LEXICAL RICHNESS AND ACCOMMODATION IN ORAL ENGLISH EXAMINATIONS WITH CHINESE EXAMINERS

JIAN ZHANG

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol for the degree of Doctor of Philosophy

Faculty of Arts, Creative Industries and Education,
University of the West of England, Bristol
November 2014

Acknowledgements

I wish to thank so many people who have helped and supported me during my life and in my PhD study.

My PhD lasted for longer than it usually takes because I applied for a suspension twice due to difficulties on the family front in China. I'm very pleased that I finally managed to complete my research during this most difficult time. And in fact after the busy hours of work and child-caring, I really enjoyed the time alone working on my PhD.

Firstly, I am very grateful to my Director of studies, Dr. Jo Angouri, who kindly took me over when my former supervisors Dr. Michael Daller and Prof. Jeanine Treffers-Daller left UWE for other universities. She discussed with me every detail of the dissertation. She provided me with very valuable suggestions on the thesis, especially the research design of the qualitative research. She read my drafts carefully and provided insightful comments on each draft. My thesis could not have reached the present level without her help.

I'm also very grateful to my former Director of study, Dr. Michael Daller, who is still in the supervision team and Prof. Jeanine Treffers-Daller, who has left the supervision team. Michael, being an expert on statistics and lexical richness, has great enthusiasm for research on such subjects and helped me a lot in statistics and research design. He is always happy and excited to discuss with me the lexical measures and their performance and is always so friendly and encouraging. Prof. Jeanine Treffers-Daller has helped me greatly in the transcription of my data and the application of the software of CLAN. She also read the first draft on literature, the pilot study and methodology chapters and provided insightful comments and suggestions. Although Dr. Michael Daller and Prof. Jeanine Treffers-Daller have moved from UWE to other universities, I'll always be indebted to them and respect them as my supervisors.

I would also like to express my love and heart-felt thanks to my parents. They are always ready to help me whenever I need them and have been unfailingly kind and generous. I could not have accomplished anything without their support and love.

I'm especially grateful to my son, Wei An, who has brought me love, joy, happy surprise and strength to move forward! It is him who encourages me to become a "superwoman" and "supermum" who will face any difficulty.

My thanks and love also goes to my husband, who has always supported me in my life and study. I could not have accomplished my PhD without him.

I'm especially grateful to my friends Ed and Liz who kindly offered me accommodation in Bristol in the first two years of my PhD before they moved to Wales.

I'm very grateful to the Higher Education Development Center of BEEA for granting me the opportunity to do the present research and helping me collect data. I'm greatly thankful to Professor Wu Guhua, Hu Jie, Zhan Chi and all the GESE examiners in Beijing who are involved in the present research and all the British colleagues from Trinity College London who have supported me for so many years.

Without the support from BIT and UWE, I would not have been able to embark upon, let alone, complete my study. I’m really indebted to all those people who made great efforts in the collaboration of UWE and BIT and made my dream of doing a PhD come true. I do greatly appreciate the support and encouragement from all my colleagues in BIT and all the stuff members and fellow PhD students in UWE. I’m also very thankful to Chen Hui and Wang Xiaoqing, who helped me with the transcription of my data.

Finally I wish to express my appreciation and gratitude to all the people who have helped me and supported me, but whose name I failed to mention here. Thank you very much indeed.

Abstract

Lexical assessment and lexical accommodation in oral examinations are new research dimensions, which have both theoretical and empirical values, however they are still much neglected. The present research aims to investigate: first, whether or not and how (if so) the measures of lexical richness can differentiate between candidates of three different grades of **Graded Examinations in Spoken English of Other Languages (**GESE) and whether or not those measures can differentiate good performers from poor performers at the same grade of GESE. Second, whether or not and to what extent (if so) Chinese examiners accommodate to the candidates at the lexical level.

180 samples from Grade 2, 5 and 7 GESE were collected. All the data were transcribed into Codes for Human Analysis of Transcripts (*CHAT*) format for the Child Language Data Exchange System (*CHILDES*) (MacWhinney 2000) for analysis. First, the lexical measures of *Token, Type, Guiraud, Guiraud Advanced (AG)* and *D* of both candidates and examiners were obtained and analyses were conducted to investigate the relationship among them. Secondly, qualitative data were collected from interviews with GESE examiners to interpret the quantitative results.

The quantitative results indicate that: 1) all the lexical measures can differentiate candidates of Grade 2 from Grade 5 and can differentiate candidates of Grade 2 from Grade 7 as well. However, there is no significant difference between Grade 5 and Grade 7 candidates' lexical variables. 2) In Grade 2 and Grade 5, all the candidates' lexical variables can distinguish between the qualified and poor performers of the same grade. Only *Type, D and AG* can differentiate between the qualified and poor candidates in Grade 7. 3) All the GESE score variables are correlated with each other, which shows a halo effect; the only GESE score variables that correlate with all candidate lexical variables in the pooled data is *Focus*. 4) The examiner variables cannot differentiate between qualified performers and poor performers in the same grade. 5) The only lexical variable that reflects the examiner's lexical accommodation to the candidate is *AG*.

The qualitative analyses indicate that the GESE examiners employ special characteristics in vocabulary assessment and the data also explain some of the quantitative results. It was found that the Chinese local examiners of GESE might apply *meaningful and relevant input* and *the general communicative ability* of the candidate as reliable overall rating strategies, and factors that affected the performance of the Grade 7 candidates are also discussed. The findings may not only shed light on a better understanding of the constructs of vocabulary knowledge and lexical richness, the accommodation the Chinese examiners conducted on candidates, but also provide insight into the design and improvement of examination procedures and training of Chinese oral examiners.

