CHAPTER ONE

INTRODUCTION

1.1 Background

English reading is an essential skill for those who use English as a second language (ESL) or as a foreign language (EFL); for many, reading is an important skill to master (Anderson, 1999; Brown, 2004). Furthermore, the ability to read English efficiently for academic purposes is widely recognized as an important skill in both ESL and EFL contexts (Grabe, 2007). This is particularly true in the educational system in Taiwan. According to Hsu (2008), in the Senior High Academic Ability Examination, a critical examination for senior high school graduates, reading is the dominant skill measured on the English portion. The subtests that assess reading ability comprise 70% of the English test. Besides, Gong (2008) points out that in order to enhance the quality of English education, many universities in Taiwan are trying to implement an English proficiency— including reading ability— threshold requirement or benchmark for university graduation. Furthermore, with the emergence of new technology, students absolutely need adequate English reading ability to extract information from the Internet. Therefore, it is necessary to provide students with assessment tools that can not only help them identify their weaknesses in reading in order to improve their English reading ability but also prepare them for future assessment, career development, and challenges in a digital society.

Meanwhile, with the advance of technology, computers have been widely integrated in assisting language learning as well as delivering language tests. This also allows for alternative types of assessment, such as diagnostic testing. A diagnostic (foreign) language test aims to identify a test-taker's specific linguistic strengths and weaknesses (Alderson, 2005; Bachman & Palmer, 1996). This can be contrasted with the purpose of traditional tests, which are often used to label a test-taker's general ability with reference to other test-takers in the normative group (Brown & Hudson, 2002). In addition, traditional tests rarely provide meaningful feedback to the test-taker or test-giver about particular aspects of language that need further improvement (Yin, 2006).

A defining feature of any diagnostic language test is the feedback it provides to the test-taker. In traditional assessment, providing feedback is equivalent to giving students their test scores after tests. However, in current research in assessment, many scholars (e.g., Wiggins, 1998, Shohamy, 2001) emphasize the importance of providing quality feedback to learners if assessment is to be used to improve performance, not just audit it. According to Wiggins (1998), feedback that is of high quality is that which is "highly specific, directly revealing or highly descriptive of what actually resulted, clear to the performer, and available or offered in terms of specific targets and standards" (p.46). Therefore, there is a great need for diagnostic tests which provide information for guiding learning, improving instruction and evaluating students' progress (Jang, 2009).

English reading comprehension tests have received considerable attention in Taiwan;

some studies on reading tests in Taiwan have focused mainly on aspects such as test-takers' reading strategy use (Chang, 2006; Hsieh, 2003, Hsu, 2003; Hsu,2008; Yang, 1997), while other studies have examined the analysis of reading comprehension tests (Lu, 2002; You, 2004). However, most of the reading tests aim to evaluate test-takers' overall reading ability without providing sufficient feedback to test-takers for future improvement. As a result, only limited research has been conducted on the diagnostic use of reading tests. Hence, there is a need to undertake further studies to investigate Taiwanese university students' perceptions about feedback from online diagnostic reading tests.

1.2 Statement of the Problem

Many researchers (Anderson, 1999, 2006; Carrell, 1993; Sims, 1996) have pointed out that reading is the most important of the four skills for learners of English as a foreign language (EFL) and English as a second language (ESL) who desire to achieve academic success in English. However, with the growing trend toward a more communicative approach in language teaching and learning, the results of many studies (e.g., Chen, 2006; Lin, 2001; Sims, 2004) have shown that Taiwanese students' English reading ability and grammar knowledge may have regressed while their listening ability seem to have shown significant progressed possibly as a result of changes in the language learning environment and education policy. Therefore, to help students improve their reading ability, it is important to have a tool that is able to diagnose students' strengths and weaknesses in reading. However, literature regarding this issue is thin.

With the integration of computer technology into language learning and teaching, more alternative types of assessments, such as diagnostic testing, have become possible. Though one of the advantages of Computer-Assisted Language Testing (CALT) is its provision of immediate feedback to the learner (test-taker), only a few language tests have been designed primarily for diagnostic purposes (Alderson, 2005; Jang, 2009) despite the fact that the importance of diagnostic language tests has gradually been recognized. Yin, Sims, and Cothran (forthcoming) have also pointed out that a relatively small number of diagnostic language tests has meant a correspondingly small amount of research about them, particularly regarding feedback.

A growing body of empirical research is now investigating the value and impact of feedback. Some studies have focused on feedback on writing (Bitchener, Young & Cameron, 2005; Ferris & Roberts, 2001, Hyland, 2003; Hyland & Hyland, 2006), and others have investigated feedback given to learners in CALL programs (Brandl, 1995; Heift, 2001, 2003, 2004; Pujola, 2001). Some aforementioned research results suggest that learners have individual preferences for certain feedback (Heift, 2001) and that feedback is handled differently by high and low language ability students (Brandl, 1995). However, there has not been research on this issue in relation to diagnostic reading test feedback. There is evidently a lack of research on how students of low and high English proficiency levels perceive diagnostic reading test feedback.

1.3 Purpose of the Study

The purpose of this study is to evaluate Taiwanese university students' perceptions of test feedback of the OEAS Reading Test – an online multiple-choice diagnostic test of English reading. The study attempts to answer both how much and in what ways the feedback is deemed useful; it also seeks to examine whether test-takers' English proficiency level is related to their perceptions of usefulness. Ultimately, the study aims to provide insights for test makers on how to develop and improve feedback on diagnostic reading tests.

1.4 Research Questions

This study seeks answers to the following three research questions:

1. How useful do test-takers perceive the Reading Test feedback to be?
2. In what ways do the test-takers perceive the feedback to be useful or not?
3. Is there a difference between test-takers of low and high English proficiency levels in how they perceive the feedback's usefulness?

1.5 Definition of Terms

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

1. Reading comprehension.

Reading comprehension can be viewed as an interaction between reader and text by which meaning is created (Anderson, 1999). In this study, a divisible view of reading comprehension is adopted (see 2.1.2). The six sub-skills measured in the OEAS Reading Test are: 1) reading for main ideas of a passage; 2) reading for main ideas of a paragraph; 3) reading for specific information; 4) guessing meaning of vocabulary from context; 5) pronoun reference; and 6) inference from reading.

2. Diagnostic (foreign) language assessment.

In language assessment, diagnostic language tests are defined as those that aim to identify learners' areas of strengths and weaknesses (Alderson et al, 1995; Bachman & Palmer, 1996; Moussavi, 2002) in order to help improve learning and teaching (see 2.3.1).

3. Feedback.

"Feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is in domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (Winne & Butler, 1994, p. 5740). In this study, feedback is conceptualized as test results and explanations for each question provided to the test-taker (see 2.3.4.3 for details about the OEAS Reading Test feedback).

4. Usefulness.

To evaluate the overall usefulness of any given test, Bachman and Palmer (1996) propose a framework including six aspects of test qualities – reliability, construct validity, authenticity, interactiveness, impact, and practicality. In this study, usefulness refers specifically to test-takers' perception of benefit to their English learning. In this study, test-takers' perceptions about the usefulness of test feedback is deemed an important factor in determining the usefulness of a diagnostic test.

5. Online English Assessment System (OEAS).

This term refers to a diagnostic language test battery at Tunghai University that is delivered on an online platform. The test battery includes a general English proficiency test, and skills tests of listening, reading, and grammar, respectively (see 2.3.4 for details).

1.6 Significance of the Study

Despite the growing demand in this area, development and implementation of diagnostic language tests is currently in its initial stages. Therefore, sufficient empirical data are needed for developing useful diagnostic language tests. One of the important factors to determine the usefulness of a test is test-takers' perceptions about its feedback.

This study aims to provide a better understanding of Taiwanese university students' perceptions about diagnostic feedback as well as their preferences towards different forms of reading test feedback. The findings of this study may reveal the importance of diagnostic test feedback and provide test-makers with a framework or recommendations for improving diagnostic reading tests so as to assist test-takers in discovering their own specific strengths and weaknesses in reading and help them improve their reading ability.

CHAPTER TWO

REVIEW OF THE LITERATURE

The present study focuses on Taiwanese university students' perceptions about the usefulness of diagnostic reading test feedback. The relevant literature is reviewed and organized into the following sections: 2.1) the reading construct, 2.2) computer-assisted language testing, 2.3) diagnostic language testing, 2.4) students' perceptions of diagnostic test feedback, and 2.5) summary and research gap.

2.1 The Reading Construct

2.1.1 Definition of Reading

The ability to read written language with good comprehension and a reasonable rate has long been recognized to be an important skill (Eskey, 1970; Carrell, 1993). Moreover, reading has been a highly emphasized skill in the EFL/ESL context. Among various definitions of reading, decoding, interpretation, and comprehension are the three most commonly used words. Thus, reading can be regarded as a process that involves decoding, interpreting, and comprehending the written materials.

Snow (2002) provides another well-articulated model of reading comprehension in the RAND Reading Study Group's report; she defines reading comprehension as "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (p. 11). Snow also proposed that reading comprehension includes three elements: "the reader who is doing the comprehending, the text that is to be comprehended, and the activity in which comprehension is embedded."

Bernhardt (1999) states that reading is the extraction and construction of a message from a written text. According to Urquhart and Weir (1998), reading is "the process of receiving and interpreting information encoded in language form via the written medium" (p. 22). Anderson (1999) states that a reader actively interacts with the reading material in the reading process. Thus, meaning does not only exist on the printed page but also in the head of the reader. That is, a reader interprets and constructs the meaning of a reading text by combining the words on the printed page and his background knowledge and experiences.

2.1.2 Processes of Reading Comprehension

Koda (2005) points out that reading, which is influenced by a range of variables, is a complex construct involving enormous component operations. In the process of reading, each single operation is dependent on a broad array of competencies. Due to its multifaceted nature, reading is such a complex mental process that it cannot be observed directly. In order to gain greater understanding of what reading comprehension is and how to measure it more appropriately, the following paragraphs review the related literature on various models and taxonomies of the reading process.

2.1.2.1 Models of the Reading Process: Bottom-up, Top-down, and Interactive

Understanding the process of reading has been the focus of much research. In the following, models of reading processes – bottom-up, top-down and interactive models – are briefly discussed.

The bottom-up model, also called the data-driven model, depends primarily on the information presented by the text (Anderson, 1999). Carver (1977) also described bottom-up reading as "a linear process from graphic symbols to meaning responses" (p. 27). Alderson (2000) further describes in detail that "the reader begins with the printed words, recognizes graphic stimuli, decodes them to sound, recognizes words, and decodes meaning" (p. 16).

In the late 1960s and 1970s, some researchers (Goodman, 1967; Smith, 1982) began to propose an alternative model called top-down processing. Top-down processing is information processing in which readers approach the text with existing knowledge, and work down to the text (Hudson, 1998). According to Stanovich (1980), reading can be seen as hypothesis testing because the reader actively engages with the written text. Goodman (1967, 1982) also calls the top-down model a psycholinguistic guessing game. This model recognizes the importance of readers' expectations of the contents of the text being processed (Urquhart & Weir, 1998). The readers draw the meaning from the text by making use of the previously acquired knowledge of the topic instead of merely focusing on letters, sounds, and words. They make predictions prior to reading the passage and test their predictions and adjust or confirm them in the reading process. As Alderson (2000) puts it, compared to the importance of meaning conveyed in the reading texts, the top-down model places much more emphasis on the importance of what readers themselves bring to the reading process.

Though the bottom-up and the top-down approach competed with each other throughout the 1970s and the 1980s, neither of these two approaches gave a satisfactory explanation of the process of reading. More recently, many theories of reading have put emphasis on the interaction between bottom-up and top-down processing (e.g. Johnston, 1984; Stanovich, 1980, 2000). Hence, a third type of reading process model has emerged: the interactive model, which regards reading as a complicated, interactive process that involves both bottom-up and top-down approaches (Carrell, 1989; Grabe, 1991; Urquhart & Weir, 1998). In this approach, the lower-level processing skills, such as word and letter-recognition processes, are just as important as the higher-level processing skills, such as the use of background knowledge and predicting. Stanovich (1980) states that in his interactive model "processes at any level can compensate for the deficiencies at any other level" (p. 36).

In addition, many researchers (Grabe, 2007; Koda, 2005; Nassaji, 2003) also agree that both lower- and higher-level processing skills work together in a complex, highly integrated set of processes while reading. Anderson (1999) points out that the interactive model is currently accepted as the most comprehensive description of the reading process.

2.1.2.2 The Reading Components: Unitary and Divisible

Apart from the three processing models mentioned above, there has long been considerable disagreement about whether reading comprehension should be viewed as a unitary process or a divisible process with separate component skills. Some scholars (Lunzer et al., 1979; Carver, 1992; Rost, 1993) view reading comprehension as a holistic process and reading only consists of a single global construct. This unitary view regards reading as a single undifferentiated ability. For example, Rost (1993) claims reading comprehension should be considered a unitary process because the subskills of reading comprehension are closely fused with each other in the process of reading. Therefore, it becomes nearly impossible to distinguish reading subskills, let alone to measure them separately. Rost also concludes that the variance in reading could be attributed to a single dimension “general reading comprehension.”

On the other hand, a number of researchers support the idea that reading comprehension is multi-divisible and can be divided into two or more components. These researchers claim that several factors, such as vocabulary (Carver, 1992; Hudson, 1996; Urquhart & Weir, 1998) and inference (Enright et al., 2000; Long et al., 1996; Munby, 1978), which influence readers’ ability to successfully read and comprehend written text, can be identified and separated. Porter and Weir (1994) argue that there are different skill components in reading itself just as the common consensus holds that there exist systematic differences between the skills of listening, reading, writing, and speaking.

A number of researchers (Grabe, 1991; Lumley, 1993; Munby, 1978; Weir et al., 2000, Koda, 2005) agree that reading is composed of multi-componential skills but differ on the number and scope, as Table 2.1 shows. Views of a few more prominent scholars are discussed below.

Munby (1978) presents a taxonomy of 54 language skills. Carroll (1980) considers Munby’s taxonomy as “one of the most fruitful sources of information for test construction” (p. 32). Among this taxonomy, Munby specifies an extensive list of nineteen reading micro-skills such as understanding conceptual meaning, distinguishing the main ideas from supporting details, skimming, basic reference skills, identifying the main point of information in discourse, understanding relations between parts of texts through lexical cohesion devices, and extracting salient points to summarize.

Grabe (1991) proposes a more concise list of six distinguishable components in the reading process: automatic recognition skills, syntactic knowledge, knowledge of formal discourse structure, content and background knowledge, synthesis and evaluation skills/strategies, and metacognitive knowledge and skill monitoring.

Additionally, Urquhart and Weir (1998) propose that reading process and skills can be conceptualized into four broad categories: (a) expeditious reading at the global level: skimming for the gist and searching for information, (b) expeditious reading at the local level: scanning for specific information through word-matching strategies, (c) careful reading at the global level: understanding explicitly stated main ideas, inferring propositional meanings and pragmatic meanings, and (d) careful reading at the local level: inferring lexical meanings and understanding syntax.

Table 2.1

Definitions of reading comprehension and components of reading (arranged chronologically)

Scholars	Definition of Reading Comprehension/ Components of Reading
Davis (1968)	Comprehension among mature readers is not a unitary mental operation. Five factors contribute to successful comprehension: 1) recalling word meanings 2) drawing inferences 3) recognizing a writer's purpose/ attitude/ tone 4) finding answers to explicit questions 5) following the structure of a passage
Munby (1978)	Reading consists of 19 skill-components, e.g.: 1) recognizing the script of a language 2) understanding conceptual meaning 3) distinguishing the main ideas from supporting details 4) scanning to locate specifically required information 5) skimming 6) deducting the meaning and use of unfamiliar lexical items 7) basic reference skills 8) understanding explicitly stated information 9) understanding information when not explicitly stated 10) extracting salient points to summarize….. etc.
Berkoff (1979)	Reading is bidivisible and is composed of two distinct factors: 1) recognition vocabulary 2) prediction and self correction
Clay (1979)	Reading involves different skills: 1) recognition vocabulary 2) prediction and self correction 3) inference 4) auditory memory 5) search for cues in text 6) know probabilities of occurrence….etc.
Coady (1979)	Reading comprehension involves the interaction of the reader's 1) conceptual abilities, 2) background knowledge, and 3) process strategies.

Table 2.1 (continued)

Scholars	Definition of Reading Comprehension/ Components of Reading
Alderson (1984)	Reading is seen as a selective process taking place between the reader and the text, in which background knowledge and various types of language knowledge interact with information in the text to contribute to text comprehension.
Greenall and Swan (1986)	Reading is a continuum that can be split into different small skills, e.g.: 1) dealing with unfamiliar words 2) extracting main ideas 3) reading fro specific information 4) understanding text organization 5) checking comprehension 6) evaluating the text and reacting to a text….etc.
Bernhardt (1991)	Reading includes various factors interacting with one another: 1) phonemic/ graphemic features 2) word recognition 3) syntactic feature recognition 4) intratextual perceptions 5) prior knowledge 6) metacognition
Grabe (1991)	Reading process consists of six components: 1) automatic recognition skills 2) vocabulary and structural knowledge 3) formal discourse structure knowledge 4) content/world background knowledge, 5) synthesis and evaluation skills/strategies 6) metacognitive knowledge and skills monitoring.
Hoover and Tunmer (1993)	Two main components contribute to reading comprehension: 1) word recognition 2) linguistic comprehension

Table 2.1 (continued)

Scholars	Definition of Reading Comprehension/ Components of Reading
Hudson (1996)	Reading involves "the interaction of a vast array of processes, knowledges, and abilities" (p. 3). Reading-processing skills cover a vast array of overlapping abilities from local textual comprehension to global text interpretations and inferencing. Following are some identified reading skills: 1) automaticity in word and sentence recognition 2) context and schema (i.e., formal, content) 3) strategies and metacognitive skills 4) reading purpose and context
Urquhart and Weir (1998)	Reading process and skills are conceptualized into four broad categories: 1) skimming for the gist and searching for information 2) scanning for specific information through word-matching strategies 3) understanding explicitly stated main ideas, inferring propositional meanings and pragmatic meanings 4) inferring lexical meanings and understanding syntax.
Koda (2005)	Successful comprehension emerges from the integrative interaction of derived text information and preexisting reader knowledge. Seven key components account for reading comprehension: 1) decoding 2) vocabulary knowledge 3) syntactic processing 4) text-structure knowledge 5) main-idea detection 6) background knowledge 7) comprehension strategies

More recently, Koda (2005) provides a comprehensive summary of various models of reading comprehension. She lists out seven key components for reading comprehension: decoding, vocabulary knowledge, syntactic processing, text-structure knowledge, main-idea detection, background knowledge, and comprehension strategies (p.254-262).

Even though the exact nature of reading comprehension remains controversial, many researchers in the field of language testing support the idea of viewing reading comprehension in terms of separate components. Grabe (1991) concludes that "a reading components perspective is an appropriate research direction because it leads to important insights into the reading process" (p.382). Moreover, according to Alderson (2000), implementing a unitary approach in the testing of reading may not fully represent the reading comprehension construct for it may not appropriately test all the relevant reading skills. In addition, Weir (2005) argues that no matter what theoretical position one takes, one inevitably measures certain reading skills upon writing individual items based on reading a passage. Thus, compared to the unitary view, the divisible view of reading may be more suitable for testing of reading comprehension, and it is the view adopted in this study.

