, J., Eignor, D. and Kirsch, I. 1998: *The relationship between computer familiarity and performance on computer-based TOEFL test takers*. TOEFL Research Reports 61. March. Princeton, NJ: Educational Testing

Service.

Tsai, C. C. & Chou, C. (2002). Diagnosing students' alternative conceptions in science. *Journal of Computer Assisted Learning*, *18*, 157-165.

Tymms, P. B. (2001). The development of a computer-adaptive assessment in early years. *Educational and Child Psychology*, *18*, 20-30.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*, 5-24.

Wang, T. H., Wang H., Wang, W. L., Huang, S. C. & Chen, S. Y. (2004). Web-based assessment and test analyses (WATA) system: development and evaluation. *Journal of Computer Assisted Learning*, *20*, 59-71.

Wei, C.L. (2007). *A comprehensive guide to internet-based TESL/TEFL*. Taiwan: Kuan Tang Press.

Widdowson, H. G. (1984). *Reading and communication.* London: OUP.

Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. In J. Coady & T. Huckins (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 5-19). Cambridge: CUP.