Table of contents

List of Tables:

List of Figures

List of Appendix

Chapter 1 Introduction

1.1 The research context

English, as an international language, has gained outstanding popularity in China with globalization of the world. Consequently, there has been a boom in English learning and teaching in China in the past two decades or so. Oral English examinations, especially the face to face oral English examinations, have also developed and grown vigorously in the past 20 years. There are oral English examinations introduced to China from English-speaking countries: for example, the speaking tests of IELTS and TOEFL, the speaking tests of Cambridge ESOL Main Suite examinations and the Graded Examinations in Spoken English for Speakers of Other Languages (GESE) of Trinity London. There are also national oral English examinations set up by local Education Bureaus in China: for example, speaking tests of College English Test Band 4 and Band 6 (CET- 4, CET- 6), speaking tests of Test for English Majors Band 4 and Band 8 (TEM-4 and TEM-8) and Public English Test System (PETS). In addition, there are examinations within the curriculums of English courses at all levels, from kindergarten to postgraduate English courses. In China, English is the compulsory course for most students from primary school to PhD levels. Except for some international English examinations, most English oral examinations in China are conducted by Chinese examiners.

However, compared with the countless oral examinations that have emerged and developed in China in the last 20 years, there is not much research on oral English examinations in the country, and research on the vocabulary in oral examination settings is very rare.

The present research was firstly motivated by the practical difficulty or the uncertainty of vocabulary assessment in GESE examinations. I had long realized the difficulty of assessing vocabulary as a GESE examiner before the present research started. An analytical rating system had been adopted in GESE since it was introduced to China in 1999 until 2010. The candidate's performance in the examination is measured by means of different assessment criteria such as readiness of the candidate, pronunciation, usage of vocabulary and grammar and etc.. The evaluation of vocabulary is an important aspect of *Usage*, one of the assessment criteria of GESE in all grades. However, in the general descriptions of the assessment criteria, only very general terms were used for vocabulary, such as the *range of vocabulary* and the *appropriate vocabulary* (GESE Syllabus from 2002). But how to rate the *range of vocabulary* or the *appropriate vocabulary* is not specifically described in the syllabus. In addition, Chinese local examiners of GESE have different ideas on how to assess vocabulary (see the results of the pilot study of the present study in Chapter 3). Literature on lexical assessment in oral English examinations, however, is rare and can provide little help in this respect.

GESE of Trinity College, London was introduced into China nearly 15 years ago and the number of candidates has increased from about 2,000 in 1999 to more than 30,000 by the end of 2012 according to the statistics of Beijing Education Examinations Authority (BEEA). However, very few studies have been conducted on GESE, especially regarding the assessment measures of vocabulary in GESE.

As an experienced teacher and GESE examiner, I have long realized that it is necessary to investigate assessment measures of vocabulary both in theory and in practice. Theoretically, the construct of vocabulary knowledge and lexical richness is not unified in literature, which requires further research. Practically, researchers have proposed different measures to quantify vocabulary knowledge in ways “other aspects of language cannot” (Milton, 2007, p.334). However, in real examinations, it is not practical for examiners to compute different measures of lexical richness when a score is required almost instantaneously after the examination, and the possible effects of these lexical measures on assessment are not clear at all. Research in this area will definitely promote our understanding of vocabulary assessment and as a result, shed lights on the assessment of oral English examinations and examiner training.

1.2 Rationale for the present research

Vocabulary knowledge is considered a fundamental component of L2 proficiency and is “the core component of all the language skills” (Long and Richards 2007, p.xii). Milton (2008) pointed out that vocabulary knowledge can be measured in ways other aspects of language cannot. “Measuring the vocabulary knowledge of learners can help give a much better impression of the scale of learning which is taking place than is possible with other measures of language proficiency” (p.334).

Lexical richness is the general term for the measure of different aspects of vocabulary use. According to Read (2000, p.200), there are 4 aspects of lexical richness: *lexical variation, lexical sophistication, lexical density* and *number of errors*. Most research on lexical richness is focused on the first two aspects of lexical variation and sophistication. Traditionally, lexical richness measures were mainly used for written discourse. Recently attention has been turned to spoken discourse, but the research is still far from enough in Read’s opinion (2000). Many researchers proposed new measures of lexical richness and proved with their own data that the measures are more valid than the traditional measures in spoken discourse. For example, there is the measure *D* proposed by Malvern and Richards and colleagues (Richards & Malvern, 2000; Malvern & Richards, 2002; Malvern et al., 2004), which is based on a single parameter of a mathematical equation that models the curve of the falling Type/Token Ration (TTR) with increasing text length N.

Guiraud Advanced (AG) was proposed by Daller, van Hout and Treffers-Daller (2003), which is the ratio of *advanced Types* shared by the square root of the total number of *tokens*. The definition of *advanced Types* is normally based on frequency lists.