2.2 Computer-Assisted Language Testing

Chapelle (2001) defines computer-assisted assessment as "testing practices requiring a computer to assist in construction, delivery, response analysis and score reporting" (p. 38). Over the past few decades, computer-assisted assessment has been developed rapidly in terms of its integration into educational settings. Many researchers (Alderson, 2000; Brown, 1997; Chalhoub-Deville, 2001; Chapelle & Douglas, 2006) agree that with the increasing accessibility and development of technology, the application of computers has had a great influence on language assessment as well as on other fields of applied linguistics.

Remarkable technological advances in the past few decades have led to a new era in computer-assisted language tests. Some researchers (Brown, 1997; Chapelle & Douglas, 2006; Dunkel, 1999; Tung, 1986) have pointed out the advantages of Computer-Assisted Language Testing (CALT). First, CALT enables the integration of multimedia into tests and the delivery of a wide variety of test tasks on-line all over world. Therefore, CALT also allows individual, time-independent language testing that can be taken at many convenient locations and times. Second, CALT offers more accurate scoring and its function of record keeping make it possible for assessing students' learning more systematically. Third, CALT provides immediate feedback, which consequently offers extra value to the user. Proponents argue that with the implementation of CALT, teachers and learners may benefit by having access to assessment which "may offer possibilities for response analysis, feedback, and record keeping beyond what is feasible with traditional assessments" (Chapelle and Douglas, 2006, p.3).

Diagnostic language testing is one kind of testing that has been facilitated by CALT. In line with the advantages mentioned above, Alderson (2005) points out that computer-based testing is particularly suitable for diagnostic testing. Moreover, He and Tymms (2005) declare that providing timely and specific information on the performance of each student is one of the most valuable benefits of using computer-assisted assessment in education. In addition to diagnosing areas where students have individual difficulties, the information can also be beneficial for guiding future instruction and learning. This claim appropriately describes the purpose of diagnostic language testing discussed in the following section.

2.3 Diagnostic Language Testing

2.3.1 Introduction to Diagnostic Language Tests

In language assessment, diagnostic language tests aim to identify learners' areas of strengths and weaknesses (Alderson et al., 1995; Bachman & Palmer, 1996; Moussavi, 2002) in order to help improve learning. In fact, Alderson (2005) claims that diagnostic test is the type of test that is "closest to being central to learning" (p. 4). However, he also points out that there is apparently a lack of clear theoretical basis for diagnosis in second language testing and thus further research is needed in this under-investigated field.

Traditional tests such as proficiency and achievement testing have been criticized for they only demonstrate learner's partial knowledge and barely represent learning processes (Brown & Hudson, 1998; Popham, 1999). Not being satisfied with proficiency tests, researchers in the field of language testing have sought ways for improvement. Some researchers (Alderson, 2005; Bailey, 1996; Kunnan & Jang, 2009, Shohamy, 1992) have suggested that assessment results should comprise more descriptive information and detailed score reporting so that the results can be used to help teachers in course design and in turn facilitate students' learning. Though being convenient and efficient, standardized tests are increasingly recognized to be insufficient in accurately measuring students' achievement and progress in specified domains, guiding learning, and designing instruction (Bejar, 1984; Brindley, 2008; Brown & Hudson, 1998). As there is a growing body of research into the impact of testing (Bailey, 1996, Messick, 1996; Shohamy, 2001; Wall, 2000), the testing community has pointed out the need for more diagnostic information that allows for meaningful interpretations of test results. Also, the proper use of test results may ultimately lead to improvement in instructional design and enhancement of students' learning (Kunnan & Jang, 2009). The following sections provide a brief introduction of various types of tests, as Table 2.2 shows, so as to help distinguish diagnostic tests from traditional tests.

According to Brown (2004), a proficiency test aims to test global competence in a language, and it is not limited to any one course, curriculum, or single skill in the language. Proficiency tests are summative and norm-referenced. They provide results with equated scores and percentile ranks taking on paramount importance; therefore, proficiency tests are usually not equipped to provide diagnostic feedback.

Davis et al. (1999) defines achievement as how well a learner masters the materials covered in the textbook or syllabus. Therefore, the content of achievement tests is mainly based on books or materials used in a curriculum within a particular time frame (Brown, 2004; Hughes, 2003). In contrast to proficiency tests, achievement tests are directly related to a course or even a total curriculum. Achievement tests primarily aim to assess the extent to which students have achieved the course objectives. Brown (2004) further points out the difference between an achievement test and a diagnostic test: “achievement tests analyze the extent to which students have acquired language features that have *already* been taught; diagnostic tests should elicit information on what students need to work on in the future” (p. 47, italics in original).

Although Bachman (1990) states that “virtually any test has some potential of providing diagnostic information” (p.60), Bejar (1984) differentiates a diagnostic test from other types of assessment by specifically pointing out that a diagnostic test is self-referencing. With achievement and proficiency tests, for instance, a student’s performance is compared to that of other students, while “in a diagnostic test the student’s performance is compared against his or her expected performance” (Bejar, 1984, p. 176). Furthermore, Cotos and Pender (2008) also mention that in addition to solely presenting immediate results, a diagnostic test should provide students with explicit feedback. With its detailed analysis of learner responses, a diagnostic test may result in improvement in learning as well as remediation in instruction.

Table 2.2

Comparison of diagnostic tests and traditional tests (adopted from Yin, 2006)

Diagnostic tests	Proficiency/ Placement tests
● Give meaningful feedback to test-takers ● Point out specific structures or sub-skills that require improvement (e.g., relative clauses, word order, listening for specific details, inference) ● Useful for students, may be useful for test administrators ● Self-referenced	● Often label test-taker with a score only ● May provide a general level of ability (e.g., low, intermediate, advanced) or point out a general area of strength or weakness (e.g., reading, writing ability) ● Useful for test administrators but not so useful for students ● Norm-referenced and summative
	Achievement tests
♦ Can complement a language course but can occur independently of one ♦ Coverage can be wider and more representative of the construct ♦ Can elicit information on what students need to work on in the future ♦ Can be given at any time (online) ♦ Well-suited for self-assessment	♦ Usually embedded in a language course ♦ Coverage is limited to course content or curriculum ♦ Determine whether students have acquired language features that have already been taught ♦ Given during and/or at end of course ♦ Students need to be enrolled in a course in order to be tested

It is evident that the importance of diagnostic testing has been identified by many researchers. Though there is a great need for research into diagnostic testing and its score reporting processes, few empirical studies have been conducted to investigate the relevant issues. As Alderson (2005) suggests, there is a remarkable lack of valid diagnostic tests or any tests that claim explicitly to be diagnostic of foreign language proficiency. Some researchers (Jang, 2009; Yin et al., forthcoming) also points out that there are so few language tests designed primarily for diagnosis purposes.

Therefore, further research is needed to develop diagnostic testing that is suited for diagnosing learners' strengths and weaknesses in the tested skills. More importantly, the feedback provided in diagnostic tests should provide positive effects on learner's future learning.

2.3.2 Diagnostic Feedback on Language Tests

Feedback has long been considered important in both encouraging and consolidating learning in educational contexts (Brandl, 1995; Hyland & Hyland, 2006). Furthermore, the significance of feedback has also been recognized by researchers in the field of second language acquisition (SLA) and assessment (e.g., Brown & Hudson, 1998; Heift, 2001, 2003; Kunnan & Jang, 2009). Many researchers (Black & Wiliam, 1998; Higgins, Hartley, & Skelton, 2002) also indicate that feedback is influential in student achievement. Hyland (2000) also points out that there is great need for feedback on how students can improve their future performance in addition to merely helping students identify their strengths and weaknesses in specific domains.

A significant feature of any diagnostic language test is the feedback it provides to the test-takers (Alderson, 2005). Therefore, to evaluate a diagnostic test's usefulness, it is necessary to include an examination of the feedback provided. Bachman and Palmer (1996) list possible questions for evaluating test usefulness. One of them is "How relevant, complete, and meaningful is the feedback that is provided to the test takers?" (p. 146).

Heift (2003) defines meaningful feedback as a "response that provides a learning opportunity for students" (p. 533). Therefore, in order to provide meaningful information, diagnostic feedback needs to be descriptive and interpretable so that it can be more oriented toward learning (Black & Wiliam, 1998; Cotos & Pender, 2008). With explicit feedback, learners can make learning plans based on their current competence level, take steps towards remediation, make gradual improvement, and achieve their desired learning goals (Black & Wiliam, 1998; Cotos & Pender, 2008; Jang, 2009). In other words, by offering explicit and meaningful feedback, a diagnostic test has the potential for enhancing learning opportunities, resulting in positive washback.

However, the language testing literature is sparse regarding the topic of diagnostic feedback. Relatively few language tests are designed primarily for diagnosing test-takers' linguistic strengths and weaknesses so as to facilitate their future language learning process. DIALANG was the first major on-line diagnostic language assessment which aimed to provide test-takers with rich and informative diagnostic feedback. It was also the model of the OEAS battery utilized in the current study. Due to its importance in diagnostic language testing, DIALANG will be introduced in the following section.

2.3.3 The DIALANG Test Battery

The DIALANG project, based on the Common European Framework of Reference (CEFR), is an on-line diagnostic language assessment system. It is the first major testing system that is oriented towards diagnosing language skills and providing feedback to users rather than certifying their proficiency (Alderson & Huhta, 2005).

2.3.3.1 Description of DIALANG Test Battery

DIALANG includes tests in five aspects of language and language use: Reading, Listening, (indirect) Writing, Grammar, and Vocabulary, in 14 European languages. Table 2.3 shows the test content of English, the only language for which meaningful data is currently available (Alderson, 2005). In the DIALANG English test, there are a total of 276 test items distributed across six CEFR (Common European Framework of Reference for Languages) levels, namely A1, A2, B1, B2, C1, and C2. The CEFR is a reference framework developed by the Council of Europe in 2001. The CEFR aims to provide comprehensive descriptions of various proficiency levels in foreign language learning and it is widely used in Europe. The CEFR is also the basis for the DIALANG test framework and part of its test specifications (Alderson & Huhta, 2005). The CEFR divides learners into three broad divisions, A, B, and C. These three divisions can be further divided into six aforementioned levels, from the lowest A1 to the highest C2.

Table 2.3

Test content of DIALANG English test (adopted from Alderson, 2005, p.55)

Skill	Examples of Sub-skills	Number of items
Grammar	• Morphology: Adjectives and Adverbs- comparison, Verbs- active/ passive … • Syntax: word order statements, simple sentences vs complex sentences…	56
Listening	• Identifying main idea • Inferencing • Listening intensively for specific detail	56
Reading	• Inferencing • Identifying main idea • Reading intensively for specific detail	50
Vocabulary	• Combination • Meaning • Semantic relations • Word formation	60
Writing	• Knowledge of accuracy (grammar/ vocabulary/ spelling) • Knowledge of register/ appropriacy • Knowledge of textual organization (cohension/ coherence…)	54 (Total: 276)

Alderson and Huhta (2005) indicate that DIALANG aims to be a tool that supports independent, life-long language learning. Meanwhile, the learners need to take more responsibility for the assessment process. Therefore, users have complete freedom to choose which language and skill they wish to be tested in, whether to take the initial Vocabulary Size Placement Test (VSPT) which estimates their approximate language ability, or whether to self-assess their language ability based on the statements of the CEFR scales. In addition, users are free to take as many tests as they wish and they can quit a test at any point.

To better serve the diagnostic purpose, DIALANG focuses both on macro and micro levels of language (Alderson, 2005). The macro level measures test-takers' overall performance in the skill being tested and relates the test result to the levels of the CEFR. On the other hand, the micro level examines test-takers' strengths and weaknesses on the specific language sub-skill being tested. Thus, based on the micro level information, test-takers may for example discover that they are good at making inferences in reading but weaker in particular grammar structures. With focuses both on macro and micro level of language, the information provided by DIALANG as feedback enables test-takers to identify their own strengths and weaknesses in language, act upon the problem, and decide on how to improve their language ability.

2.3.3.2 Test Feedback of DIALANG

One of the main innovative features of DIALANG is the breadth of its feedback (Alderson & Huhta, 2005, p. 305). After taking each test of DIALANG, test-takers are provided with various kinds of feedback: 1) test result; 2) item review; 3) explanatory feedback; and 4) advisory feedback.

On completing the test, test-takers will first be presented with an overall "test result" in terms of the six levels of the CEFR – from A1 (the lowest) to C2 (the highest). Along with the test result, test-takers are also given a brief description of what learners at that CEFR level can do. Second is “item review”; a chart provides test-takers with brief explanations of items answered correctly and incorrectly. The items and explanations are classified according to the sub-skills being tested. Third, “explanatory feedback” states whether there is any mismatch between the test-taker’s self-assessed CEFR level and the DIALANG-assessed CEFR level. In this section, test-takers are able to understand possible reasons for the mismatch and are given warnings about the risks of over- and under-assessment of one’s ability. Fourth, “advisory feedback” provides explanations of what a test-taker at a given CEFR level can do as well as some advice on how test-takers can progress to the next level. The advisory feedback is presented in a series of tables. The aim of this is to encourage test-takers to reflect on what is involved in language learning (Alderson, 2005) and to help test-takers move forward in further language learning.

2.3.3.3 Reading Construct of DIALANG

To assess test-takers’ reading comprehension, the items in the Reading Test of DIALANG cover the following three reading sub-skills: “1) the ability to understand or identify the main idea, 2) the ability to find specific details or specific information, and 3) the ability to make inferences on the basis of the text by going beyond the literal meaning of the text or by inferring the approximate meaning of unfamiliar words” (Alderson, 2005, p. 125). The task types of the DIALANG reading test include multiple-choice questions, short-answer questions, and two kinds of gap-filling tasks.

The next section will be dedicated to the introduction of another diagnostic test battery, the Online English Assessment System (OEAS), which is designed primarily for self-assessment (Yin, 2006; Yin et al., forthcoming) and is aimed at Taiwanese university students for diagnosing their linguistic strengths and weaknesses.

2.3.4 The OEAS Test Battery

In 2005, the Taiwan Ministry of Education (MOE) began “Teaching Excellence” Projects (教學卓越計畫) with an aim to promote teaching quality in higher education. The Online English Assessment System (OEAS) was a sub-project under one of the projects proposed by Tunghai University. The goal of the OEAS project was to construct and validate a diagnostic language test that would assist the university’s students in their language learning (Yin, 2006). The OEAS intends not only to provide test-takers with better understanding of their English proficiency level but also to inform test-takers of their linguistic strengths and weaknesses in regards to the content tested. Eight faculty members in the Foreign Languages and Literature Department (FLLD) worked together in the OEAS test construction. There were two advantages of having teachers as test creators: first, these teachers were experienced in test construction and item writing; and second, they were familiar with Taiwanese students’ linguistic strengths and weaknesses (Yin, 2006).

The OEAS test battery was designed based on the model of DIALANG and it consists of two main sections: a general test and a group of skill tests. These two sections are introduced as follows.

2.3.4.1 The General Test

The OEAS general test comprises 60 multiple-choice questions – 20 each of grammar, listening, and reading. The main purpose of the general test is macro-diagnostic (see 2.3.3.1). The test results provide test-takers with information on their overall English proficiency level. As shown in Yin’s (2006) study, there were high correlations between the scores of the OEAS general test and the first tests of the intermediate and high intermediate General English Proficiency Tests (GEPT), respectively. Therefore, test-takers can estimate their possible performance on the GEPT based on the results the OEAS general test.

2.3.4.2 The Skill Tests

The skill tests form the second section of the OEAS test battery. Test-takers can take one or more multiple-choice skill tests regarding specific aspects of language: grammar, reading, and listening. These tests are designed to give test-takers specific "micro-diagnostic" information (see 2.3.3.1) about their strengths and weaknesses in that linguistic area as well as to give test-takers feedback upon the language skills tested. The content of the skill tests are shown in Table 2.4: 1) the grammar test presents test-takers with 60 questions to assess test-takers' knowledge of 15 grammatical structures; 2) the reading test consists of 4 testlets and each testlet contains a general expository reading passage and 8-11 questions covering the 6 reading sub-skills tested, the number of questions varies due to the length and difficulty of the reading passages; 3) the listening test contains 8 testlets which correspondingly include a listening passage and a question for each of the three listening sub-skills.

Table 2.4

OEAS skill tests introduction

Type of skill test	Contents	Number of questions	Examples of sub-skills tested
Grammar Test	15 structures x 4 questions	60	Past tense, modals, noun clauses
Reading Test	4 testlets x 8 to 11 questions	36	Reading for main idea, inference (see 2.3.4.3 for details)
Listening Test	8 testlets x 3 questions	24	General comprehension, inference

Regarding the aforementioned diagnostic purpose of OEAS, the forms of its feedback are designed based on the principles of diagnostic testing so as to help test-takers understand their linguistic strengths and weaknesses. After test-takers complete the test, they are provided with informative and enriched feedback both in Chinese and English. In brief, test-takers receive a summary report which lists percentages correct for each structure or sub-skill. This is meant to inform the test-takers which aspects of language they are strong or weak in. Additionally, if test-takers click on each structure or sub-skill label, they receive a general explanation of the sub-skill tested in both reading and listening tests and specific item explanations in grammar, reading and listening tests (see 2.3.4.3). The following section will focus on the OEAS Reading Test and its feedback since it will be utilized in the current study.

2.3.4.3 The OEAS Reading Test and Its Feedback

Table 2.5 illustrates the content of the OEAS Reading Test, including the topics of the passages, length of the reading passages, and each of the six reading sub-skills tested and its corresponding test items. The more detailed introductions are presented below.

First, the OEAS Reading Test consists of 4 testlets, each of which contains 8-11 items associated with a reading passage. The Reading Test comprises a total of 36 multiple choice items assessing test-takers' understanding of four reading passages of different lengths. The test-takers are allowed to spend 40 minutes to complete the Reading Test and the maximum score for the Reading Test is 36.

Second, each of the four testlets varies in topic since test-takers' background knowledge is a vital factor that influences reading comprehension (Murtagh, 1989; Anderson, 1999; Koda, 2005). That is, to prevent test-takers' performance from being influenced by prior knowledge or lack thereof, the four testlets cover various topics.

Third, each testlet contains a general expository reading passage and 8-11 questions covering the six reading sub-skills tested (see Table 2.5). According to Eskey (1986), expository passages best “represent the genre of discourse that EFL students often deal with in their academic studies” (cited in Sims, 2004, p. 316). Also, as suggested by Hughes (1989), a reading exam should include questions that involve both “macro-skills,” such as making generalizations or locating specific information, and “micro-skills,” such as identifying pronoun reference or using context clues to guess unknown vocabulary. Therefore, the reading passages in the OEAS Reading Test are composed of both macro questions and micro questions and cover the following six reading sub-skills: 1) reading for main ideas of a passage; 2) reading for main ideas of a paragraph; 3) reading for specific information; 4) guessing meaning of vocabulary from context; 5) pronoun reference; and 6) inference from reading.