Since *D* was proposed, it has been widely used as an effective measure of vocabulary use in the field of child language development and SLA (for example Jarvis, 2002; Duran et al., 2004; David, 2008, Yu, 2009 and Lu, 2011). However, some researchers also suggest that *D* was not as reliable as the creators claimed. McCarthy and Jarvis (2007, p.482) argued that *D* is also affected by text length and its reliability “is limited to specific and quite short text lengths”, which might be between 100 and 400 tokens according to their research results, and “which is in line with the claims of Malvern *et al*. (2004)" . McCarthy and Javis (2010) examined the validity of the measure of lexical diversity (*MLTD*) by comparing it with other competing indices of lexical diversity. They suggested that the three indices of lexical diversity of *MLTD*, *voc-D* (or *HD-D*) and Maas used in the studies seemed to have "captured unique lexical information", and researchers should bear in mind that "lexical diversity can be assessed in many ways and each approach may be informative as to the construct under investigation" (p. 381). Researchers should apply several valid measures instead of using any single index in their studies in order to foster a better understanding of the construct of lexical richness.

In the present research, different lexical richness indices of *Type, token, D, Guiraud, Guiraud Advanced (AG*) as well as *Mean Length of Utterance* (*MLU*), a general indicator of language proficiency (Brown, 1973) were applied to examine the lexical features of both GESE examiners and candidates. Hopefully some unique lexical information about both GESE examiners and candidates of different levels can be captured by different indices of lexical richness. This is the pioneer research on lexical richness in GESE conducted in China.

In addition to lexical richness, lexical accommodation is another key concept in the present research. Accommodation (Giles & Powesland, 1975) is a characteristic of natural communication when a person changes his or her speech to adapt to the interlocutor or to show difference from the interlocutor. The purposes of accommodation might be to get approval or to keep some distance from the interlocutors, and another important purpose of accommodation is to promote understanding. Accommodation occurs in oral examinations (Ross, 1992; Lazaraton, 1996; Malvern & Richards, 2002; Lorenzo-Dus and Meara, 2005) when examiners try to facilitate the examinee when the latter has troubles in the process of communication. Ross (1992) believed that accommodation in oral interview is very important for both the reliability and validity of the exam and even proposed that in addition to the abstract definitions of proficiency, the manner and quantity of interview accommodation necessary for the interview should be included in the assessment process. Lazaraton (1996, p.167) also proposed that "it is critical that more studies on oral test interaction, whether they be statistical or discourse analytic or, ideally, both, be undertaken on other widely used proficiency examinations", so we will have a better understanding of the "validity of oral proficiency assessment itself". GESE conducted in China has been developed into a very popular examination, yet research has rarely been undertaken up to now. The present research is expected to bridge the gap between GESE practice and research.

Compared with accommodation in sound and discourse, lexical accommodation is not much studied, and studied even less in respect of lexical accommodation of non-native examiners. Research on accommodation and vocabulary are two distinct inquiries of applied linguistics. However, by examining the measures of lexical richness of both examiners and candidates and the correlation between the two groups of measures, this present study links the research into lexical richness and accommodation.

Richards and Malvern (2000) and Malvern and Richards (2002) did some pioneer work on linking the study of lexical richness and lexical accommodation. However, their conclusions were made on a small data set, with only 34 UK students taking French as a second language. Lorenzo-Dus and Meara (2005) also tried to investigate the relationship between examiner support and examinee vocabulary based on the analyses of 30 Spanish oral examinations. What is in common in the above mentioned research is that they used rather a small data set (less than 35) collected on the basis of availability. Quantitative analysis of large-scale random sampling is rare in the field of lexical accommodation of L2 speakers in oral interview settings. The current research is carried out on data which were randomly chosen from a much larger corpus of GESE examinations. 180 data sets were collected from the corpus. It is believed that random sampling on a large scale can present a more complete and representative picture of the population.

The research results of Malvern and Richards (2002), Richards and Malvern (2000) and Lorenzo-Dus and Meara (2005) all indicate that the relationship between candidates' use of vocabulary and examiner accommodation is not simple and straightforward. More studies are necessary to explore the relationship before we take accommodation into assessment criteria as Ross (1992) proposed.

For most international oral examinations such as IELTS and TOEFL, the examiners are native examiners, but GESE is special that it has examiners who are both native English speakers and non-native English speakers in China and India. Literature on how non-native English examiners of an international examination adapt their language level to the examinee at the lexical level is very rare. It is hoped that the present research will start some initial work in this area and it may provide insights into the administration and examiner training of the oral examinations conducted by non-native examiners all around the world.

1.3 The general research purpose

Based on the data drawn from 3 different levels of an English oral examination of GESE conducted in Beijing by Chinese local examiners, this research mainly investigates the lexical richness of both candidates and examiners of different levels and the lexical accommodation the examiners may perform towards GESE candidates. The first focus of the present research is to investigate whether or not and how (if yes) the measures of lexical richness can differentiate between candidates of 3 different grades and whether or not those measures can differentiate good performers from poor ones within the same grade. Another focus of the research is to investigate whether or not and to what extent (if yes) Chinese local examiners accommodate to the examinees at the lexical level.

1.4 The overall structure of the thesis

Following the first chapter of **Introduction**, which sets out the research context, the rationale of the present research and the general purposes of the thesis, the rest of the thesis is structured as follows. **Chapter 2 is the Literature Review** which looks at the key literature on input, accommodation and lexical richness and research in these areas is presented and critically reviewed. The gap in the literature is discussed and the general research questions of the present research are proposed based on the literature review. **Chapter 3** is the **Pilot Study**. First, a preliminary pilot study was conducted to investigate the Chinese local GESE examiners' viewpoints on the assessment of vocabulary and the relationship between vocabulary and other aspects of language proficiency, and then quantitative analysis was conducted on a small-scale examination data set for further investigation. The pilot study helps formulate the research questions of the thesis; helps choose measures of lexical richness to be applied in the main research, and helps choose instruments to be applied in the main research. The research methods of the main study are presented in **Chapter 4**, **Research Methodology**, where elaborated research questions are developed on the basis of general research questions proposed at the end of the literature review, and the subjects, instrument and research procedures are also described and discussed. The main results of the **Quantitative Analyses** are presented in **Chapter 5**. **Chapter 6** presents the qualitative analyses of the interviews with three experienced examiners, and the results of Chapter 6 also provide insights into the question concerning the unexpectedly low indexes of lexical variables of Grade 7 compared with Grade 5. The problems with Grade 7 candidates in the interactive tasks are specially discussed with possible factors that may have caused the problems. It also partially interprets some of the quantitative results presented in Chapter 5. **Chapter** 7 is the **Conclusion** and it summarises and synthesizes the findings of this research. The implications and the contribution to knowledge made by the research is also discussed. Finally, the chapter concludes by discussing some of the limitations of the present research and suggestions for future research are also provided.

Chapter 2 Literature Review

2.1 Introduction

In this chapter, the key literature on input, interaction and accommodation in second language (L2) acquisition is reviewed and the main findings in the field of lexical richness are presented as well.

This chapter starts with language input in second language acquisition (SLA). Although different importance is placed on second language input by different schools of thought, different linguistic views such as the Behaviourist view of SLA, the mentalist views of language acquisition and the interactionist view of SLA all recognize that input is a necessary factor in learning a second language. Following the review on early research of mother tongue and L2 input, the development of research on L2 interactions is also discussed.

Following the discussion on input and interaction, the development of Accommodation Theory is reviewed. Accommodation Theory was first proposed as a socio-psychological theory (Giles and Powesland 1975). According to this theory, the speaker adjusts the way of speech to his or her interlocutor in order to win social approval or to promote understanding. In language proficiency oral interviews, accommodation at different levels may occur between the interlocutors, or the examiner and the candidate / examinee.

Lexical accommodation is one of the important aspects of accommodation which has not received much attention in studies of oral examinations. In Malvern and Richards (2002), the only teacher variable that was most responsive to student variable is the lexical diversity index *D* developed by the researchers. (*D* is based on a single parameter of a mathematical equation that models the curve of the falling Type/Token Ration (*TTR*) with increasing text length N.) The examiners who are also the students' teachers did accommodate to the students at the lexical level, but they only accommodated to the general level of the whole class instead of to individuals. One hypothesis of the present thesis is built on the research results of Malvern and Richards and their colleagues (Richards and Malvern, 2000; Malvern and Richards, 2002; Duran et al., 2004).

Lexical richness has been considered a quite “illuminative predictor” of a learner’s language proficiency and an important indicator of the quality of a learner’s speaking and writing performance (Yu, 2009, p. 236). Lexical diversity (or lexical variation) and lexical sophistication (or lexical difficulty) are the two aspects of lexical richness which have attracted much attention from researchers. It is generally accepted that these two aspects can indicate how well an L2 learner actively uses the vocabulary. An L2 learner’s vocabulary knowledge and lexical richness is also discussed in this chapter. In the final part of this chapter, the gap in literature is discussed and the research questions of the present research are proposed.

2.2 Input and interaction in SLA

The studies on L2 input and interaction have developed in the past 40 years from the description of the nature of modified input in the earlier stage to the exploration of the link between input and L2 acquisition.

As early as nearly three decades ago, Ellis defined L2 input and interaction as follows:

> Input is used to refer to the language that is addressed to the L2 learner either by a native speaker or by another L2 learner. Interaction consists of the discourse jointly constructed by the learner and his interlocutor; input, therefore, is the result of interaction (Ellis, 1985, p.127).

In the 1985 definition of input and interaction proposed by Ellis, only input and interaction are involved in the process. However, in the up-dated version of interactionist account, more factors such as selective attention, output and feedback are added to interaction.

In the context of conversations and oral interviews, the speech from one person, or the output of the speaker, is also the input for his or her interlocutor. The conversation or interaction is the co-construction of interlocutors.

2.21 Early research on accommodative features of L2 input and interaction

The earliest research on L2 input in the 1970s and1980s was mainly concerned with two questions; first, what are the features of L2 input the L2 learners typically receive and second, what are the functions of L2 input in L2 acquisition.

The research on L2 input was greatly influenced by the research on the L1 input children receive. Based on a large quantity of research, both Ellis (1994, p.251) and Larsen-Freeman and Long (2000, p.115) conclude that the L1 input addressed to language-learning children are fine-tuned. In other words, the L1 input, or the caretaker talk, which refers to the language addressed to the children by the parents and other caretakers, is well-formed and well adapted to the children's language ability, especially their understanding abilities or comprehension. The typical language that addressed to children has some special features compared with language addressed to adults according to Long (1991): syntactically, shorter and less varied utterance and higher ratio of content words are used. Phonologically, higher pitch, clearer articulation, exaggerated intonation, slower speed of delivery and other features are found in caretaker talk. In the area of semantics, more restricted vocabulary or less varied vocabulary is featured in caretaker talk; the topics are restricted to here and now, so a higher frequency of nouns and present tense verbs are used. Ellis (1994) summarizes that caretaker talk is 1) more grammatical 2) simpler and 3) more redundant than speech addressed to adults.