Table 2.5

Sub-skills and corresponding items on the OEAS Reading Test

Testlet (Topic) [Word length of text] Item Sub-skill	1 (Promoting Earthworms) [362]	2 (Introducing Astrology) [474]	3 (Electricity and Fish) [374]	4 (The Industrial Revolution) [616]
	Number	Number	Number	Number
Sub-skill 1 (Main ideas of a passage)	1	9	17	26
Sub-skill 2 (Main ideas of a paragraph)	2,3	10,11	18,19	27,28
Sub-skill 3 (Specific information)	4,5	12,13	20,21	29,30
Sub-skill 4 (Guessing meaning of vocabulary from context)	6	14	22,23	31,32,33
Sub-skill 5 (Pronoun reference)	7	15	24	34,35
Sub-skill 6 (Inference)	8	16	25	36

To determine the Flesch Reading Ease and Flesch-Kincaid Grade Level of the four reading passages, the readability was calculated by using the Readability Statistic under spell check in Word for Windows. As shown in Table 2.6, the results indicated that the four reading passages of the OEAS Reading Test were of moderate difficulty and grade level.

Table 2.6

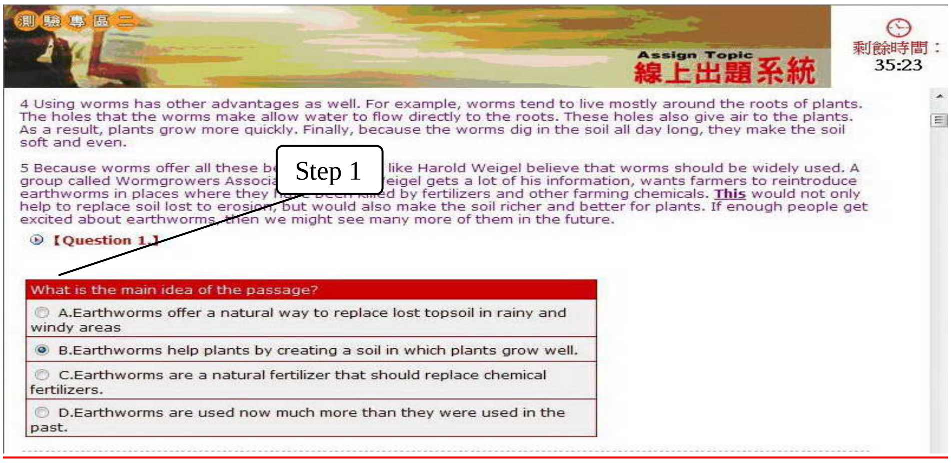
Readability of the four OEAS reading passages

Testlet	Flesch Reading Ease	Flesch-Kincaid Grade Level
1. Promoting Earthworms	58.1	9.3
2. Introducing Astrology	41.5	12.0
3. Electricity and Fish	49.8	10.9
4. The Industrial revolution	52.5	11.1

To check the validity of the OEAS Reading Test, two kinds of evidence have been gathered: First, at least 4 committee members, who didn’t participate in the test construction, were invited to check content validity of the Reading Test. After careful examination, these judges agreed that there was a great consistency between test specifications and test content. Second, the Spearman's correlation between the reading section of the General Test and the Specific Test of reading is 0.44. It is a positive correlation, though it only indicates a medium level of correlation. Some possible reasons for the medium level of correlation might be: (1) there is an imbalance in the questions in the two tests**;** the general test gives 2 passages with 10 questions each while the specific test gives 4 passages with 8-11 questions each, (2) the topics of the texts in the two tests are all different, and (3) for the same group of students taking both tests, the average score was 63% on the general test reading section but only 45% on the reading skill test (Yin, personal communication). In other words, the differences in the length of reading texts (see Table 2.5), test-takers’ background knowledge about the texts, and difficulty levels of texts may account for the medium level of correlation.

The following paragraphs illustrate the steps of how test-takers take the OEAS Reading Test and how test feedback is presented to the test-takers. First, in Step 1, test-takers read each of the 4 reading passages and answer associated questions, as shown in Figure 2.1. Second, as Step 2 shown in Figure 2.2, on completing the 36 questions of the OEAS Reading test, test-takers click on the button of "送出答案" (Submit answers).

Figure 2.1 *Steps of operating the OEAS Reading Test (1)*



4 Using worms has other advantages as well. For example, worms tend to live mostly around the roots of plants. The holes that the worms make allow water to flow directly to the roots. These holes also give air to the plants. As a result, plants grow more quickly. Finally, because the worms dig in the soil all day long, they make the soil soft and even.

5 Because worms offer all these be like Harold Weigel believe that worms should be widely used. A group called Wormgrowers Associa eigel gets a lot of his information, wants farmers to reintroduce earthworms in places where they have been killed by fertilizers and other farming chemicals. **This** would not only help to replace soil lost to erosion, but would also make the soil richer and better for plants. If enough people get excited about earthworms, then we might see many more of them in the future.

【Question 1.】

What is the main idea of the passage?

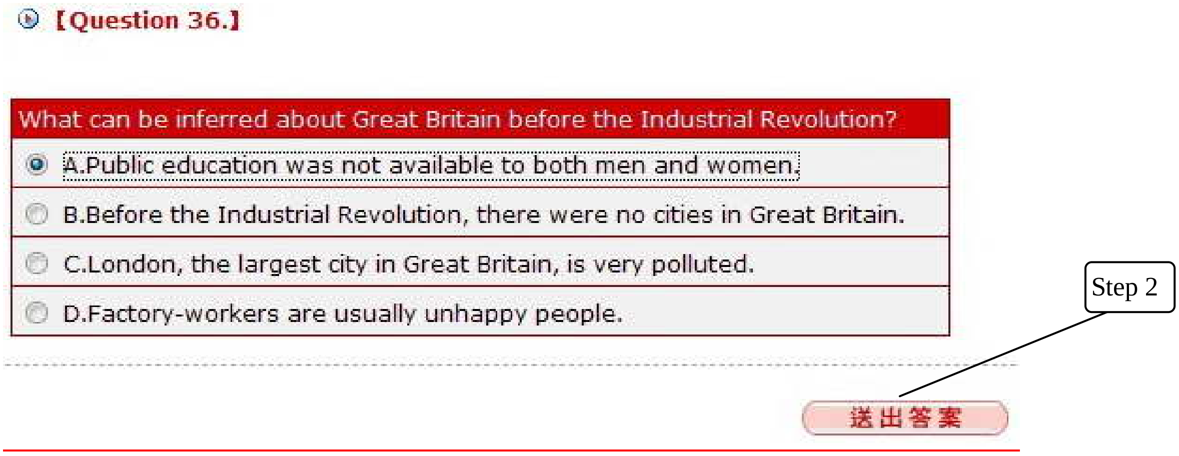
○ A.Earthworms offer a natural way to replace lost topsoil in rainy and windy areas

● B.Earthworms help plants by creating a soil in which plants grow well.

○ C.Earthworms are a natural fertilizer that should replace chemical fertilizers.

○ D.Earthworms are used now much more than they were used in the past.

Figure 2.2 *Steps of operating the OEAS Reading Test (2)*



【Question 36.】

What can be inferred about Great Britain before the Industrial Revolution?

● A.Public education was not available to both men and women.

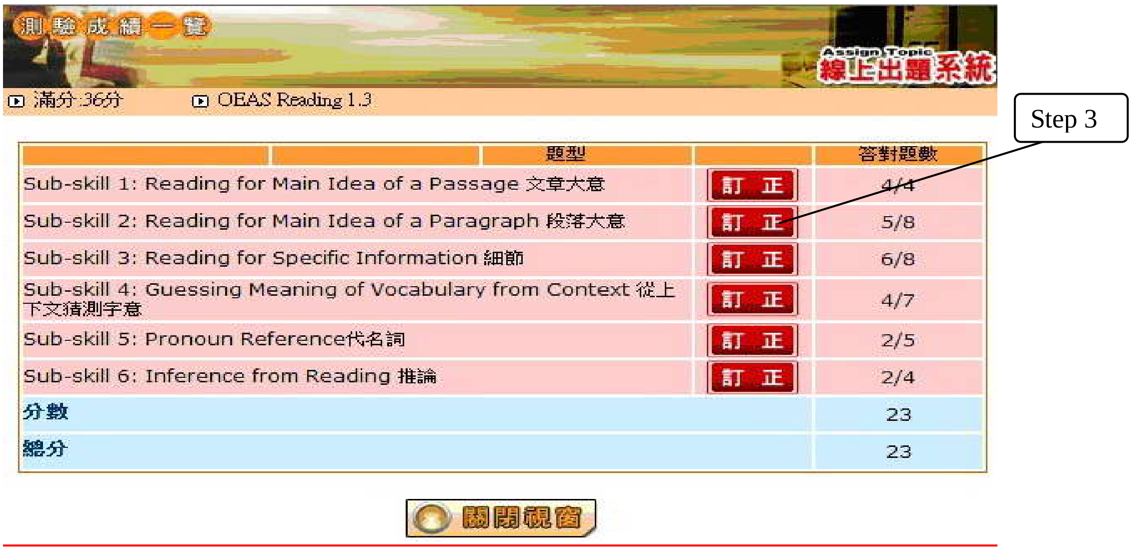
○ B.Before the Industrial Revolution, there were no cities in Great Britain.

○ C.London, the largest city in Great Britain, is very polluted.

○ D.Factory-workers are usually unhappy people.

After submitting their answers, test-takers will be presented with an overall report (see Figure 2.3) which is grouped according to the 6 reading sub-skills tested. The overall report indicates both how many items the test-takers answered correctly in each reading sub-skill and their total scores of the OEAS Reading Test. Then in Step 3, also shown in Figure 2.3, test-takers click on the "訂正" (Correction) button and will be led to more detailed feedback for each reading sub-skill.

Figure 2.3 *Steps of operating the OEAS Reading Test (3)*



滿分:36分　　OEAS Reading 1.3

題型		答對題數
Sub-skill 1: Reading for Main Idea of a Passage 文章大意	訂 正	4/4
Sub-skill 2: Reading for Main Idea of a Paragraph 段落大意	訂 正	5/8
Sub-skill 3: Reading for Specific Information 細節	訂 正	6/8
Sub-skill 4: Guessing Meaning of Vocabulary from Context 從上下文猜測字意	訂 正	4/7
Sub-skill 5: Pronoun Reference代名詞	訂 正	2/5
Sub-skill 6: Inference from Reading 推論	訂 正	2/4
分數		23
總分		23

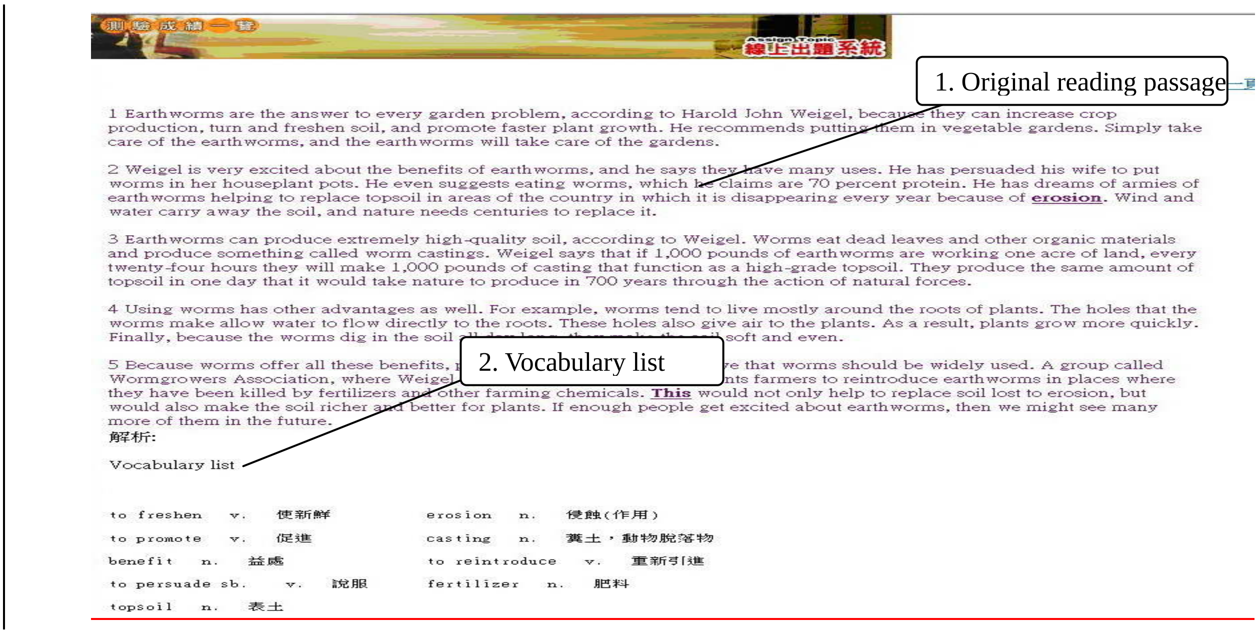
Canale (1984) suggests that in addition to assessing reading performance in a reliable, valid way, good reading tests should also provide "clear, rich, relevant, and generalizable feedback" (p. 351). The OEAS offers seven kinds of feedback on the Reading Test (Table 2.7). One is an overall report of the 6 reading skills tested with the number of questions answered correctly by the test-taker (Figure 2.3). This is meant to give the student an overall picture of which reading skills the student is weaker or stronger in. Clicking on one type of the reading skills leads to the other six kinds of feedback (Figures 2.4~2.7), which shows the test-taker (1) the original reading passages, (2) a vocabulary list for each reading passage, (3) equivalent Chinese translation of each reading passage, (4) the item itself with the choice made by the test-taker and the correct answer, (5) a general explanation, in both English and Chinese, of the answer and corresponding reading skill, and (6) specific explanations for each test item.

Table 2.7

Types of the OEAS Reading Test feedback

1. An overall report of 6 sub-skills
2. Reading passages
3. Vocabulary lists of each reading passage
4. Chinese translations of each reading passage
5. The item itself with the choice made by the test-taker and the correct answer
6. General explanations for each sub-skills
7. Specific explanations for each item

Figure 2.4 *Feedback of the OEAS Reading Test (1)*



1 Earthworms are the answer to every garden problem, according to Harold John Weigel, because they can increase crop production, turn and freshen soil, and promote faster plant growth. He recommends putting them in vegetable gardens. Simply take care of the earthworms, and the earthworms will take care of the gardens.

2 Weigel is very excited about the benefits of earthworms, and he says they have many uses. He has persuaded his wife to put worms in her houseplant pots. He even suggests eating worms, which he claims are 70 percent protein. He has dreams of armies of earthworms helping to replace topsoil in areas of the country in which it is disappearing every year because of **erosion**. Wind and water carry away the soil, and nature needs centuries to replace it.

3 Earthworms can produce extremely high-quality soil, according to Weigel. Worms eat dead leaves and other organic materials and produce something called worm castings. Weigel says that if 1,000 pounds of earthworms are working one acre of land, every twenty-four hours they will make 1,000 pounds of casting that function as a high-grade topsoil. They produce the same amount of topsoil in one day that it would take nature to produce in 700 years through the action of natural forces.

4 Using worms has other advantages as well. For example, worms tend to live mostly around the roots of plants. The holes that the worms make allow water to flow directly to the roots. These holes also give air to the plants. As a result, plants grow more quickly. Finally, because the worms dig in the soil all day long, they make the soil soft and even.

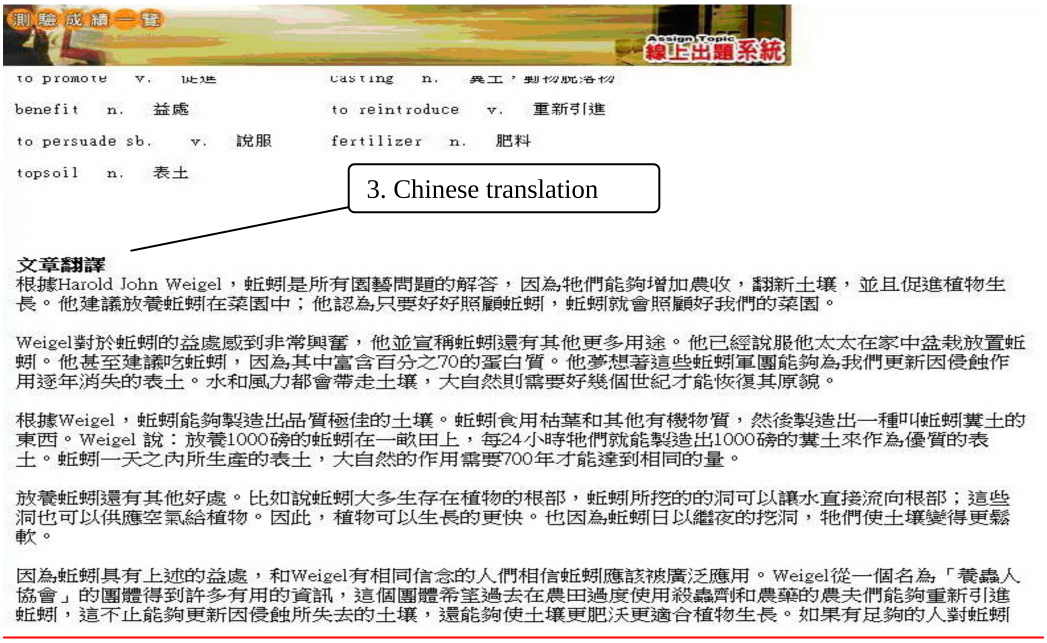
5 Because worms offer all these benefits, p ve that worms should be widely used. A group called Wormgrowers Association, where Weigel nts farmers to reintroduce earthworms in places where they have been killed by fertilizers and other farming chemicals. **This** would not only help to replace soil lost to erosion, but would also make the soil richer and better for plants. If enough people get excited about earthworms, then we might see many more of them in the future.