Many researchers (Long, 1983; Yano, Long and Ross, 1994; Gass and Varonis,1985; Krashen,1985; Parker and Chaudron,1987; Ellis,1994) have concentrated on the features and functions of L2 input. Some researchers addressed the ungrammatical modification of the input in the 1970s and 1980s. Ferguson (1975, cited in Ellis, 1994) used elicited written data to investigate how native speakers switched to ungrammatical forms when talking to non-native speakers. He claimed that the ungrammatical talk of native speakers is a variety of speech which he named foreigner talk (FT). But soon more and more researchers began to report that ungrammatical modification is not the norm in SLA. Long (1983) argued that the use of ungrammatical foreigner talk is very limited and only appears if two or more of the following conditions are met:

(1) the non-native speaker has very low or no proficiency in the language of

> communication; (2) the native speaker is, or thinks s/he is, of higher status than the non-native speaker; (3) the native speaker has considerable prior foreigner talk experience, but of a very limited kind; and (4) the conversation occurs spontaneously, i.e. not as part of a laboratory study (p.126).

Larsen-Freeman and Long (2000, p.119) argued that L2 input is well formed and the findings were similar to those in caretaker talk. "Modified but grammatical speech to foreigners tends to be a more regular version of the language, avoiding forms which constitute exceptions to general rules in the language concerned". Hatch, Shapira and Wagner-Gough (1978, cited in Ellis, 1994) also found that grammatical foreigner talk is the norm in most classrooms. Teachers, especially language teachers, use that kind of language to organize and manage classroom activities. So the grammatical foreigner talk is also called teacher talk or language teacher talk. According to Ellis (1994, p.254), grammatical foreigner talk or teacher talk is characterized by three modification processes: simplification, regularization and elaboration. Simplification is achieved by avoiding the use of difficult items in the target language. Table 2.1 shows all the grammatical linguistic modifications which contribute to simplification.

Table 2.1 Simplification in grammatical foreigner talk

Type of simplification	Comment
Temporal variables	Speech to non-native speakers (NNs) is often slower than that addressed to native speakers (NSs) - mainly as a result of longer pauses.
Length	FT makes use of shorter sentences (fewer words per T-units)
Syntactic complexity	FT is generally less syntactically and propositionally complex, i.e. fewer subordinate clauses of all kinds (adjectival, noun, and adverbial), greater use of parataxis (e.g. simple coordinate construction), and less preverb modification.
Vocabulary	FT manifests a low type-token ratio and a preference for high frequency lexical items.

(adopted from Ellis 1994, p.256)

It can be found that in simplification, the linguistic modification has much in common with caretaker talk syntactically, phonologically and semantically, and the main purpose is to facilitate understanding. It needs to be mentioned that type-token ratio (TTR) was used to describe vocabulary use of foreign talk. TTR was widely used in both child language and SLA research as the measure of vocabulary use in the 20th century, but it is not in favour any more. The reasons are discussed in detail in the later part of this chapter.

Regularization means using some forms that are very explicit. For example, the full forms rather than contracted forms are preferred; explicit markers of grammatical relations; lexical items with more general meaning rather than specific meaning; and the avoidance of idiomatic expressions. Regularization can help to make the meaning of utterances more transparent.

Elaboration means making the sentences longer in order to make the meaning clear. For example paraphrases, synonyms are often used to make the meaning easy to understand.

In addition to linguistic features of grammatical foreigner talk, researchers also investigated the interactional features of it. Many researchers found that there were special features of foreigner talk at the discourse level. It was the modifications of the discourse that were used more often in foreigner talk and was also more consistently observed, so many researchers turn to investigate the foreigner discourse or what Ellis (1994, p.257) and Long (1981) called interactional modification. Later on it was also referred to as negotiation by Long and other researchers. Long (1981) proposed that interaction modifications include clarification requests, confirmation checks and comprehension checks, which he later categorized as strategies of utterance repair. Pica et al. (1991) argued that terms such as clarification request, confirmation checks and comprehension checks implied that the research could identify the intention of the speakers, which is seldom the case, so they used the terms signals for listener utterance and trigger from speaker utterance. Ellis (1994, p.258) summarized the interactional modification in foreigner talk and categorized interactional modification into discourse management and discourse repair. The purpose of the former is to simplify the discourse to facilitate communication and the latter takes place when there is a need to repair communication breakdown or learner errors. Negotiation of meaning is classified as repair of communication breakdown, which is one of the two types of discourse repair. The main features of interactional modification are shown in Table 2.2.

No matter how different terms are used or categorized, the features of interactional modification discussed by the researchers are the same. As Pica (1994, p.497) remarked: "Whatever labels are used, these features of negotiation portray a process in which a listener requests message clarification and confirmation and a speaker follows up these requests, through repeating, elaborating, or simplifying the original message." Many of the features mentioned in Table 2.2 were also used by researchers such as Ross (1992) and Lazaraton (1996) to investigate accommodation in oral proficiency interviews.

Table 2.2 Interactional modifications in foreigner talk

Interactional modification	
Discourse management	Discourse repair
Types of discourse management	Repair of communication breakdown
Amount and type of information conveyed	Negotiation of meaning (requests for clarification; request for confirmation; self-and-other repetitions)
Use of questions	
Here-and-now orientation	
Comprehension checks	Relinquishing topic
Self-repetition	

(adopted from Ellis, 1994, p.258)

The shift of attention from L2 input to L2 discourse in the 1980s and 1990s also promoted the importance of interaction in L2 learning. According to Long (1991), the negotiation of meaning takes place between native speakers (NS) and non-native speakers (NNS), and the foreigner talk is a "dynamic, constantly being adjusted to what the learner is perceived to be understanding" (p.126), so the analysis of interactional features needs to look at the speech and previous speech of both participants in a conversation.

Gass and Varonis also contributed a lot to the shift from L2 input to L2 discourse in L2 research. They carried out a study to investigate the nature of discourse involving non-native speakers (NNS), to be more specific, variables influencing native speaker (NS) foreigner talk and the form that speech modification takes. 80 taped telephone interviews between NNS at 2 distinct proficiency levels and NSs and 20 NS-NS interviews were investigated. Five variables were considered: 1) negotiation of meaning, 2) quality of speech, 3) amount of repair, 4) elaborate responses and 5) transparent responses. Based on analyses of the data, the researchers finally concluded that the speech of NS changes as a function of an NNS's ability to understand and be understood. NNS' understanding of NS' speech is an important factor that triggers NS speech modification. The authors also found that transparency is common in both NNS and NS speech. By transparency, it means "giving information in a less compact, and thus potentially more easily interpretable manner" (Gass and Varonis 1985, p.50). Examples of transparency include carefully articulated speech, full clauses or decreased number of non-finite verbs, etc. The authors suggested that transparency might be a general cognitive principle underlying aspects of both foreigner talk and L2 acquisition.

2.22 Different views on the role of input and interaction in SLA

Ellis (1985) classified three different views on the role of input in SLA: the behaviourist view, the mentalist view and the interactionist view. In this section of the chapter, the review of literature follows this line of discussion, but the content goes beyond what Ellis discussed.

In the behaviourist model, L2 input serves as both stimuli and feedback in the language learning process. In the case of stimuli, the learner imitates what his or her interlocutor says and internalizes the forms and patterns. In this sense, input is a determining factor in L2 learning. In the case of feedback, it reinforces the correct forms and patterns of the utterance and corrects those that are incorrect. As Ellis (1985, p.128) puts it, in the behaviourist model "the regulation of the stimuli and the provision of the feedback shape the learning that takes place and leads to the formation of the habit”.

The mentalist view of SLA emphasizes the internal factors of the learner, such as the black-box Language Acquisition Device (LAD) from Chomsky in the 1960s and Universal Grammar (UG) since the 1980s.

Chomsky emphasized the determining function of the innate mechanisms in language learning. He argued that humans are innately endowed with universal language-specific knowledge, or what he calls UG. According to UG theory,

> What we know innately are the principles of the various subsystems of S0 [the initial state of the child’s mind] and the manner of their interaction, and the parameters associated with these principles. What we learn are the values of the parameters and the elements of the periphery (along with the lexicon to which similar considerations apply). (Chomsky, 1986, p.150).

According to UG theory, all human beings are born with language knowledge which consists of a universal set of principles and parameters. The principles are universal and not varying but the parameters possess variations, and the parametric variations characterize the differences between languages. This knowledge of language does not need to be learnt but it needs to be triggered. The role of input is only to trigger the UG, and the nature of input does not affect acquisition at all. The input the child receives when learning his or her mother tongue is poor or degenerate in nature, however the child can produce or create limitless sentences he or she has never heard before, which is often referred to as the logical problem of language acquisition.

UG is a linguistic theory of natural languages and it would be very difficult to deny that L2 is not a natural language (Mitchell and Myles, 2004). In this case, there are two different possibilities of the role of UG in SLA. First, second language learners are UG- constrained and have full access to UG as first language learners and the second is that the second language learners only have partial access to UG, because some parts of UG are no longer available to them. Input functions in SLA in both cases, but its role is not as crucial as in UG.

Krashen puts much more emphasis on the role of input. Krashen believes that L2 acquisition is driven by the language environment rather than by the mind. He emphasizes "the nature of the input rather than the processes of the mind" (Cook, 1993, pp.54-58). According to Krashen's Input Hypothesis, "humans acquire language in only one way – by understanding messages or by receiving comprehensible input" (Krashen, 1985, p.2). Here comprehensible input is the crucial factor in acquiring a first and second language in his model. It must be neither too difficult nor too easy to understand, which can be shown in a formula *i* +1. Here *i* is the current level and *i* + 1 is the next level the learner will go to. If the input is slightly beyond the current level of the learner, he or she will progress continuously along the stages from *i* to *i* +1. Krashen (1983, pp.138-139) once proposed that L2 acquisition involves three stages to turn input into intake, which is input that has become part of the interlanguage system of the learner: understanding the L2 *i*+1 form, noticing the gap between L2 *i* + 1 and the interlanguage rule the learners controls and finally the reappearance of the *i*+1 form. But in other versions of the hypothesis, the concept of noticing is not addressed. It seems that "the acquisition takes place when the learner understands language contains *i* +1. This will automatically occur when communication is successful" (Ellis, 1985, p.157).

Krashen's Input Hypothesis has been criticized for several reasons. First, the hypothesis is not easy to testify; second, there is no precise definition of "comprehensible input"; thirdly, the terms such as the current level of the learner, the *i*+1 level of the comprehensible input are not described in a characterized way and they are not easy to quantify. In addition, many important factors that may affect the language study such as social environment, the internal language acquisition device of the learner are not discussed (Mitchell and Myles, 2004; Gass and Selinker, 2008).

While the behaviourist view of SLA regards language progress as caused by external factors, mentalist views of language acquisition emphasize the inner ability of the learner, and the interactionist view of SLA account for learning through input, output and feedback that comes as a result of interaction (Mitchell and Myles, 2004). With the development of this approach, it gradually puts stress on both the inner ability and the language environment. The nature of a learner's mental organism (e.g. noticing, attention) both determines and is determined by the nature of input. According to this viewpoint, not only utterance, but the discourse between the learner and the interlocutor is also important.