解析:

Vocabulary list

to freshen	v.	使新鮮	erosion	n.	侵蝕(作用)
to promote	v.	促進	casting	n.	糞土，動物脫落物
benefit	n.	益處	to reintroduce	v.	重新引進
to persuade sb.	v.	說服	fertilizer	n.	肥料
topsoil	n.	表土			

Figure 2.5 *Feedback of the OEAS Reading Test (2)*



to promote	v.	促進	casting	n.	糞土，動物脫落物
benefit	n.	益處	to reintroduce	v.	重新引進
to persuade sb.	v.	說服	fertilizer	n.	肥料
topsoil	n.	表土			

3. Chinese translation

文章翻譯

根據Harold John Weigel，蚯蚓是所有園藝問題的解答，因為牠們能夠增加農收，翻新土壤，並且促進植物生長。他建議放養蚯蚓在菜園中；他認為只要好好照顧蚯蚓，蚯蚓就會照顧好我們的菜園。

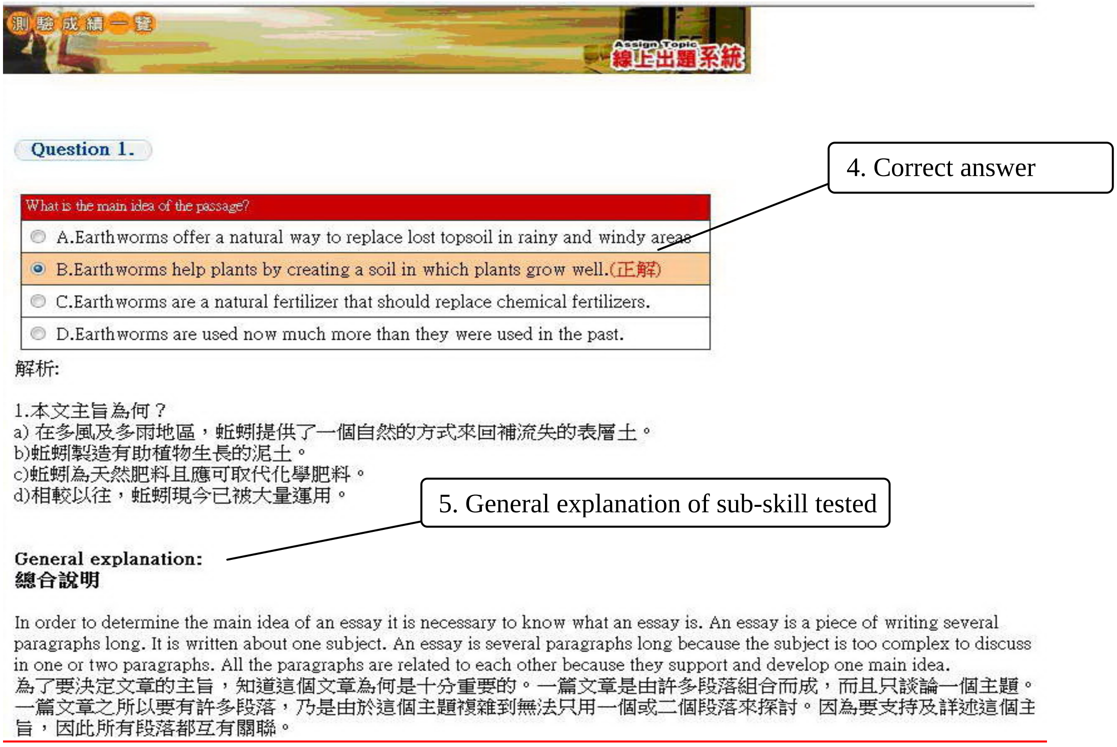
Weigel對於蚯蚓的益處感到非常興奮，他並宣稱蚯蚓還有其他更多用途。他已經說服他太太在家中盆栽放置蚯蚓。他甚至建議吃蚯蚓，因為其中富含百分之70的蛋白質。他夢想著這些蚯蚓軍團能夠為我們更新因侵蝕作用逐年消失的表土。水和風力都會帶走土壤，大自然則需要好幾個世紀才能恢復其原貌。

根據Weigel，蚯蚓能夠製造出品質極佳的土壤。蚯蚓食用枯葉和其他有機物質，然後製造出一種叫蚯蚓糞土的東西。Weigel 說：放養1000磅的蚯蚓在一畝田上，每24小時牠們就能製造出1000磅的糞土來作為優質的表土。蚯蚓一天之內所生產的表土，大自然的作用需要700年才能達到相同的量。

放養蚯蚓還有其他好處。比如說蚯蚓大多生存在植物的根部，蚯蚓所挖的的洞可以讓水直接流向根部；這些洞也可以供應空氣給植物。因此，植物可以生長的更快。也因為蚯蚓日以繼夜的挖洞，牠們使土壤變得更鬆軟。

因為蚯蚓具有上述的益處，和Weigel有相同信念的人們相信蚯蚓應該被廣泛應用。Weigel從一個名為「養蟲人協會」的團體得到許多有用的資訊，這個團體希望過去在農田過度使用殺蟲劑和農藥的農夫們能夠重新引進蚯蚓，這不止能夠更新因侵蝕所失去的土壤，還能夠使土壤更肥沃更適合植物生長。如果有足夠的人對蚯蚓

Figure 2.6 *Feedback of the OEAS Reading Test (3)*



Question 1.

4. Correct answer

What is the main idea of the passage?

○ A.Earthworms offer a natural way to replace lost topsoil in rainy and windy areas

● B.Earthworms help plants by creating a soil in which plants grow well.(正解)

○ C.Earthworms are a natural fertilizer that should replace chemical fertilizers.

○ D.Earthworms are used now much more than they were used in the past.

解析:

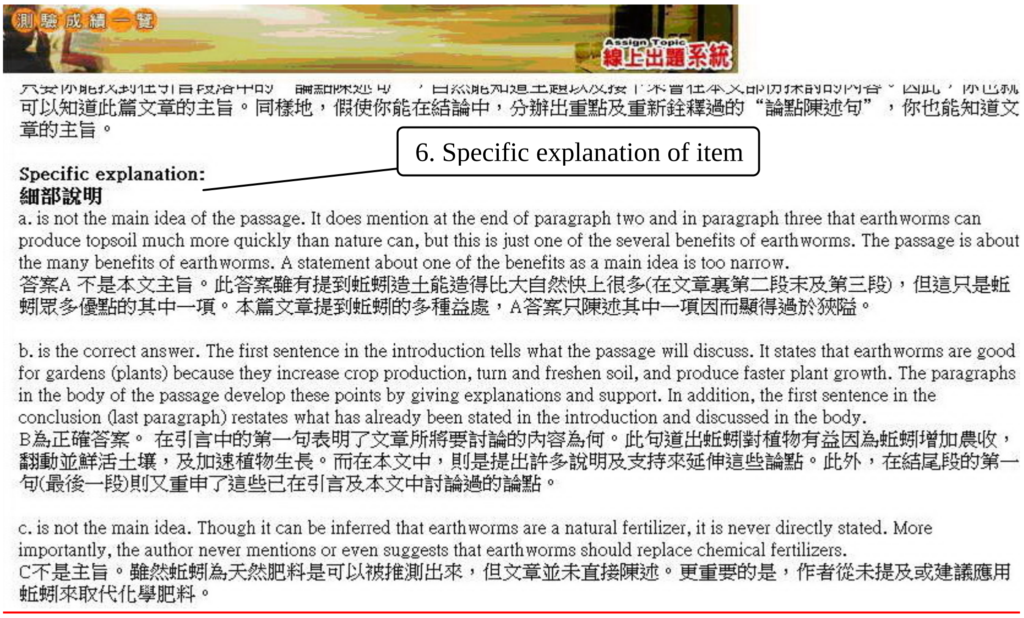
1.本文主旨為何？
a) 在多風及多雨地區，蚯蚓提供了一個自然的方式來回補流失的表層土。
b)蚯蚓製造有助植物生長的泥土。
c)蚯蚓為天然肥料且應可取代化學肥料。
d)相較以往，蚯蚓現今已被大量運用。

5. General explanation of sub-skill tested

General explanation:
總合說明

In order to determine the main idea of an essay it is necessary to know what an essay is. An essay is a piece of writing several paragraphs long. It is written about one subject. An essay is several paragraphs long because the subject is too complex to discuss in one or two paragraphs. All the paragraphs are related to each other because they support and develop one main idea.
為了要決定文章的主旨，知道這個文章為何是十分重要的。一篇文章是由許多段落組合而成，而且只談論一個主題。一篇文章之所以要有許多段落，乃是由於這個主題複雜到無法只用一個或二個段落來探討。因為要支持及詳述這個主旨，因此所有段落都互有關聯。

Figure 2.7 *Feedback of the OEAS Reading Test (4)*



只要你能找到在引言段落中的 "論點陳述句" ，自然能知道主題以及接下來會在本文部份探討的內容。因此，你也就可以知道此篇文章的主旨。同樣地，假使你能在結論中，分辨出重點及重新銓釋過的 "論點陳述句" ，你也能知道文章的主旨。

Specific explanation:
細部說明

a. is not the main idea of the passage. It does mention at the end of paragraph two and in paragraph three that earthworms can produce topsoil much more quickly than nature can, but this is just one of the several benefits of earthworms. The passage is about the many benefits of earthworms. A statement about one of the benefits as a main idea is too narrow.
答案A 不是本文主旨。此答案雖有提到蚯蚓造土能造得比大自然快上很多(在文章裏第二段末及第三段)，但這只是蚯蚓眾多優點的其中一項。本篇文章提到蚯蚓的多種益處，A答案只陳述其中一項因而顯得過於狹隘。

b. is the correct answer. The first sentence in the introduction tells what the passage will discuss. It states that earthworms are good for gardens (plants) because they increase crop production, turn and freshen soil, and produce faster plant growth. The paragraphs in the body of the passage develop these points by giving explanations and support. In addition, the first sentence in the conclusion (last paragraph) restates what has already been stated in the introduction and discussed in the body.
B為正確答案。 在引言中的第一句表明了文章所將要討論的內容為何。此句道出蚯蚓對植物有益因為蚯蚓增加農收，翻動並鮮活土壤，及加速植物生長。而在本文中，則是提出許多說明及支持來延伸這些論點。此外，在結尾段的第一句(最後一段)則又重申了這些已在引言及本文中討論過的論點。

c. is not the main idea. Though it can be inferred that earthworms are a natural fertilizer, it is never directly stated. More importantly, the author never mentions or even suggests that earthworms should replace chemical fertilizers.
C不是主旨。雖然蚯蚓為天然肥料是可以被推測出來，但文章並未直接陳述。更重要的是，作者從未提及或建議應用蚯蚓來取代化學肥料。

2.4 Student's Perceptions of Diagnostic Test Feedback

According to Bachman and Palmer (1996), one of the essential elements to evaluate the usefulness of a test is to investigate "how relevant, complete, and meaningful is test feedback that is provided to test-takers" (p.146). Also, Kunnan and Jang (2009) point out that both teachers' and students' opinions are equally important and should be all taken into consideration if one wants to investigate the use of diagnostic feedback. They further point out that diagnostic feedback may have different effects due to various factors such as the learners' competency levels, cognitive and metacognitive learning styles, and learning context.

Despite the fact that only a few diagnostic tests are specially designed for providing diagnostic feedback (Alderson, 2005; Hughes, 2003; Jang, 2009), some projects, such as DIALANG and OEAS discussed in the previous section, have been developed to serve the purpose of diagnostic testing. Also, Jang (2009) used the existing tutorial courseware, Educational Testing Service's *LanguEdge*^{TM}, to investigate test-takers' perspectives on the usefulness of the diagnostic feedback. These tests aim to help test-takers discover their own linguistic strengths and weaknesses as well as provide feedback that would facilitate test-takers in their future language learning. This section will focus on the recent studies that have examined the test-takers' perceptions about the usefulness of test feedback.

As cited in Alderson (2005), some researchers (Huhta & Figueras, 2001; Yang, 2003) have conducted studies to examine the test feedback of DIALANG. In Yang's (2003) study, 12 overseas graduate students at Lancaster University were interviewed after taking DIALANG. The result showed that students considered Item Review and Advisory Feedback as the most useful because they provided the test-takers with specific explanations of the items as well as some advice on how test-takers can improve in their language abilities and progress to a higher level. In a larger study conducted by Huhta and Figueras (2001), test-takers involved were from a variety of educational institutes in Finland, Germany, and Spain. The results showed that test-takers in general held positive attitudes toward the feedback and they had preference for richer feedback such as the Advisory Feedback.

In another study, Jang (2009) utilized Educational Testing Service's *LanguEdge*^{TM} courseware to assess students' reading comprehension. The courseware was developed as an instructional tool for English as second language (ESL) classrooms and it included a prototype of the Internet-based Test of English as a Foreign Language (iBT TOEFL) in two reading comprehension test forms. Since the tests were not designed specifically for diagnostic purposes, Jang applied a Cognitive Diagnostic Assessment (CDA) approach to make inferences about test-takers' competency in the tested skills and to develop meaningful diagnostic feedback for test-takers. The feedback was presented in a report card (see Jang, 2009, p.72-73) which provided detailed information about test-takers' performance on the reading comprehension test. First, the report card included a chart which listed out the correct answers as well as whether the test-taker answered correctly on each item. In addition, the difficulty level (in this case easy, medium, and hard respectively) of each item was also shown. Also, the total score and the individual score at each level were presented. Second, a bar graph indicated the test-taker's level of mastery of nine tested reading skills. Third, a chart provided descriptions of each reading skill as well as symbols indicating skills that the test-taker was weaker in and thus needed further improvement upon.

To investigate test-takers' perspectives on the usefulness of the diagnostic feedback, Jang employed both quantitative and qualitative methods. Twenty-eight students of two TOEFL preparation courses in the USA took the reading tests pre- and post-instruction. After taking the tests, the students received the report cards at both the beginning and end of the two-month instructional term. The 28 students provided their opinions about the usefulness of the diagnostic feedback. The participants' opinions on the diagnostic reports were gathered using questionnaires and interviews. The finding showed that 89% of the students thought that the report card was useful and it accurately reflected their reading skills. Furthermore, those who performed weakly in the reading tests expressed their desire to study harder and asked for more guidance on how to improve their reading skills. The findings of Jang's study suggested that the diagnostic reports had an overall positive impact on students' learning.

Yin, Sims, and Cothran (forthcoming) evaluated the perceived pedagogic usefulness of feedback to Taiwanese university students on an online multiple-choice diagnostic test of English grammar. Yin et al. also employed a mixed-method approach to gather quantitative and qualitative data separately in two stages. In Stage I, 94 university students took the English grammar test and filled in the questionnaire as they read the feedback. In Stage II, 5 students, not involved in Stage I, were interviewed individually while reading test feedback. The test feedback students received included the item itself with the choice made by the student and grammatical explanations both in Chinese and English. The findings showed that students as a whole regarded the feedback to be very useful, and that the Chinese feedback was considered to be more useful than the English. As for students of different English proficiency levels, low-scorers did not deem the feedback as useful as the high-scorers did. In the study of Yin et al. (forthcoming), students felt good feedback should have certain characteristics. For example, students liked explanations that gave clear examples of correct and/ or incorrect usage and they also preferred explanations that pointed out common misunderstandings.

In sum, the results of the aforementioned studies suggested that test-takers as a whole perceive feedback to be useful. However, their preferences for feedback might differ for various reasons. For example, the forms of the feedback and the test-takers' proficiency level all might have impact on their preferences. Though the abovementioned studies have evaluated test-takers' perceptions of diagnostic test feedback, there are still not fully generalizeable. For instance, some of the research (Yang, 2003; Huhta & Figueras, 2001) examined test-takers' overall perceptions of the test feedback provided by DIALANG without focusing on specific language aspects such as reading, grammar, or listening. Although Jang's study (2009) investigated test-taker's perspectives on the usefulness of the diagnostic feedback of reading comprehension tests, the study was conducted in an ESL context and the tests were not designed particularly for diagnostic purposes. The study of Yin et al. (forthcoming) was conducted in an EFL context in Taiwan but it investigate test-takers' perceptions of diagnostic grammar test feedback. Hence, there is a lack of research on Taiwanese students' perceptions of diagnostic reading test feedback and more empirical studies focusing on this issue are necessary.

2.5 Summary and Research Gap

This chapter has reviewed literature related to the current study in several areas. First, various models of reading process and definitions of reading construct were introduced. Second, an introduction of computer-assisted language testing was made and its advantages were presented. Third, diagnostic language tests, such as DIALANG and OEAS, were described. Meanwhile, an overview of the importance of test feedback and a variety of diagnostic feedback, which aims to facilitate test-takers' future language learning, was presented. Finally, several studies investigating students' views about test feedback were discussed so as to understand learners' perceptions of the diagnostic test feedback.

While the literature shows that diagnostic testing has drawn growing attention in the field of language assessment, there are still some limits. First, many researchers (Alderson, 2005; Jang, 2009) point out that in contrast to other fields of language assessment, there are relatively few tests specifically designed for the purpose of diagnostic language testing. Second, though Computer-Assisted Language Testing has been widely implemented, Alderson (2005) suggests that the complexity inherent in computer-assisted diagnostic assessment calls for further research. Third, it is important to examine the usefulness of test feedback provided by diagnostic language tests; however, little research has focused on students' perceptions about feedback of diagnostic language tests (Yin et al., forthcoming), not to mention that of reading tests. Therefore, in view of these limits the current study has been designed to fill the research gap and aims to investigate Taiwanese university students' perceptions about diagnostic reading test feedback.

CHAPTER THREE

Methodology

This chapter introduces the research methodology for this study. This study employs a mixed methods design. Quantitative and qualitative data were collected separately in two stages to evaluate students' thoughts about the usefulness of the Reading Test feedback. The content of this chapter is presented in the following sequence: 3.1) research questions and overall design, 3.2) Stage I methodology, 3.3) Stage II methodology, 3.4) validation, 3.5) summary of the research design, 3.6) pilot study, and 3.7) summary of the chapter.

3.1 Research Questions and Overall Design

The present study aimed to address the following three research questions:

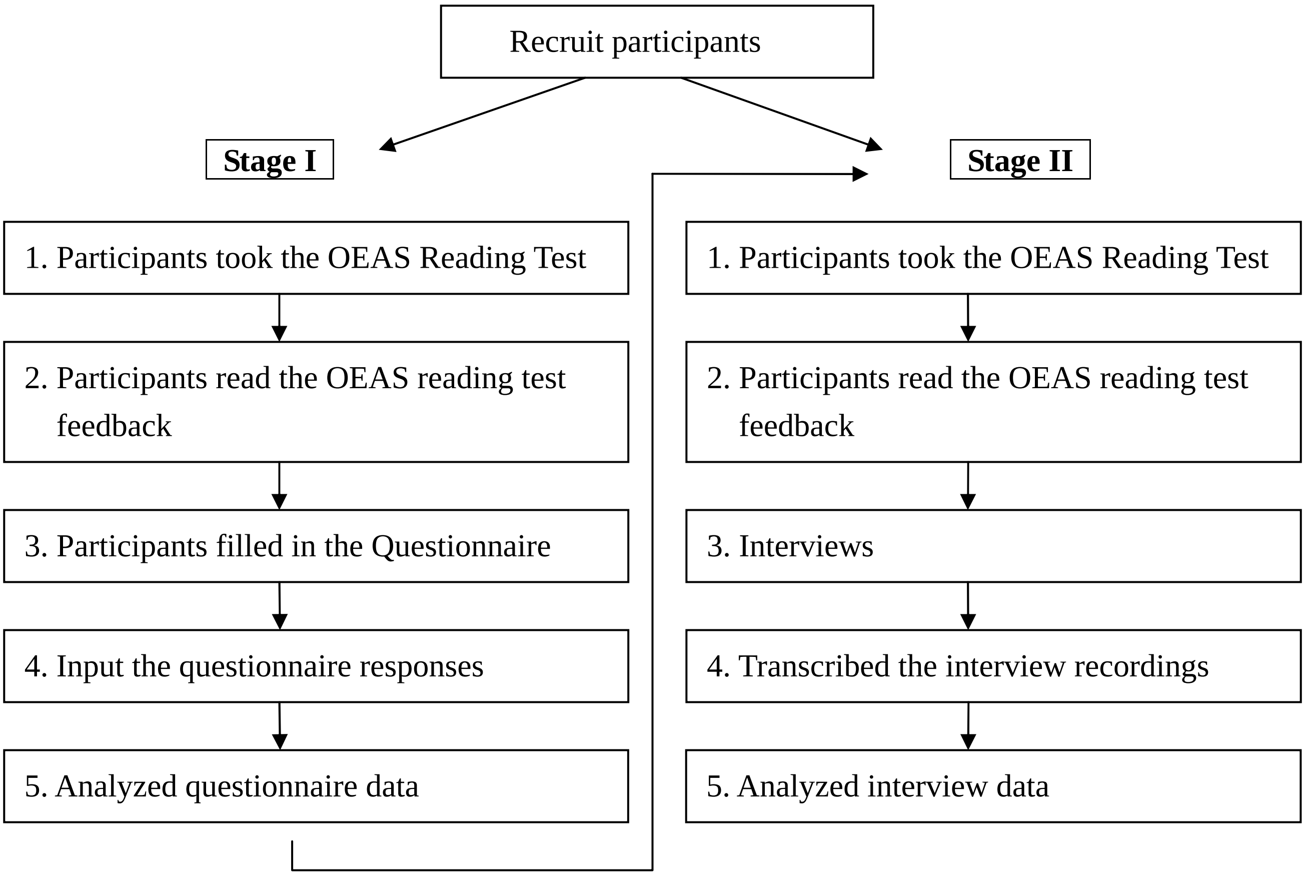
1. How useful do test-takers perceive the Reading Test feedback to be?
2. In what way do the test-takers perceive the Reading Test feedback to be useful or not?
3. Is there a significant difference between low- and high- English proficiency level test-takers in their perception of the Reading Test feedback's usefulness?