2.23The development of the interactionist views

The development of the Interactionist view on input and interaction is reflected on the change of Interaction Hypothesis (Long 1981. 1996).

Long (1981) first proposed the Interaction Hypothesis which developed the Input Hypothesis of Krashen. The basic claim of the Interaction Hypothesis is that L2 acquisition is promoted if learners solve communication problems by means of conversational modification. Long (1981) conducted a study of 16 native speaker pairs and 16 native speaker vs. non-native speaker pairs performing the same face to face oral tasks. He found that the major difference did not lie in grammatical complexity, but that the native vs. non-native pairs were more likely to use some communicational tactics such as repetitions, confirmation checks, comprehension checks or clarification request to solve communication problems. The role of these communicational tactics or interactional modification is, as Larssen-Freeman and Long (1991, p.144) later argued, " a better candidate for a necessary (not sufficient) condition for acquisition. The role it plays in negotiation for meaning helps to make input comprehensible while still containing unknown linguistic elements, and, hence, potential intake for acquisition."

The earliest version of the Interaction Hypothesis claims that modifying conversational structure while negotiating solutions to communication problems helps make input comprehensible to learners. In addition to simplified input and contextual support, negotiated interaction has been found to be equally important.

However, empirical studies on the role of interaction in acquisition have given rather mixed results. Some studies have shown rather positive evidence of the interactional modification. Pica et al. (1987) and Loschky (1994) proved that interactional modification can improve comprehension of L2 learners, but they failed to prove that increased comprehension can lead to acquisition. While some other researchers (eg. Issidorides and Hulstijn, 1992; Gass and Varonis, 1994) found that modified input and interaction could not promote comprehension and task success of L2 learners. The mixed results of the studies "show a need for a stronger theoretical model clarifying the claimed link between interaction and acquisition" (Mitchell and Myles, 2004, p.173).

With the development of the interactionist research, terms such as selective attention, output and negative feedback are proposed to update the old version of the hypothesis. Long reformulated his version of Interaction Hypothesis in 1996:

> It is proposed that environmental contributions to acquisition are mediated by selective attention and the learner's developing L2 processing capacity, and that these resources are brought together most usefully, although not exclusively, during *negotiation for meaning.* Negative feedback obtained during negotiation work or elsewhere may be facilitative of L2 development, at least for vocabulary, morphology and language-specific syntax, and essential for learning certain specifiable L1-l2 contrasts (Long, 1996, p.414).

> *Negotiation for meaning*, and especially negotiation work that triggers interactional adjustments by the NS or more competent interlocutor, facilitates acquisition because it connects input, internal learner capacities, particular selective attention, and output in productive ways (Long, 1996, pp.451-452).

Mitchell and Myles (2004, p. 174) pointed out that in the updated version of the Interaction Hypothesis, Long "highlights the possible contribution to L2 learning of negative evidence… (and) also highlights the attempt to clarify the processes by which input becomes intake , through introducing the notion of selective attention".
Ellis (2005, p.219) stated that according to the new version of the Interaction hypothesis, "the interactional modifications arising help to make input comprehension, provide corrective feedback, and push learners to modify their own

output in uptake".

As part of interaction, output has also drawn much attention from researchers. Swain proposed the Output Hypothesis (1985, 1995) based on her study of the children in a French immersion class in Canada. She found that input only is not sufficient for language learning. She suggested that the reason why the children in the immersion environment lacked development in their second language after years of study is that they lacked the opportunity to use the language productively. Output pushed the learner to be understood and it is also a learning process. Swain stressed the crucial role of output in language learning and suggested that in addition to the traditional practice function, output has three further functions, they are "noticing, hypothesis–testing and metalinguistic or reflective function" (1995, p.128). Production makes the learners become aware of the gap or problems in their current second language system, which may help the learners "notice the items in input that they did not notice before" or try to 'fill the gap' through a lucky guess, trial and error, use of analogy, first language transfer or problem solving, and the learner may also "deliberately seek to find the item by reference to outside sources like teachers, peers or dictionaries" (Nation, 2007, p.5). Output is different from input and provides different opportunities for learning. Output provides learners with opportunities to experiment with new structures and forms and then maintain or modify them on the basis of feedback. The third function of output is the reflective function, and it provides opportunities to reflect on the problems in their L2. According to Nation (2007), the meta-linguistic (reflective) function of output "involves largely spoken output being used to solve language problems in collaboration with others" (p.6).

Output can also help the development of implicit knowledge. Implicit knowledge is generally regarded as underlying the ability to communicate fluently and confidently in an L2. Ellis (2005) classified the theories of implicit knowledge into skill-building theory and emergentist theories. Although the theories express different opinions on how implicit knowledge develops, "there is consensus that learners need the opportunity to participate in communicative activity to develop implicit knowledge. Thus, communicative tasks need to play a central role in instruction directed at implicit knowledge" (p.210).

Schmidt (1990,1994, 2001) is an influential researcher in promoting the crucial importance of noticing. He uses the term *noticing* to refer to the process of bringing some stimulus into focal attention, whether voluntarily or involuntarily. His strong claim is that "noticing is the necessary and sufficient condition for the conversion of input to intake for the learning" (Schmidt, 1994, p.17), but the more widely accepted is the weaker version of the claim "more noticing leads to more learning" (Schmidt, 1994, p.18). Schmidt's idea is in line with Long's (1996) statement suggesting the important role of attention and Gass's statement that "attention, accomplished in part through negotiation, is one of the crucial mechanisms in this process (of learning)" (1997, p. 132).