In order to find answers to these research questions, this study adopted a two-stage mixed methods design. According to Creswell and Garrett (2008), a mixed methods design is capable of providing an in-depth understanding of research problems by combining quantitative and qualitative data, and explaining the quantitative statistical results in more detail with qualitative data. Therefore, a mixed method approach was utilized in this study. In Stage I, the researcher administered questionnaires to collect quantitative data with an aim to answer research questions (1) and (3). In Stage II, interviews were conducted to provide qualitative data to answer questions (2) and (3).

The following figure shows the overall research design of this study and the data collection procedures:

Figure 3.1

Overall design of the current study


Recruit participants
Stage I
Stage II
1. Participants took the OEAS Reading Test
2. Participants read the OEAS reading test feedback
3. Participants filled in the Questionnaire
4. Input the questionnaire responses
5. Analyzed questionnaire data
1. Participants took the OEAS Reading Test
2. Participants read the OEAS reading test feedback
3. Interviews
4. Transcribed the interview recordings
5. Analyzed interview data

3.2 Stage Ⅰ Methodology

This section presents Stage I methodology. First, the participants in Stage I of the study are introduced. Next, the instrument - a questionnaire - employed in this study and data collection procedures are presented. Finally, data analysis procedures are described.

3.2.1 Participants

In Stage I of this study, the participants were 48 freshmen from Tunghai University in central Taiwan. In this university, all the freshmen have to take an English placement exam during freshman orientation. The purpose of the placement exam is to divide students into classes based on their language ability. Based on the results of this placement exam, students are placed into one of approximately 100 sections of Freshman English for Non-Majors (Sims, 2004). Since there are existing English classes based on the results of the placement test, the participants were recruited mainly from three different sources (see Table 3.1) to have a representative sample of university freshmen in Taiwan. The first two sources of collecting participants were: Freshman English for Non-Majors (FENM) program, and English majors from the Foreign Languages and Literature Department (FLLD). Meanwhile, the researcher placed an advertisement on the website of the English Language Center as the third source of recruiting participants. Even though some researchers (Chen, 1997; Luo, 2005) pointed out that Taiwanese university students' English ability has declined dramatically, Sims (2004) and Chen (2006) both suggested that the overall English ability of the incoming freshmen at Tunghai University had been quite consistent over the past few years. In other words, regardless of students' year of study in school, the results of the above research indicated no significant differences in their total measure of English ability when taking the placement test. Therefore, the researcher considered that freshmen were suitable and representative for the present study.

To recruit participants from the existing English classes, the researcher first explained the purpose of the study and the data collection procedures to the English instructors of the university. With the instructors' consent, the students in these classes were informed about the nature of this study. Then, students from each class signed up voluntarily to participate in this study. The participants received 150 NT compensation for completing both the OEAS Reading Test and the questionnaire.

Table 3.1

Three major sources of recruiting participants

1. Students recruited from Freshman English for Non-Majors program
2. Students recruited from the website of English Language Center
3. Students (English majors) recruited from the FLLD

Using the above three approaches to collect participants was meant to have students of diverse backgrounds. In this way, the participants collected varied in gender and major. Additionally, to have participants of different English proficiency levels, the researcher selected target participants among the students who signed up for this current study. At first, the mean score of 2009 Tunghai English Placement Exam was calculated. Afterwards, students who scored above and below one standard deviation were selected. In addition, parts of the high-level students in this study were also chosen from high-level students of the English department. Apart from the Tunghai English Placement Exam, the FLLD groups the English majors into three groups, namely high, intermediate, and low based on the result of a placement test administered to English majors only.

3.2.1.1 Background Information of Stage I Participants

The following table illustrates background information of the 48 participants in Stage I. As shown in Table 3.2, all the 48 participants were university freshmen, including 11 (22.9%) males and 37 (77.1%) female, but varied in their field of study and English proficiency levels. Additionally, the participants' English proficiency levels were confirmed again by the results of the OEAS Reading Test and they were classified as high- and low-level group respectively in this study (see 4.1.1 for details).

Table 3.2

Stage I participants' characteristics

Characteristic	Number (total n=48)
Gender	
Male	11
Female	37
Year in school	
1	48
Major of study	
Accounting	1
Animal Science	2
Business Administration	1
Chemical & Materials Engineering	1
Economics	4
Electrical engineering	2
Finance	3
Fine art	2
Food science	7
Foreign languages & literature (including English)	5
International Trade	8
Law	1
Life Science	4
Mathematics	1
Political Science	1
Social Work	2
Sociology	3

3.2.2 Instrument

To collect the data for Stage I, a questionnaire (Chinese version, see Appendix E) was employed to gather participants' thoughts about the OEAS Reading Test feedback. Many researchers (e.g. Brown, 2001; Dörnyei, 2003; Seliger & Shohamy, 1989) have suggested that there are a number of positive features of questionnaires. For example, questionnaires allow data to be collected for large-scale study on a one-shot basis; they are more efficient and economical. Also, the data produced is number-oriented and capable of being processed in a statistical manner. Further, with uniform instruments and controlled factors, data can be collected in a standardized manner across all participants. Additionally, since questionnaires are generally used to collect data from a large group of subjects, the result is more representative and more likely to be generalizable.

Survey of Students' Perceptions of OEAS Reading Test Feedback

In the current study, a paper-and-pencil questionnaire was administered to the participants, who completed it right after taking the OEAS Reading Test. The questionnaire asked test-takers to rate the usefulness of each item's feedback on a 5-point Likert scale (see Appendix E), with 1 being not useful at all and 5 being very useful. The design of the questionnaire followed the Reading Test feedback described in 2.3.4. The questionnaire included two parts and consisted of a total of 55 questions (see Table 3.3). Part I included 5 questions (1~5) about participants' personal background information.

Part II of the questionnaire, with a total of 50 items, was further divided into two sections (A & B) to survey participants' thoughts about the reading test feedback. In section A, the only question (A1) was to examine participants' perceptions about the overall test result report. In section B, the other 49 questionnaire items were further classified into seven categories. The first six categories were in accordance with the six reading sub-skills tested (see 2.3.4.3). To be more specific, the first question in each of the six categories (B1-1, B2-1, B3-1, B4-1, B5-1, B6-1) was to survey participants' thoughts about general explanations for each reading sub-skill tested and the other questions were to investigate participants' thoughts about specific test item feedback. Though each item feedback of the OEAS reading test consisted of an English and Chinese version, each questionnaire item asked for participants' opinion of feedback including Chinese and English together. Yin et al. (forthcoming) found that students consistently preferred Chinese feedback, so it was decided that there was no need to make the questionnaire too long by asking about Chinese and English feedback separately.

After that, the last category, including 7 items (B7-1~B7-7), was designed to investigate participants' general attitudes towards the reading test feedback. In addition to rating the usefulness of the test feedback, test-takers could also make additional comments after each questionnaire item. At the end of the questionnaire, one more open-ended question that asked the participant to share anything they want to say about the Reading Test or feedback was added for the sake of soliciting more responses, if any.

Table 3.3

Description of questionnaire items

	Sections / Categories	Item Number
Part I	Participants' personal background information	1~5
Part II	A— Overall test result report	A1
	B— Reading sub-skill 1: reading for man idea of a passage	B1-1~B1-5
	Reading sub-skill 2: reading for main idea of a paragraph	B2-1~B2-9
	Reading sub-skill 3: reading of specific information	B3-1~B3-9
	Reading sub-skill 4: guessing meaning of vocabulary from context	B4-1~B4-8
	Reading sub-skill 5: pronoun reference	B5-1~B5-6
	Reading sub-skill 6: inference from reading	B6-1~B6-5
	Perceptions about general aspects of feedback	B7-1~B7-7

3.2.3 Data Collection Procedures

In Stage I, 48 participants signed up for slots spread over three days to take the OEAS Reading Test in a computer classroom. The researcher administered the Reading Test for each time and the whole process of data collection in Stage I was about 1.5 hours. The steps of data collection in Stage I are shown in Table 3.4. Prior to the test, the researcher first explained the purpose of the study to the participants and gave them instructions (see Appendix C) on how to use the OEAS program. To avoid the interfering effect of participants' English abilities, all the instructions were given in their native language, Mandarin Chinese. After that, participants filled in the personal background information questionnaire (Part I). Then, the participants started to take the OEAS Reading Test for about 40 minutes. As soon as the participants finished the OEAS Reading Test, they were asked to complete the second part of the questionnaire while they read the reading test feedback (see 2.3.4.3 for details).

Table 3.4

Steps of Stage I data collection

Step	Descriptions
1	The researcher explained the purpose of the study and give instructions on using the OEAS program
2	The participants filled in questionnaire Part I
3	The participants took the OEAS Reading Test for 40 minutes
4	The participants filled in questionnaire Part II as they read the feedback

3.2.4 Data Analysis Procedures

The computer software package SPSS 17.0 for Windows was used to organize, compute, and analyze the quantitative data gathered from the questionnaire. To answer RQ1, descriptive statistics was conducted on each questionnaire items in the *Survey of Students' Thoughts about OEAS Reading Test Feedback* to obtain frequency, average and standard deviation of the scores for each of the items. The result helped to understand how test-takers rate the usefulness of various aspects of the Reading Test feedback. Furthermore, to answer RQ3, independent-sample t-tests were run to examine whether there are significant differences between the high and low English proficiency groups in their responses about the usefulness of the Reading Test feedback.

3.3 Stage II Methodology

The purpose of Stage II is to understand more deeply what forms of the reading feedback test-takers' perceive to be useful and how students of different English proficiency levels perceive the feedback. The following sections illustrate the methodology of Stage II. First, the participants in Stage II of the study are introduced. Next, the data collection procedures are presented. Finally, data analysis procedures are described.

3.3.1 Participants

The participants in Stage II were six university students who were not involved in the Stage I research. As mentioned in Stage I, the participants were recruited both by word of mouth and through the advertisement on the school website. These six participants varied in gender, major, year of study, and English proficiency levels. Considering the phenomenon of "data saturation" (Morgan, 2002, as cited in Yin et al., forthcoming), which is common in qualitative usability studies (Luoma & Tarnanen, 2003), only 6 participants were interviewed. In some previous studies (Bunce, Guestt & Johnson, 2006; Yin et al., forthcoming), researchers originally intended to interview a large number of participants. Nevertheless, after the first few interviews, participants' responses displayed great similarity. Further successive interviews generated little new data. Therefore, based on the phenomenon mentioned in the previous studies, only 6 participants were selected and interviewed in the current study.

3.3.1.1 Background Information of Stage II Participants

The interviewees in this study were 6 university freshmen, not involved in Stage I study, who varied in gender and fields of study (see Table 3.5). The interviewees also varied in their English proficiency level. Their average score on the OEAS Reading Test was 20.83 out of 36, with an SD of 6.46. To be more consistent with the Stage I standard, the interviewees were further classified into different English proficiency levels based on the cut points in the Stage I study. In other words, interviewees who scored in the top 25% were classified as having high English proficiency level and those who scored in the bottom 25% were classified as having low-level. The others were grouped as intermediate level students, as shown in Table 3.5

Table 3.5

Stage II interviewees' characteristics

Student	Major	Gender	Proficiency Level
S1	English	Male	high
S2	Environmental Science and Engineering	Male	low
S3	Public Management and Policy	Female	intermediate
S4	Music	Female	high
S5	Computer Science	Male	intermediate
S6	Chemistry	Female	intermediate

3.3.2 Data Collection Procedures

In Stage II, the researcher employed semi-structured interviews to get the participants' oral report of their perceptions of the usefulness of the Reading Test feedback in depth. It is widely mentioned that qualitative data can be used to supplement, validate, explain, illuminate, or reinterpret quantitative data gathered from the same subject or site (Miles & Huberman, 1994). The use of interview "allows for greater depth" than other methods of data collection (Cohen & Manion, 1994, p. 272). Therefore, semi-structured interviews were adopted by the researcher to get a better understanding and to collect more explicit information about participant's perceptions about the usefulness of the Reading Test feedback in depth.

The interviews were scheduled over a period of about two weeks. At first, all the participants signed up for the most appropriate time for them to have the interview. At the scheduled time, each participant was interviewed one by one in a professor's office. In order to maintain consistency, the researcher conducted all the interviews. Each session of exam and interview lasted 1.5~2 hours.

The interviews started with greetings to create an easy and friendly atmosphere to ease interviewees' worry and anxiety. Since understanding how the interviewee thinks is at the center of the interview (Bogdan & Biklen, 2003), it is crucial to have interviews in which the interviewees are at ease and talk freely about their point of view (Brigg, 1986). Then, the researcher introduced the purpose of the study and the procedures (see Table 3.6) to follow. Each participant was asked to fill in a basic personal background questionnaire and then sit down at a computer running the OEAS test platform and take the Reading Test. After taking the reading test, the participant was asked to read each type of feedback and to say whether it is useful and explain why or why not. Meanwhile, the researcher took interview notes. In addition, based on the participant's answer and comments, the researcher asked some follow-up questions to elicit more information (see Table 3.7).

The whole process of the interviews was conducted in Chinese and was audio taped for transcription into written language for further analysis. In the meantime, the interview process was also video taped using a screen capture program – *Camtasia Studio*, since interviewees might refer to the item or feedback shown on the screen when they responded to the interview questions. Therefore, the use of *Camtasia Studio* would be helpful for transcription and data analysis. In this way, the researcher was able to double-check and make sure what the interviewees were exactly referring to during the interview process.

Table 3.6

Steps of Stage II data collection

Step	**Descriptions**
1	The researcher explained the purpose of the study and give instructions on using the OEAS program.
2	The participants filled in a basic personal background information questionnaire.
3	The participants took the OEAS Reading Test for 40 minutes
4.	The participants read the test feedback and explained in Chinese whether the test feedback was useful or not.
5.	The researcher asked participants follow-up questions.

Table 3.7

Sample follow-up questions for interviews

1. Do you like the test feedback provided in the Reading Test? Why or why not?
2. Do you like the test feedback displayed in English and Chinese side by side? Why or why not?
3. Do you think the vocabulary lists provide you with enough information?
4. Is there any other type of feedback you might like to receive in the Reading Test feedback?

3.3.3 Data Analysis Procedures

The qualitative data collected from the interviews were selectively transcribed and translated; for example, content not related to the study (e.g., greetings and chitchat) was not included. In qualitative analysis, researchers have to look for commonalities, regularities, or patterns across the various data (Brown, 2001; Seliger & Shohamy, 1989). To deal with the data, the researcher in the current study reviewed and categorized the interview responses with an aim to focus on patterns of useful information. Bogdan and Biklen (2003) also pointed out that developing a coding system is a crucial step in qualitative data analysis. In this study, the coding scheme used to categorize the participants' responses is based on the perspective held by subjects. That is, the participants' thoughts or opinions which show commonalities toward a certain aspect of the Reading Test feedback were assigned into the same categories. Table 3.8 shows examples of coding categories in this study. Then, the categorized data were analyzed for recurring patterns or themes in order to answer RQ2 and RQ3.

To check inter-coder reliability, two coders independently coded the qualitative data. The researcher first coded the entire set of interview transcripts. After that, two transcripts (33% of the total) were double coded by a second coder who was sufficiently trained and familiar with the task. The researcher first gave the coder transcripts with underlined parts but no code labels. Then the coder labeled those underlined parts. The inter-coder reliability was calculated by the percentage of agreement between the two coders and the result indicated a high level of agreement (84 %). Therefore, the coding scheme used in the current study was quite reliable.

To have a more comprehensive understanding of the data collected, videos of each participant's computer screen capture were checked while analyzing the data. This was to make sure which part of the test feedback the participants were exactly referring to when they responded to the interview questions (see 3.3.2).

Table 3.8

Examples of coding categories

1. Participants' attitude toward having an overall report categorized into six reading sub-skills
2. Participants' attitude toward having English reading passages again
3. Participants' attitude toward Chinese translation of each reading passage
4. Participants' attitude toward vocabulary list
5. Participants' attitude toward general explanations of each sub-skill
6. Participants' attitude toward specific explanations for each test item
7. Participant's proficiency levels and their attitude towards the feedback
8. Participants' overall opinions about the Reading Test
9. Participants' difficulties with feedback

3.4 Validation

Various methods were adopted to increase the validity of the findings in this study. First, the researcher employed a two-stage mixed methods approach which aimed to probe into the research problems by combining quantitative and qualitative data. Explaining the quantitative statistical results with qualitative data helped to provide detailed or in-depth information to understand the research problems. Also, the data from one source could help confirm or disconfirm the tentative findings attained from the other source of data. Second, to assure the validity of the questionnaire, a professor in the MA program in Teaching English as a Foreign Language (TEFL) was invited to review the questionnaire items to give expert advice as well as some suggestions on revision of the items. Third, to increase inter-coder reliability and analyze collected data in a consistent manner, a coder was invited to examine part of the interview data and code the data based on the researcher's coding categories. Then, the coder's coding was compared with that of the researcher and the results showed a fairly high level of agreement (see 3.3.3).

3.5 Summary of the Research Design

The purpose of this study is to examine university students' perceptions about the feedback of the OEAS Reading Test. To answer the three research questions, the present study adopted a two-stage mixed methods approach to collect both quantitative and qualitative data. In Stage I, a questionnaire was administered to collect quantitative data targeting at answering RQ1 and RQ3. On the other hand, the Stage II research employed a semi-structured interview to gather qualitative data with an aim to answer research RQ2 and RQ3 (see Table 3.9).

Table 3.9

Summary of the research design

Research Questions	Type of data collected	Type of analysis
RQ1: How useful do test-takers perceive the Reading Test feedback to be?	questionnaire	quantitative
RQ2: In what ways do the test-takers perceive the feedback to be useful or not?	interview	qualitative
RQ3: Is there a difference between test-takers of low and high English proficiency levels in how they perceive the feedback's usefulness?	questionnaire and interview	quantitative and qualitative

3.6 Pilot Study

The purpose of the pilot study was mainly to check the feasibility of the methodology adopted in this study and to check the reliability and the validity of the questionnaire for revisions, if necessary, and find out potential problems, if any. Similar to the main study, the pilot study was conducted in two stages. In Stage I, 6 university freshmen took the OEAS Reading Test and then filled in the Chinese version of the questionnaire (see Appendix A) in March, 2009. In Stage II, 2 university freshmen were individually interviewed as soon as they finished the OEAS Reading Test.

3.6.1 Stage I Methodology and Results for Pilot Study

In Stage I, 6 university freshmen from the Tunghai FENM program (see Table 3.10) were invited to participate in the pilot study. These students all took the OEAS Reading Test in a computer classroom and then filled in the questionnaire (see 3.2.2 for detail). The questionnaire consisted of two sections: Part I includes 5 items about personal background information, and in Part II, a survey containing 49 items in eight categories is used to investigate students' thoughts about Reading Test feedback (see Table 3.3).

Table 3.10

Participants of Stage I pilot study

Student	**Major**	**Gender**
S1	Business Administration	Female
S2	Computer Science	Male
S3	Social Work	Female
S4	Finance	Female
S5	Hospitality Management	Male
S6	Animal Science and Biotechnology	Female

The data collected were computed and analyzed using SPSS 17.0. Cronbach's alpha was calculated for the 49 items in Part II, and reliability reached 0.965, which means the questionnaire adopted in this study is very reliable. Means of each category in Part II were calculated and the results are shown in Table 3.11 (see Appendix D for detail). First, the mean for the first category, overall test feedback report, is 4.00 (SD= .632). The result shows that the pilot test-takers considered the overall test report very useful. Second, means for the feedback on each reading sub-skill fell between 3.83 and 4.27. The result indicates that test-takers had a positive attitude towards the usefulness of each reading sub-skill feedback. Third, the overall mean for test-takers' perceptions about general aspects of Reading Test feedback was 4.08. To sum up, the test-takers' in general regarded the Reading Test feedback to be very useful.

Table 3.11

Means and standard deviations of the eight categories of Part II questionnaire for pilot study (n=6)

Categories	Item Number	Means	SD
Overall test feedback report	A1	4.00	.632
Reading sub-skill 1	B1-1~B1-5	3.83	.581
Reading sub-skill 2	B2-1~B2-9	3.88	.666
Reading sub-skill 3	B3-1~B3-9	4.06	.693
Reading sub-skill 4	B4-1~B4-8	3.92	.635
Reading sub-skill 5	B5-1~B5-6	3.89	.855
Reading sub-skill 6	B6-1~B6-5	4.27	.682
Perceptions about general aspects of feedback	B7-1~B7-6	4.08	.750

3.6.2 Stage II Methodology and Results for Pilot Study

In Stage II, 2 university freshmen, not involved in Stage I, were interviewed respectively. Each interview was about 1.5 hours. At first, the interviewees had to take the OEAS Reading Test. Then, the researcher asked the interviewees questions based on the questionnaire and the interviewee orally expressed their opinions and offer additional comments, if any. In addition, interviewees were asked some follow-up questions to clarify their opinions and elicit further information about their perceptions of the test feedback.

The results of the interview showed that both interviewees considered the Reading Test feedback to be useful and the results were consistent with the findings of the Stage I pilot study. Only one questionnaire item was added since one interviewee mentioned that it might be helpful if sample sentences were provided along with the vocabulary. Therefore, the item “我認為若能提供單字表中所列出的單字例句是有幫助的” (I think it would be helpful if sample sentences along with the vocabulary could be provided.) was added to the questionnaire.

3.7 Summary of the Chapter

In the beginning of this chapter, the research questions and overall research design of this study are presented. This study employed a two-stage mixed methods research design; therefore, participants, instruments, data collection procedures, and data analysis procedures are introduced respectively for each of the two stages. Furthermore, a pilot study was conducted to check the feasibility of the methodology and the reliability of the questionnaire adopted in this study. The results of the pilot study showed that the overall research design is suitable and the questionnaire, with Cronbach’s alpha of 0.965, is quite reliable. In addition, the participants in general regarded the Reading Test feedback to be very useful. After the pilot study, one more questionnaire item was added for the formal study.

CHAPTER FOUR

RESULTS

This chapter presents results of the study described in Chapter 3. The results include 1) the quantitative analysis of the Stage I questionnaire; and 2) qualitative results of the Stage II interviews.

4.1 Results for Stage I Study

This section presents analysis results of all the items on the *Survey of Students' Perceptions of the OEAS Reading Test Feedback*, including 1) high- and low-English proficiency groups, 2) university students' perceptions of the reading test feedback, and 3) the top- and bottom-10 rated feedback items.

4.1.1 High- and Low-English Proficiency Groups

Stage I results of the study was based on 48 students' responses to a paper-and-pencil questionnaire – *Survey of Students' Perceptions of the OEAS Reading Test Feedback* (see Appendix G for details), which participants completed after taking the OEAS Reading Test. The participants were recruited by three major means (see 3.2.1 for details), including students recruited from the Tunghai Freshman English for Non-Majors (FENM) program, English majors from the Foreign Languages and Literature Department (FLLD), and respondents to an advertisement on the website of the English Language Center. The 48 participants in Stage I, were all university freshmen but varied in genders and their field of study (see Table 3.2 for details).

The results the OEAS Reading Test also confirmed that the participants in Stage I varied in their English reading comprehension level. Stage I participants' average score on the OEAS Reading Test was 17.08 out of 36, with an SD of 5.9 and they were normally distributed. In order to examine whether there were significant differences in the participants' perceptions of the OEAS Reading Test feedback related to differences in English proficiency level, participants were divided into low-scoring and high-scoring groups based on the test results of the OEAS Reading Test. Test-takers whose OEAS Reading Test scores were in the bottom 25% (n=12) were grouped and defined as having a low English proficiency level and in the top 25% (n=12) as having a high English proficiency level (see Table 4.1). There was a significant difference between the high- and low level groups on their OEAS Reading Test scores (24.33 vs. 9.33, t= 12.94[1], sig. at p＜ 0.05).

Table 4.1

Stage I participants' average OEAS Reading Test scores

	Mean	SD
All test-takers (n=48)	17.08	5.90
Low-level (n=12)	9.33	2.93
High-level (n=12)	24.33	2.74

4.1.2 Students' Perceptions of the OEAS Reading Test Feedback

Part II of the Questionnaire of Students' Perceptions of the OEAS Reading Test Feedback, including fifty 5-point Likert scale items, was used to investigate the participants' perceptions of usefulness of the feedback provided by the OEAS Reading Test. The reliability of the whole questionnaire was calculated, resulting in a Cronbach's alpha of 0.95, which indicated the questionnaire was fairly reliable.

[1] For the purpose of this study's analysis, it was assumed that assumptions for using a T-test were fulfilled.

4.1.2.1 Students' Perceptions of the Overall Test Result Report

The first item (Item A1) in Part II of the questionnaire was to examine students' thoughts about the overall test result report. As Table 4.2 shows, test-takers generally perceived the overall test result report to be useful, with an average rating of 3.60. In addition, the high-level test-takers viewed the overall test report as slightly more useful (3.83) than the low-level test-takers did (3.67). However, no significant difference was found between high- and low-level test-takers' perceptions of the overall test report (3.83 vs. 3.67, t= 0.66, n.sig. at p＜0.05, two-tailed).

Table 4.2

Average ratings of the overall test result report (Item A1)

	Frequency (# of responses)					Mean	SD
	1[a]	**2**	**3**	**4**	**5**		
All test-takers (n=48)	0	2[b]	17	27	2	3.60	0.64
Low-level (n=12)	0	0	4	8	0	3.67	0.49
High-level (n=12)	0	0	4	6	2	3.83	0.72

Note:
a. 1= not helpful, 2= not quite helpful, 3= somewhat helpful, 4= helpful, 5= very helpful
b. Frequencies were based on the total responses for that item.

4.1.2.2 Students' Perceptions of the General Explanations of the Six Reading Sub-skills

As mentioned in the previous chapter (see 2.3.4.3), the OEAS Reading Test aimed to test six reading sub-skills: 1) reading for main ideas of a passage, 2) reading for main ideas of a paragraph, 3) reading for specific information, 4) guessing meaning of vocabulary from context, 5) pronoun reference, and 6) inference from reading. Hence, the test feedback was also provided accordingly.

Table 4.3 presents the means (M), standard deviations (SD), and frequency of participants' responses to the "General Explanation" for each of the six reading sub-skills tested, including Items B1-1, B2-1, B3-1, B4-1, B5-1, and B6-1. With a grand mean of 3.84, it indicates the participants as a whole seemed to consider the General Explanations to be fairly useful. Among these six items, participants as a whole considered Item B4-1 (General Explanation for reading sub-skill 4: guessing meaning of vocabulary from context) to be most useful (M= 4.02) and Item B2-1 (General Explanation for reading sub-skill 2: reading for main ideas of a paragraph) to be least useful (M= 3.67). The perceptions of the high-level test-takers were also consistent with the above results (Item B4-1, M= 4.33 vs. Item B2-1, M= 3.67), while the low-level test-takers thought Item B6-1 (General Explanation for reading sub-skill 6: inference from reading) to be least useful.

Table 4.3

Descriptive statistics of general explanations for 6 reading sub-skills

(Overall, n=48; Low-level, n=12; High-level, n=12)

No.	**Item Description** I think the **General Explanation** for _______ is helpful.	**Proficiency groups**	**Frequency (# of responses)**					**M**	**SD**
			1	**2**	**3**	**4**	**5**		
B1-1	Reading sub-skill 1 (Reading for main ideas of a passage)	Overall	0	0	11	35	2	3.81	.49
		Low-level	0	0	3	9	0	3.75	.45
		High-level	0	0	2	10	0	3.83	.39
B2-1	Reading sub-skill 2 (Reading for main ideas of a paragraph)	Overall	0	0	20	24	4	3.67	.63
		Low-level	0	0	3	9	0	3.75	.45
		High-level	0	0	4	8	0	3.67	.49
B3-1	Reading sub-skill 3 (Reading for specific information)	Overall	0	2	11	28	7	3.83	.72
		Low-level	0	1	3	5	3	3.83	.94
		High-level	0	0	5	5	2	3.75	.75
B4-1	Reading sub-skill 4 (Guessing meaning of vocabulary from context)	Overall	0	1	11	22	14	4.02	.79
		Low-level	0	0	4	5	3	3.92	.79
		High-level	0	0	1	6	5	4.33	.65
B5-1	Reading sub-skill 5 (Pronoun reference)	Overall	0	2	12	24	10	3.88	.79
		Low-level	0	0	5	4	3	3.83	.84
		High-level	0	0	0	9	3	4.25	.45
B6-1	Reading sub-skill 6 (Inference from reading)	Overall	0	1	14	26	7	3.81	.70
		Low-level	0	1	4	6	1	3.58	.79
		High-level	0	0	1	7	4	4.25	.62
	Average	**Overall**						**3.84**	**.69**
		Low-level						**3.78**	**.76**
		High-level						**4.01**	**.71**

In the main, the high-level test-takers perceived the General Explanations to be more useful than the low-level test-takers did (4.01 vs. 3.78). Regarding the frequency of responses, Table 4.3 also shows that a larger number of high-level test-takers considered the General Explanations to be useful. In addition, a significant difference was found between high- and low-level test-takers' perceptions of the general explanations for reading sub-skill 6 (4.25 vs. 3.58, t=2.29, sig. at $p < 0.05$, two-tailed).

Because the high- and low-level test-takers' perceptions of Item B6-1 (General Explanation for reading sub-skill 6: inference from reading) showed a significant difference, the researcher further explored the relationship between the participants' perceptions of the General Explanations of the OEAS Reading Test feedback and the participants' reading test scores. A positive significant Spearman's correlation coefficient was found between participants' perceptions of General Explanation of reading sub-skill 6: inference from reading (Item B6-1) and their reading test scores (r= 0.42). at $p < .05$. In other words, test-takers with higher reading test scores tended to give the general explanation of reading sub-skill 6 – inferencing from reading – higher usefulness ratings.

4.1.2.3 Students' Perceptions of the Specific Item Explanations of the Reading Test

This section presents how helpful test-takers perceived the specific item explanations of the six reading sub-skills to be. Table 4.4 presents the average means (M), standard deviations (SD), and frequency of participants' responses to the "Specific Explanations" for the 36 reading test items covered in the six reading sub-skills.

In general, participants considered the specific item explanations to be quite useful, with an average mean of 3.70. As Table 4.4 shows, both participants as a whole (M=3.76) and the low-level test takers (M=3.78) considered specific explanations for reading sub-skill 3 (Item B3-2~B3-9) to be most useful while the high-level test-takers regarded those to be least useful (M=3.77). And the high-level test-takers perceived specific explanations for reading sub-skill 5 (Item B5-2~B5-6) to be most useful. Again, as with the general explanations (see 4.1.2.2), the high-level test-takers gave comparatively higher usefulness ratings to the "Specific Explanations" than the low-level test-takers did and this difference was statistically significant (3.86 vs. 3.66, t=3.85, sig. at p＜0.05, two-tailed).

Table 4.4

Descriptive statistics of specific explanations for 6 reading sub-skills

(Overall, n=48; Low-level, n=12; High-level, n=12)

No.	**Item Description** I think the **Specific Explanation** for _______ is helpful.	**Proficiency groups**	**Frequency (# of responses)**					**M**	**SD**
			1	**2**	**3**	**4**	**5**		
B1-2~ B1-5	Reading sub-skill 1 (Reading for main ideas of a passage)	Overall	0	5	56	114	17	3.75	.64
		Low-level	0	3	11	28	6	3.77	.73
		High-level	0	0	13	32	3	3.79	.56
B2-2~ B2-9	Reading sub-skill 2 (Reading for main ideas of a paragraph)	Overall	0	20	123	206	35	3.67	.72
		Low-level	0	5	28	59	4	3.65	.64
		High-level	0	5	19	63	9	3.79	.66
B3-2~ B3-9	Reading sub-skill 3 (Reading for specific information)	Overall	2	14	115	198	55	3.76	.75
		Low-level	2	3	25	51	15	3.78	.82
		High-level	0	3	27	55	11	3.77	.69
B4-2~ B4-8	Reading sub-skill 4 (Guessing meaning of vocabulary from context)	Overall	0	18	103	172	43	3.71	.76
		Low-level	0	5	27	50	2	3.57	.64
		High-level	0	2	16	50	16	3.94	.72
B5-2~ B5-6	Reading sub-skill 5 (Pronoun reference)	Overall	0	15	72	118	35	3.72	.79
		Low-level	0	42	21	28	7	3.63	.77
		High-level	0	3	10	34	13	3.95	.75
B6-2~ B6-5	Reading sub-skill 6 (Inference from reading)	Overall	0	17	62	91	22	3.62	.80
		Low-level	0	5	12	23	6	3.58	.90
		High-level	0	0	10	31	7	3.94	.58
	Average	**Overall**						**3.70**	**.74**
		Low-level						**3.66**	**.75**
		High-level						**3.86**	**.66**

Note:

a. Frequencies were based on the total responses for items covered in that sub-skill.

4.1.2.4 The Top-10 and Bottom-10 Rated Specific Explanations

This section presents the distribution of the top- and bottom-10 rated specific item feedback provided by the OEAS Reading Test. The top- and bottom 10 rated feedback were based on all Stage I participants' responses on the usefulness of the feedback. In order to further explain why test-taker's perceived some feedback to be more useful than others, the following paragraphs discuss some shared characteristics and possible factors relating to test-takers' preferences towards certain feedback.

Table 4.5 shows the top- and bottom-10 rated specific item feedback provided by the OEAS Reading Test along with respective overall frequencies, means, standard deviations, item difficulty, and length of Chinese and English explanation of each item. As seen in Table 4.5, the average means of the top- and bottom-10 rated specific item feedback were 3.82 and 3.59 respectively. There was a significant difference found between the usefulness ratings of top- and bottom-10 rated feedback (3.82 vs. 3.59, t= 8.37, sig. at $p < 0.05$, two-tailed). Table 4.6 presents examples of specific item explanations with high and low ratings.

Length of explanations was one characteristic related to students' preferences toward the reading test feedback. In this study, the length of each explanation was calculated respectively for Chinese and English feedback by using Word Count in Word for Windows. As shown in Table 4.5, the average length of Chinese explanations of the top-10 rated item was shorter than that of the bottom-10 rated item explanation (144 vs. 252), and this difference was statistically significant (t=-2.27, sig. $p < 0.05$, two-tailed). Similarly, a significant difference was found between the length of English explanations of the top-10 and bottom-10 rated items (97 vs. 159, t=-2.27, sig. $p < 0.05$, two-tailed). That is, participants in Stage I perceived feedback with shorter length to be more useful. Potential reasons are discussed in Chapter 5.

Table 4.5

Analysis of top-10 and bottom-10 rated specific explanations (n=48)

Item	Frequency (# of responses)					Average usefulness rating	SD	ID[a]	Length of Chinese explanation[b]	Length of English explanation	Test item	Testlet [c]
Top-10 rated items	**1**	**2**	**3**	**4**	**5**							
B33	0	1	11	26	10	3.94	.73	.81	178	138	Q5	1
B39	0	3	8	28	9	3.90	.78	.33	121	82	Q30	4
B14	0	1	10	32	5	3.85	.62	.65	185	135	Q17	3
B32	0	0	16	24	8	3.83	.69	.40	221	174	Q4	1
B52	0	3	12	23	10	3.83	.83	.54	25	20	Q7	1
B44	0	3	12	25	8	3.79	.80	.67	158	101	Q22	3
B15	0	2	14	24	8	3.79	.77	.58	77	43	Q26	4
B28	0	1	14	28	5	3.77	.66	.45	242	128	Q27	4
B34	0	2	13	28	5	3.75	.70	.54	135	80	Q12	2
B38	0	3	15	23	8	3.75	.81	.48	99	67	Q29	4
Average						**3.82**	**.74**	**.55**	**144**	**97**		
Bottom-10 rated items	**1**	**2**	**3**	**4**	**5**							
B12	0	1	18	27	2	3.63	.61	.42	303	209	Q1	1
B22	0	2	18	24	4	3.63	.70	.56	413	286	Q2	1
B26	0	3	17	23	5	3.63	.76	.38	150	101	Q18	3
B54	0	3	17	23	5	3.63	.76	.21	93	73	Q24	3
B62	0	5	17	17	9	3.63	.91	.31	481	155	Q8	1
B35	0	3	15	23	7	3.60	.68	.65	185	128	Q13	2
B45	0	4	16	24	4	3.58	.77	.48	161	138	Q23	3
B23	0	4	19	19	6	3.56	.82	.77	376	264	Q3	1
B36	1	1	20	22	4	3.56	.77	.38	124	86	Q20	3
B64	0	4	22	19	3	3.44	.74	.10	230	149	Q25	3
Average						**3.59**	**.76**	**.43**	**252**	**159**		

a. ID= item difficulty

b. Length is in terms of Chinese characters; an English word used in the Chinese explanation was counted as one word as well.

c. Testlet 1: Promoting Earthworms; 2: Introducing Astrology; 3: Electricity and Fish; 4: Industrial Revolution

Table 4.6

Examples of specific item explanations with high and low ratings

Item number & sub-skill tested	Item	Rating of explanation	Key & item explanation in English and Chinese
B33 (Q5) sub-skill 3: reading for specific information	Which of the following statements is NOT true? A. Mr. Weigel's wife's houseplants have earthworms in their pots. B. Earthworms generally live around plant roots. C. Earthworms produce good soil faster than nature. D. Fertilizers have no effect on earthworms.(正解)	3.94 (high)	Specific explanation: 詳解 a. is a true statement according to the passage and so is not the correct answer. We know the statement is true from the second sentence in paragraph 2, "He has persuaded his wife to put worms in her houseplant pots." 根據文章，a 為真實的陳述，故不是正確答案。我們從第二段第二句 " 他說服他妻子放些蚯蚓在她的盆栽裏"得知此句為真。 b. is a true statement according to the passage and so is not the correct answer. The second sentence of paragraph 4 is: For example, worms tend to live mostly around the roots of plants. 根據文章，b 為真實的陳述，故不是正確答案。第四段第二句為"例如，蚯蚓多半傾向於住在植物的根周圍" 。 c. is a true statement according to the passage and so is not the correct answer. We know this from the last sentence in paragraph 3. 根據文章的第三段最後一句可知，c 為真實的陳述，故不是正確答案。 d. is not true according to the passage and so is the correct answer. It states in the second sentence of the last paragraph that earthworms have been killed by fertilizers and other farming chemicals. 根據文章，d 不是真實的陳述，故是正確答案。最後一段第二句中提到"蚯蚓會被肥料及其它農藥所殺滅" 。
B39 (Q30) sub-skill 3: reading for specific information	According to the passage, which of the following statements is NOT true? A. Before the Industrial Revolution most people lived in the same village their whole lives. B. Factory-owners during the Industrial Revolution were known to treat workers very well.(正解) C. Electricity was not available in rural Great Britain at the start of the Industrial Revolution. D. Putting together machines was a job performed in factories during the Industrial Revolution.	3.90 (high)	Specific explanation: Choice A is not the correct answer because this idea is mentioned in paragraph 2, sentence number 3. Choice C is not the correct answer because this idea is mentioned in the last sentence of paragraph 3. Choice D is also not the correct answer because it mentions twice in paragraph 5 that machines were made in factories. Choice B is the correct answer. It states in the second sentence of paragraph 5 that there were many factory-owners who cheated their workers. 詳解： A 選項不是一個正確的選擇，因為這個概念在第二段的第三句提過。C 選項不正確，因為這個概念在第三段的最後一句也提到了。D 選項也是不正確的，因為機器是工廠所製造在第五段已提過兩次。B 選項為正確答案；在第五段的第二句有提到，有許多的工廠主人會欺騙他們的員工。

B64(Q25) sub-skill 6: inference from reading	What can be inferred about electric eels? A. They are related to torpedoes. B. They are not related to the torpedoes. C. Scientists will find a way to use them as a source of energy. D. Horses are not seriously injured during their capture in South America.(正解)	3.44 (low)	Specific explanation: 詳解 A. No. Even though they both use electricity for defense and protection, one cannot assume that they are related species. In fact, they are several clues that that may not be related. For example, torpedoes are found in salt water and electric eels in freshwater and both have very different ways of producing electricity. We do not have enough information to make this inference. A. 錯誤。即使二種都用電來防衛及保護，我們不能假定它們為相近的品種。事實上，有不少線索指出它們或許不相似。例如，電魟在鹽水域中被找到，而電鰻則在淡水域，並且兩者發電的方式大不相同。我們沒有足夠的資訊來做這種假設。 B. No. Even though A) may not be true, we still cannot assume B). Both generate electricity and are fish which may mean that they are related in some way. We do not have enough information to make this inference. B. 錯誤。雖然 A 句錯誤，我們也不能假設 B 句是對的。同樣會發電且又都是魚類，或許代表二者在某方面上是相似的。我們沒有足夠的資訊來做這種假設。 C. No. There is no mention of electric eels being a potential source of energy in the passage. C. 錯誤。文章中沒有提到電鰻可以成為能源的來源。 D. Yes. One can assume that horses are valued higher than electric eels by South Americans, and as such, they would not risk serious injury to their horses. D. 正確。我們可以假定，在南美，馬的價值高於電鰻，因此，人們不會冒險讓馬受到重傷。
B36(Q20) sub-skill 3: reading for specific information	According to the passage, why must a person touch a torpedo in two places to get a shock? A. because current passes from head to tail B. because the fish is negative and you are positive C. because the electric plates are flat D. because otherwise the circuit is not complete(正解)	3.56 (low)	Specific explanation:詳解 Note: You should immediately find the paragraph pertaining to torpedoes (paragraph 2). 注意：你應該馬上找到有關於電魟的文章(第二段)。 A. This is false. This is how current passes in an electric eel (paragraph 3) A 此句是錯的。這種電流的走向是電鰻。(第三段) B. This is not mentioned in paragraph 2. B 此句並未在第二段中提及。 C. This is in paragraph 2 and is true, but it does not explain why it will cause a person to get a shock. C 此句出現在第二段而且正確。但此句並未解釋為何它會使人感電。 D. The correct answer. [“Generally it is necessary to touch the fish in two places, which then completes the circuit, in order to receive an electric shock”] D 正確答案。(大致上來說，為了要通電，電魟的二個地方要被碰到，用來完成電流迴路，是十分必要的)

With regard to item difficulty, the average item difficulty of the top-10 items was higher than that of the bottom-10 items (0.55 vs. 0.43). The result indicated that the test items of the top-10 rated feedback were inclined to be easier than those of the bottom-10 rated feedback. Though no significant difference was found (0.55 vs. 0.43, t= 1.54, n.sig. at p＜0.05, two-tailed), the result was consistent with the overall tendency stated below.

Table 4.7 summarizes the Spearman's correlation analysis of the link between the participants' perceptions of the usefulness of the OEAS Reading Test feedback and the item difficulty of the reading test. As seen in the table, a positive significant Spearman's correlation coefficient (r= .351) was found at p＜.05. In other words, feedback of test items with higher item difficulty tended to receive higher usefulness ratings. That is, Stage I participants considered feedback of easier test items to be more useful.

Table 4.7

Correlation between overall usefulness ratings and item difficulty (n=48)

Variables	Means	SD	Spearman's Correlation Coefficient	Sig.
Usefulness	3.71	.74	.351	.037
Item difficulty	.47	.17		

* correlation is significant at the 0.05 level (2-tailed)

4.1.2.5 Students' Perceptions of the General Aspects of the Reading Test Feedback

The last section of the Part II questionnaire consisted of 7 items with an aim to investigate participants' thoughts about the general aspects of the OEAS Reading Test Feedback. As shown in Table 4.8, the participants as a whole perceived these aspects to be very useful. Among these 7 items, Item B7-1 had the highest mean (4.29) and the lowest SD (.71). That is, most participants consistently expressed that it is helpful to have the original reading passages again when they read the OEAS Reading Test feedback. Meanwhile, Item B7-4 had the lowest mean (3.90) and the highest SD (.97). In other words, though the participants thought it might be helpful to have sample sentences along with the vocabulary list, their opinions varied.

Table 4.8

Descriptive statistics of "general aspects" of the OEAS Reading Test Feedback

(Overall, n=48; Low-level, n=12; High-level, n=12)

No.	**Item Description**	**Proficiency groups**	**Frequency (# of responses)**					**M**	**SD**
			1	**2**	**3**	**4**	**5**		
B7-1	How helpful is having the reading passages again?	Overall	0	1	4	23	20	4.29	.71
		Low-level	0	1	0	5	6	4.33	.89
		High-level	0	0	0	6	6	4.50	.52
B7-2	How helpful is having Chinese translation of the passage?	Overall	0	0	8	19	21	4.27	.74
		Low-level	0	0	3	0	9	4.50	.91
		High-level	0	0	2	5	5	4.25	.75
B7-3	How helpful is having vocabulary list of the reading passage?	Overall	0	1	6	21	20	4.25	.76
		Low-level	0	1	0	3	8	4.50	.91
		High-level	0	0	2	3	7	4.42	.79
B7-4	How helpful if sample sentences are provided along with vocabulary list?	Overall	0	5	10	18	15	3.90	.97
		Low-level	0	1	2	4	5	4.08	.99
		High-level	0	1	2	4	5	4.08	.99
B7-5	In general, how helpful are the English explanations?	Overall	0	0	13	20	15	4.04	.77
		Low-level	0	0	2	7	3	4.01	.87
		High-level	0	0	3	3	6	4.25	.67
B7-6	In general, how helpful are the Chinese explanations?	Overall	0	1	8	27	12	4.04	.71
		Low-level	0	1	2	4	5	4.08	.99
		High-level	0	0	2	8	2	4.00	.60
B7-7	How helpful is it to have bilingual feedback together?	Overall	0	0	15	16	17	4.04	.82
		Low-level	0	0	4	2	6	4.17	.94
		High-level	0	0	3	4	5	4.17	.84
	Average	**Overall**						**4.12**	**.78**
		Low-level						**4.19**	**.93**
		High-level						**4.24**	**.74**

Participants as a whole perceived the English explanations, Chinese explanations, and bilingual feedback to be equally important (M= 4.04). When it comes to students of different English proficiency levels, it was found that low-level students considered both Chinese translation and Chinese explanations to be more useful than English explanations. And they valued the Chinese feedback more than the high-level students did. On the other hand, the high-level students viewed English explanations as more useful than Chinese ones (4.25 vs. 4.00). Further examination showed no significant differences between high-and low-level students' perceptions of usefulness regarding English (Item B7-5: 4.25 vs. 4.01, t= 0.81, n.sig. $p＜0.05$, two-tailed) and Chinese (Item B7-6: 4:00 vs. 4.08, t= -0.35, n.sig. $p＜0.05$, two-tailed) feedback since both groups gave these items fairly high ratings.

In summary of Stage I, the results from the questionnaire aimed to answer research questions (1) and (3). The findings indicated that the test-takers as a whole perceived the reading test feedback to be fairly useful and that different proficiency levels did differentiate test-takers in their opinions towards some types of feedback. Additionally, the Stage I findings provided partial answer to research question (2). The results showed that test-takers preferred shorter feedback and that feedback of easier test items tended to receive higher usefulness ratings.

4.2 Results for Stage II Study

To know more about students' perceptions of the usefulness of the reading test feedback, 20 interview questions were designed based on the questionnaire used in Stage I (see 3.3.2). This section presents the interview results of students' perceptions of the OEAS Reading Test feedback and it was further divided into four parts: students' perceptions of 1) the overall test result report; 2) the general explanations of each reading sub-skill; 3) the specific item explanations; and 4) the general aspects of the reading test feedback

4.2.1 Students' Perceptions of the Overall Test Result Report

All the interviewees perceived the overall test result report to be very useful. They thought the classification of the overall test report could help them understand which reading sub-skills they were weaker or stronger in (S1[2]: "We might not sense some items tested specific reading skills when we were taking the test. This overall report helped me understand my weaknesses (in reading)…", also S2, 3, 4, 5, 6). Besides, some students expressed that the overall test report could provide them with information for further improvement (S3: "The immediate test report is quite different from the usual scores we got; it gave me information about which parts (skills) I need to improve…"; also S1, 2, 4).

4.2.2 Students' Perceptions of the General Explanations of Each Reading Sub-skill

The results of the interviews indicated that there were mainly two kinds of opinions about the general explanations. Half of the students considered the general explanations of each reading sub-skill helpful because these explanations helped them understand the purposes of the test design and familiarized them with specific reading sub-skills. For example, S1 mentioned "I think it helpful because it teaches me how to read and analyze a reading text or deal with some types of test items. It would be useful when I read similar articles…" (also S2, 5). In contrast, the other students didn't view the general explanations useful and just browsed through this part. For instance, S4 expressed that "I tend to ignore this part because it teaches how to take tests"; S6 pointed out that "I knew most of these

[2] S [number] denotes which student made the comment or similar comments.

reading skills already; I care more about the questions I answered incorrectly"; also S3 stated "It's a little bit too long and it contains some terms…I don't spend too much time reading it." It seemed that these students thought general explanations were not so useful because they offered overall introduction of certain reading sub-skill but were not directly related to test items.

4.2.3 Students' Perceptions of the Specific Item Explanations of the Reading Test

All of the six interviewees perceived the specific explanations for each item to be very useful because this kind of feedback explained each distractor and helped clarify some ambiguity. For example, S3 stated "It's helpful because it's directly related to the test items and helped me understand why I answered some questions wrong and why a certain answer should be chosen" (also S1, 4, 5, 6). Moreover, S1 mentioned that the specific explanations were very clear and helped him clarify some uncertainties he had while reading (also S3, 6). In addition, S5 pointed out that specific explanations were comparatively more useful than the general explanations because specific explanations could truly reflect and clarify what he didn't understand so as to help him improve. The finding was consistent with the result of another interview question. When asked "Which is the most useful kind of feedback to you?", most interviewees (five out of six) expressed that the specific explanations were the most useful to them.

The results disagreed with the Stage I findings. As shows in Table 4.4 and Table 4.5, participants in the Stage I considered the general explanations more useful than the specific item explanations (3.84 vs. 3.70). The reason for this seemingly contradictory opinion might be that the Stage II interviewees didn't provide usefulness ratings of feedback item by item. They were asked interview questions and expressed their overall feeling of usefulness of the general explanations and specific item explanations.

4.2.4 Students' Perceptions of the General Aspects of the Reading Test Feedback

In general, students all thought it helpful to have original reading passages again because in this way students could match the explanations with the contents of the reading texts and check their own understanding (S4: "Having the reading passages again makes it easier for me confirm or clarify my own thinking when I read the explanations…"; also S1, 2, 3). Though students considered the equivalent Chinese translation of reading passages useful, most of them only read part of the Chinese translation (S3: "I didn't read through the translation, I only use it to check the meaning of certain sentences," also S4, 5, 6). As for the vocabulary list, students as a whole viewed it to be very helpful (S4: "It helped me confirm my guessing while I read..," and S3: "It was useful because it was provided immediately after the reading test and I could learn some words that I didn't understand"). In addition, students (S1, 3, 4, 5, 6) thought they could check the usage of specific vocabulary in the reading passages and there was no need to provide sample sentences along with the vocabulary list.

More information of students' perceptions of the general aspects of the reading test feedback is presented below.

a. Chinese VS. English Feedback

Generally speaking, students reported that it is helpful to have bilingual explanations. However, they did express different opinions toward Chinese and English feedback. Most students directly went to Chinese explanations when reading feedback. The reason might be as follows: first, Chinese is students' native language and they are more used to it (S2, 3, 4, 5, 6); second, English explanations contain new words which may in turn increase the difficulty of reading the feedback (S3, 5); and third, the Chinese explanation is clear enough (S6). Though most students seemed to prefer Chinese explanations to English ones, the high-level students (S1, 4, see Table 3.5) stated that English explanations were clearer in terms of explaining the content of English reading comprehension and enabled them to learn new words.

b. The Length of Feedback

The OEAS Reading Test provided test-takers with enriched feedback and students did express some opinions concerning their preference for the length of feedback. All students thought that explanations should be clear and explicit. In addition, they should be able to clarify students' confusion or misunderstandings (S1, 2, 3, 5). Furthermore, half of the students reported that they tended to skip or skim longer feedback since they had just completed four long passages in the reading comprehension test (S2, 5, 6) Meanwhile, a majority of students indicated that they preferred feedback that was clear and brief. This was in accordance with the phenomenon found in Stage I, which indicated that students had preference for feedback with shorter length (see 4.1.2.4). When asked about the proper length of a feedback, some students (S3, 4, 6) stated that feedback not exceeding three lines would be more readable and appropriate.

c. Students' Way of Reading Feedback

Interviewees commonly expressed that they didn't read the feedback one by one. They mainly focused on explanations of the items that they answered incorrectly or those they didn't understand while taking the reading test. All students tended to selectively read the explanations of the items answered incorrectly because they thought it was more beneficial, some even thought that reading the explanations of the items that they answered correctly seemed to be a waste of time (S3, 5).

d. Students' Difficulties towards Reading Feedback

Based on the results of the interviews, students' difficulties with reading feedback could be classified into two parts: content and format of the feedback.

Concerning the content of the feedback, students pointed out some difficulties they encountered while reading the test feedback. First, the content of the feedback, especially the general explanations, contains some special terms which may interfere with students' understanding of the feedback. S1 mentioned that only students who had received instructions in reading comprehension might be familiar with the terms such as "thesis statement" and "body," and that these terms might cause extra burden to students when reading the feedback (also, S3). Second, some explanations seemed ambiguous and didn't clarify students' questions. S3 stated that a few explanations seemed vague (also S2) and she could only help her reject what she originally thought without eliminating her confusion.

When it comes to the format of the feedback, students also expressed some difficulties they had. First, most of the interviewees thought the font size was too small and this made the process of reading tiring (S1, 2, 3, 5, 6). Second, for many students, reading online material was quite different from reading printed texts and this difference caused them some difficulties while reading the test feedback on the screen. For example, they couldn't circle or underline some key words and it turned out that they had to spend more time locating the parts that the explanations were referring to (S2, 3, 4). Third, students voiced that they should be given a choice of which part of feedback that they wanted to read. Though the overall test result report was classified into six reading sub-skills and clicking on one type of the reading skills led to the other six kinds of feedback (see 2.3.4.3 for details), a test-taker was still presented with a large amount of information in the feedback for each reading sub-skill. Some students considered it time-consuming to drag all the way to certain item explanations (S2, 3, 5). Therefore, students thought that it would be more beneficial and efficient if they could click on the item number that they answered incorrectly and then be led to that specific item explanation.

e. Other Issues Regarding OEAS Reading Test and Its Feedback

Students gave other opinions of the test and feedback that did not fit easily into the above categories.

First, some interviewees expressed that the reading passages were a bit too long and thus made it hard for them to finish the reading test within 40 minutes. For example, S3 thought that the length of the reading passages should be shortened or some passages should be replaced with shorter ones. Some interviewees also reported that seeing the countdown on the screen caused them extra pressure (S2, 3, 5, 6).

Second, some interviewees pointed out that the knowledge of reading sub-skills was not necessarily the only factor that influenced their reading test performance. For instance, some interviewees said that they might know most of the reading skills tested; however, insufficient vocabulary hindered their comprehension of the reading passage (S2, 5, 6).

Third, an interviewee raised an interesting issue. S3 mentioned that she regarded reading test feedback less useful compared to feedback or explanations of grammar or cloze test. She further explained that the reading test feedback only enabled her to confirm or reject her thinking/ prediction while reading. On the other hand, the feedback of grammar or cloze test usually contained a formula or new grammatical usage and consequently could make her learn more.

In summary of Stage II, the results of the interview data aimed to answer research questions (2) and (3). The interviews results showed in which ways test-takers perceive feedback to be useful or not. The overall results will be summarized and discussed in the following chapter.

CHAPTER FIVE

DISCUSSION & CONCLUSIONS

This study investigated Taiwanese university students' perceptions of the feedback of a diagnostic reading test. The researcher also examined characteristics relating to students' preferences towards reading test feedback. In addition, the research examined differences in the perceptions of the feedback of a diagnostic reading test between high- and low-English proficiency level students.

This chapter concludes the study by first summarizing its major findings following the order of the three research questions of the study. Then the results are discussed and the pedagogical implications for university teachers and diagnostic test developers are presented. Finally, the chapter ends with limitations of the study and suggestions for further research.

5.1 Summary of the Findings

This section presents and interprets the results of the study in the hope of answering all the research questions presented in Chapter 1.

5.1.1 Answer to Research Question 1: How useful do test-takers perceive the Reading Test feedback to be?

The results of the Survey of University Students' Perceptions of the OEAS Reading Test Feedback indicated that, in general, the participants of this study perceived the reading test feedback to very useful (M=3.78), which was in accordance with the result found in the pilot study (M=3.99). To be more precise, participants gave a positive usefulness ratings to each type of test feedback, including the overall test result report (3.60), the general explanations (3.84), the specific item explanations (3.70), and general aspects of the reading test feedback (4.12).

5.1.2 Answer to Research Question 2: In what ways do the test-takers perceive the feedback to be useful or not?

Based on the overall usefulness ratings of Stage I, the researcher listed out the top- and bottom-10 rated specific item feedback (see Table 4.6). The results showed that there was a significant difference between the lengths of the top- and bottom-10 rated specific item feedback. In general, the length of the top-10 rated feedback was shorter than that of the bottom-10 rated feedback. That is, test-takers preferred shorter feedback and tended to give it higher usefulness rating. Moreover, the interviewees in Stage II also expressed similar opinions that they preferred feedback that was clear and brief.

In addition, it was also found that item difficulty played an important role in participants' determining the usefulness of the reading feedback. The result of the study indicated a positive correlation between the participants' perceptions of the usefulness of the OEAS Reading Test feedback and the item difficulty of the reading test. Hence, feedback of test items with higher item difficulty tended to receive higher usefulness ratings. That is, students considered feedback of easier test items to be more useful.

The results also indicated that Stage I test-takers' considered Chinese and English feedback to be equally useful, both with a mean of 4.04. Stage II interviewees further pointed out that having bilingual feedback was useful to them. However, when presented with both Chinese and English feedback, most interviewees would read the Chinese feedback first.

Furthermore, study participants reported some other preferences for certain feedback (see 4.1.2.4, 4.1.2.5, 4.2.1, 4.2.3, 4.2.4 for details):

- Vocabulary list
- Immediate feedback
- Feedback with distractor explanations
- Feedback with shorter length
- Specific item explanation

To sum up, university students as a whole favored immediate feedback with distractor explanation and the length of feedback should be shorter. Besides, feedback of easier items was considered more useful. Additionally, the usefulness of bilingual feedback was identified by the participants.

5.1.3 Answer to Research Question 3: Is there a difference between test-takers of low and high English proficiency levels in how they perceive the feedback's usefulness?

The results of the study indicated that there were only few significant differences between perceptions of the low- and high English proficiency level students. The reason might be that both low- and high-level students gave positive ratings on most of the reading test feedback items. Even so, some noticeable findings were presented below.

First, the means of high-level students' perceptions of the usefulness of the reading test feedback were consistently higher than those of low-level students (see Tables 4.3, 4.4, 4.5, and 4.6). In other words, compared to the low-level students, the high-level students in general perceived the reading test feedback to be more useful.

Second, the high-level students perceived English feedback to be more useful than Chinese feedback (4.25 vs. 4.00) while low-level students considered Chinese feedback and Chinese translation to be more useful. In addition, high-level students (S1, 4) in Stage II interviews tended to have more positive attitudes towards the reading test feedback and expressed that they could learn new things by reading the English feedback.

Third, as seen in 4.1.2.1, there was a significant difference found between high- and low-level students' perceptions of Item B61 (general explanation for reading sub-skill 6: inference from reading).

5.2 Discussion and Implications

This section discusses the results of the study as well as presents some pedagogical implications of diagnostic reading tests. The discussion is divided into the following five sections: 1) overall positiveness about diagnostic reading test feedback, 2) factors affecting usefulness ratings: item difficulty and length of feedback, 3) high- and low-level test-takers' perceptions of diagnostic reading test feedback, 4) comparison between the current study and previous research, and 5) improving diagnostic reading test construction.

5.2.1 Overall Positiveness about Diagnostic Reading Test Feedback

In the current study, participants generally perceived the OEAS Reading Test feedback to be very useful. The positive results also support the findings of some previous studies (Jang, 2009; Yin et al., forthcoming). A number of possible interpretations might account for the results.

First, the feedback provided by the OEAS Reading Test was quite different from the test results that participants are used to getting in traditional tests. In addition to the test score, the OEAS Reading Test also presented each test-taker with an overall test result report classified into six reading sub-skills tested along with the number of questions answered correctly in each sub-skill. The diagnostic overall test report provided in this study is similar to what Spolsky (1990) viewed as "profiles," which show multi skills tested in addition to the overall test score. Many researchers (Shohamy, 1992; Alderson, Clapham & Wall, 1995; Jang, 2009) have proposed that reporting test results in "profiles" is better than merely presenting one overall score. The overall test result report enabled test-takers to better understand their strengths and weaknesses in certain reading skills tested.

Second, the OEAS Reading Test offered abundant and comprehensive feedback to the test-takers. In addition to giving test-takers an overall picture of their reading ability, the OEAS Reading Test provided general explanation of each corresponding reading sub-skill as well as specific explanation of each test item covered in that sub-skill. Besides, distractors in each item were also explained to clarify misunderstanding. Yin et al. (forthcoming) likewise found that test-takers showed preferences towards item feedback with distractor explanations.

Third, the feedback was provided immediately, right after the reading test. The content of the reading texts, unknown words, and uncertainties of test items were still all fresh in test-takers' minds. As a result, test-takers perceived the immediate feedback useful in confirming their interpretation of the reading texts as well as in clarifying some misunderstanding they had while taking the reading test. Moreover, the importance of immediate diagnostic feedback has been recognized by many researchers (e.g. Alderson, 2005; Jang, 2009).

These findings imply that English teachers may incorporate diagnostic reading tests to better understand students' reading comprehension ability so as to provide proper instruction. Moreover, teachers may utilize diagnostic reading tests to raise students' awareness of their strengths and weaknesses in reading skills which in turn provides students' a basis for future improvement. In addition to informing students' about their current reading proficiency level, diagnostic reading test feedback should contain a wild range of feedback such as a detailed test result profile, re-presentation of the original test materials, indications of whether a test-taker's response is correct or incorrect, and explicit explanations of the correct response as well as the distractors.

5.2.2 Factors Affecting Usefulness Ratings: Item Difficulty & Length of Feedback

This study found that both item difficulty and length of feedback were influential factors relating to students' perceptions of the OEAS Reading Test feedback. First, the results showed that easier items tended to receive higher usefulness ratings. One possible reason might be that easier items tend to be understood by more test-takers and thus test-takers could relate specific item explanation to their own interpretation of the test item and the reading passage. In other words, reading passages and test items that are too difficult and far beyond students' level may make certain feedback incomprehensible and less useful to students. Some interviewees in Stage II also mentioned that their knowledge of reading skills might not be the only factor influencing reading. They pointed out that too many unknown words in the reading passages made the whole reading process difficult.

From a sociocultural theory perspective (Yin, 2010), it could be said that easier items are within a test-takers' Zone of Proximal Development (ZPD), and the feedback is like a teacher "embedded" in the computer who is providing mediation in order to scaffold the test-taker's learning; thus, the test-taker will see the feedback as very useful–he/she could almost respond to the test items correctly, and only needs the assistance from the feedback in order to understand where he/she went wrong. More difficult items, on the other hand, would be beyond a test-taker's ZPD; the fact that the feedback in the computer is "frozen" and inflexible enough to adjust to the test-taker's lower level means the test-taker is unable to make good use of the feedback and thus get no assistance (Yin, personal communication).

With regard to the length of feedback, statistical results indicated that test-takers preferred shorter one. Some interviewees in the Stage II study also expressed similar opinions. Half of the interviewees mentioned that they tended to simply skim or even skip longer feedback since they had just completed four long passages in the reading comprehension test (see 4.2.4). Similarly, Van der Linden (1993) investigated learners' reactions to feedback in CALL programs and she found that lengthy feedback, which exceeded three lines, were not being read.

The aforementioned findings imply that teachers or test developers should pay attention to the difficulty level of reading passages and test items when constructing a diagnostic reading test. One option is to include reading passages of different length, topics, and difficulty levels so as to minimize the influence of other factors, such as test-takers' background knowledge and vocabulary size, in reading comprehension. Another option is to make the test difficulty level appropriate to the students' level. For instance, some universities in Taiwan use English placement test to assign students into classes of different levels (Sims, 2004; Tsai, 2008). Therefore, students in a class are rather homogeneous in their English proficiency and teachers may adjust the difficult level of diagnostic reading tests to meet student' level.

Though seemingly contradictory, participants in this study viewed detailed feedback useful while they showed preferences for shorter feedback. To deal with the conflicting opinions, teachers and test developers may first list out succinct explanations and then present a more comprehensive feedback. In this way, test-takers are free to look at part or all of the feedback that they find useful.

5.2.3 High- and Low-level Test-takers' Perceptions of the Reading Test Feedback

Though participants as a whole perceived the Reading Test feedback to be fairly useful, this study found some significant differences between the high- and low-level students' perceptions of the reading test feedback. First, the high-level students consistently perceived the reading test feedback to be more useful than the low-level students did. The results corresponded with those found in previous research on grammar test feedback (Yin et al., forthcoming). In addition, a possible explanation might be test-takers' willingness to undertake the process of error correction. Brandl (1995) examined high- and low-achievement students' preferences for error feedback in a German CALL program and found that high-achievement students were more willing to engage in the error correction process than low-achievement students.

Second, the low-level students considered Chinese feedback and Chinese translation to be more useful. One possible explanation might be that reading Chinese feedback was more secured and convenient since it is the participants' native language. In contrast, the high-level students deemed English feedback more useful than Chinese feedback. The interview results also supported the statistical results. Besides, high-level students (S1, 4) in Stage II interview tended to have more positive attitudes towards the reading test feedback and expressed that they could learn new things by reading the English feedback, which corresponded to the finding of Yin et al. (forthcoming). One possible reason might be that students with higher proficiency levels are likely to show stronger motivation in learning English (Oxford & Shearin, 1994; Peng, 2002; Chen, 2007), in this case, gaining more English knowledge while reading the English feedback.

Third, a significant difference was found between the high- and low-level test-takers' perceptions of the general explanation of reading sub-skill 6: inference from reading. Some results of previous reading research might account for this finding. Purpura (1998; 1999) investigated the effect that strategy use had on high- and low-ability test-takers' L2 test performance. The results indicated that inferencing was one of the strategies that high-ability test-takers utilized more frequently than the low-ability test-takers did. In addition, Hsu (2008) also indicated similar findings in the research on Taiwanese senior high school students' English knowledge, strategy use, and multiple-choice reading test performance. If high-level students use inferencing more often, then they would understandably get more from the feedback.

These findings imply that test-takers' proficiency levels did differentiate their perceptions of certain types of feedback. It has been suggested that the proper use of meaningful diagnostic feedback may not only lead to the improvement in test-takers' future language learning (Alderson, 2005; Cotos & Pender, 2008; Jang, 2009) but also the enhancement of instructional design (Kunnan & Jang, 2009). Therefore, when constructing a diagnostic reading test, teachers may first relate the content of a diagnostic reading test to the materials covered in the curriculum so that it is more relevant to the students. After the test, teacher may help students, low-level students in particular, understand the value of diagnostic feedback. Moreover, based on the results of diagnostic reading test, teachers may adjust their instruction and give proper suggestions for students to act upon their current competency level.

5.2.4 Comparison between the Current Study and Previous Research

As seen in Table 5.1, some findings of the current study are in accordance with the results of Yin et al.'s (forthcoming) research on test-takers' perceptions of diagnostic grammar test feedback. Meanwhile, there are some differences found between the results of the two studies.

Table 5.1

Similarities and differences in findings between the current study and Yin et al. (forthcoming)

Similarities
1. High-level test-takers perceived the feedback to be more useful than low-level test-takers did.
2. Test-takers in general favored feedback with distractor explanations.
Differences
1. In Yin et al., test-takers considered the Chinese feedback to be more useful that the English feedback. The current study found that test-takers generally perceived the Chinese and English feedback to be equally useful but the low-level test-takers deemed the Chinese feedback more useful.
2. Yin et al. found that test-takers in general preferred longer explanations, while the current study showed that test-takers favored shorter reading test feedback.

The two studies both indicated that high-level test-takers perceived the feedback to be more useful than low-level test-takers did. The possible reasons discussed in 5.2.3

might account for the findings. In addition, test-takers preferred detailed explanations which clarified common misunderstandings and explained distractors.

As for the differences in findings of the two studies, the first issue is concerning the Chinese and the English feedback. Yin et al. (forthcoming) found that test-takers perceived the Chinese feedback more useful while the current study showed that the Chinese and the English feedback of a reading feedback were equally useful. One possible reason might be that the diagnostic tests utilized in the two studies tested different language aspects, grammar and reading. The diagnostic grammar test feedback often included examples or formulas of grammatical usages and thus contained more terms. As a result, the Chinese feedback might be easier for test-takers to understand and was deemed more useful. On the other hand, the diagnostic reading test feedback usually covered some content of the reading passages; though the Chinese feedback helped clarify misunderstandings, the English feedback might make it easier for test-takers to refer back to parts the original reading passages. Therefore, test-takers' considered that the Chinese and the English feedback of a reading feedback were equally useful in a diagnostic reading test.

The second issue is regarding to the length of feedback. Yin et al. (forthcoming) found that test-takers had preferences for longer grammar test feedback while the present study found that test-takers considered shorter reading test feedback more useful. One possible reason might be the test format. In Yin et al., the grammar test consisted of 60 individual items whereas the Reading Test utilized in this study consisted of four reading passages and 36 test items (see 2.3.4.3 for details). The Reading Test was comparatively longer and test-takers might feel exhausted reading longer feedback after completing the reading test. Another possible reason might be the content of feedback. The diagnostic grammar test feedback often included formula or grammatical usages; therefore, students preferred longer or detailed feedback because it might enable them to learn more. In fact, one student in this study's Stage II also mentioned that she viewed grammar test feedback as more useful because it contained formula or grammatical usages and could make her learn more (see 4.2.4). On the other hand, the diagnostic reading test feedback usually referred to and covered part of the reading content and this resulted in lengthy feedback. As a consequence, test-takers might feel overwhelmed and thus showed preferences for shorter feedback.

5.2.5 Improving Diagnostic Reading Test Construction

5.2.5.1 Potential Ways to Improve Diagnostic Reading Test Validity

Yin et al. (forthcoming) suggest that usefulness of item feedback can be seen as evidence to support or disconfirm the validity of a diagnostic test. This study's investigation of feedback usefulness brought up data along these lines. On the support side, many students' responses showed the feedback targeted the problems they had (see 4.1, 4.2). On the other hand, some data raised questions about test validity. For instance, some students mentioned that the time was a pressure for them and when they saw the countdown on the screen they tended to randomly guess answers since there was not enough time (see 4.2.4). Thus, the results of the test might not fully represent the test-takers' reading comprehension ability. In addition, some students also reported that they might know most of the reading skills but too much vocabulary hindered their comprehending the reading passages. As a result, it seemed to be a vocabulary problem and not a reading problem per se. Therefore, teachers and test designers should pay attention to students' thoughts about feedback to improve the validity of the test itself.

5.2.5.2 Students' View for Improving Diagnostic Reading Test Design

To evaluate the usefulness of any given diagnostic language test, it is important to investigate test-takers' perceptions of the diagnostic feedback (Bachman & Palmer, 1996, Alderson, 2005; Jang, 2009; Kunnan & Jang, 2009, Yin et al., forthcoming). The participants in this study also provided some constructive suggestions for improving an online diagnostic reading test. First, in order not to make the reading process tiring, the font size of the reading texts and the feedback should be enlarged. Second, the online assessment system should enable test-takers to circle or underline key words and sentences when reading the passages. Many participants expressed that this function would reduce the difficulty of taking an online reading test and save them a lot of time reading the passage again to locate some information related to the test items. Third, students expressed the desire to have more control over which part of the feedback they wanted to read.

Therefore, test designers may take this into consideration and rearrange the way that feedback is presented. Take the overall test result report for instance; in addition to its original design (see 2.3.4.3 for details), item numbers that answered correctly and incorrectly by a test-taker could be listed out. Clicking on the item number will lead to specific explanation of that item. Thus, test-takers can choose which feedback they want to read instead of browsing back and forth to locate the information they need. In addition, test designers may also make feedback able to be cut-and-paste so that test-takers can choose the feedback that they deem useful and save it as a file or even send it to their e-mail. In other words, the feedback can be individualized and tailored to students' needs.

In this way, online diagnostic reading test designers may increase the flexibility of the system from the learners' point of view so that it better serves as a self-assessment tool and enables students to control their own learning path.

5.3 Limitations of the Study

This study aimed at investigating university students' perceptions of diagnostic reading test feedback. Although the present study has produced substantive findings and the research questions have been answered, the study was carried out and completed subject to the following limitations.

First of all, the findings here are generated from university students in central Taiwan. They were all freshmen from one university. Due to the limited number of participants, the results might not fully reflect the differences between high- and low-level students' perceptions of reading test feedback. As a consequence, limited representativeness of the sample may hinder the generalization of the findings of this study.

Second, the data used and analyzed in Stage I were collected entirely through self-report questionnaires. In addition, the time that participants read each of the feedback was not recorded. Hence, the participant might not have offered completely honest responses or might have provided answers hastily while filling out the questionnaires.

5.4 Suggestions for Further Research

On the basis of the findings and the aforementioned limitations of the current study, the researcher provides the following suggestions for further research.

First, researchers may include a larger and more representative sample to have a more comprehensive understanding of students' perceptions of a diagnostic reading test feedback. To do so, researchers need to recruit their participants of different school years and from universities in different parts of Taiwan.

Second, researchers may try recording the time that students spend reading each feedback. By doing so, researchers will be able to examine whether students read the feedback or not as well as whether time differences affect students' perceptions of the diagnostic reading test feedback.