Attention is also emphasized in classroom teaching and learning. Interactionists promote the notion of *focus on form*, which is different from *focus on forms*. *Focus on forms* refers to the traditional teaching methods of teaching of grammatical features in accordance with structural syllabus. *Focus on form* refers to noticing of specific linguistic items, as they occur in the input the learners are exposed to. Long (1991, pp.45-46) stated that *focus on form* "overtly draws students' attention to linguistic elements as they arise incidentally in lessons whose over-riding focus is on meaning or communication".

As discussed above, the interactionist research has focused either on the characteristics or functions of input and interaction. A great deal of early research on input and interaction were descriptive, focusing on characteristics of input/interactions. The studies on the relation between different types of language input / interaction and L2 learning have generated mixed results. It seemed that a theoretically stronger linguistic model is needed to link environmental stimuli, the internal system of the learner and L2 language learning.

Input processing theory developed by Van Patten and colleagues (Van Patten, 1996, 2002) is one of those attempts to theorize how environmental L2 input becomes intake. Intake here is defined as "the linguistic data actually processed from the input and held in the working memory for further processing" (Van Patten, 2002, p.757). Input Processing theory tries to explain the processing strategies the learners tend to use when they parse sentences in a restricted way. It offered a series of principles rather than a complete theory or model to explain how learners parse sentences in comprehension. According to the input processing theory, the learners prefer semantic processing over morphological processing. They pay attention to meaning, and content words in the input are processed first. The second principle is that learners process lexical items rather than grammatical items and thirdly, they prefer to process a form that is meaningful or with "high communicative value" rather than a non-meaningful form or a form with a "low communicative value" (Van Patten, 1996, p.24). However, the weakness of input processing theory is that it does not explain how intake is processed further and developed into the inter-language system of the learner.

Mitchell and Myles (2004, p. 191) remarked that "attempts at modelling this interaction are still very fragmentary and incomplete." Although researchers have different ideas on the role of input and interaction, nobody can deny the importance of input and interaction in SLA. More recently, the input and interaction in L2 oral proficiency interviews also arouse much attention in the field.

Recently, the co-constructed interactions between two or more interlocutors have been studied from new and wider perspectives. For example, Zhu Hua (2010) explored how interculturality emerges through interactions among people of different cultural backgrounds. Nakatsuhara (2011) studied the influence of the interlocutor's extraversion levels' and oral proficiency levels' conversational styles in group oral tests. It seems that a mobile process and the co-construction of the interactions is more stressed in recent research.

After the review of the development of the interactionist research on input and interaction, the literature on Accommodation Theory is reviewed in the next section.

2.3 Accommodation Theory

While researchers in the field of SLA explain the features and functions of modified input mainly from the linguistic perspective, Accommodation Theory (AT) attempts to explain the modification or variation in communication from a broader perspective.

Boves (1992) divided the development of Accommodation Theory (AT) into two phases: Speech Accommodation Theory (SAT) and Communication Accommodation Theory (CAT). Speech Accommodation Theory was first proposed by Giles in the early 1970s to explain some aspects of speech variation in interpersonal encounters. Boves (1992) remarked that AT was first a social psychological theory, in which research areas of social perception, impression formation and speech variation are closely related. Then AT was modified and expanded in the following few decades. The up-dated versions are referred to as Communication Accommodation Theory after 1987, which has been moving in a more interdisciplinary direction and the focus has changed from exploring specific linguistic variables to broader mentions of social interactions such as non-verbal variation.

In the present study, the term *Accommodation Theory (AT)* is used as an umbrella term and it includes both SAT and CAT. In this section of the chapter, the development of AT is first reviewed, and then studies on accommodation in oral interview settings are presented, and finally accommodation at lexical level in oral English examinations is fully discussed.

2.31 The Development of Accommodation Theory

When Accommodation Theory (AT) was first proposed by Giles in 1973, the focus of the research on AT lay in "the social psychological research on similarity-attraction which suggests that a person can induce another to evaluate him more favourably by reducing dissimilarities between them" (Giles and Powesland 1975, p.233). It was believed that the reasons behind the accommodation act might be a person's desire to win social approval. According to AT, the accommodative act provides the sender with rewards of the receiver's approval. It can be regarded as an attempt to modify or disguise his or her persona (social identity) to make it more acceptable to the interlocutor. The following is the schema of accommodation .

There is a dyad consisting of speakers A and B

> Assume that A wishes to gain B's approval
>
> A then:
>
> Samples B's speech and
>
> draws inferences of the personality characteristics of B (or at least the characteristics which B wishes to project as being his)
>
> assumes that B values and approve of such characteristics
>
> assumes that B will approve of A to the extent that A displays similar characteristics
>
> Chooses from his speech-repertoire patterns of speech which projects characteristics of which B is assumed to approve. (Giles and Powesland 1975, p.234)

According to Coupland (Coupland and Giles, 1991), Speaker A tries to make his or her speech similar to that of B and thus speech convergence takes place. Convergence refers to "the ways in which speakers modify their language (and other behaviour differences) to reduce differences between them" (p.26) and if B goes through a similar process, then there is mutual convergence. Contrary to convergence, if the speakers modify their speech (and non-verbal behaviours) to increase the difference between themselves and others, then divergence takes place. So the speaker's orientation to the listener can be said to be convergent or divergent. There is also a third state between convergence and divergence named maintenance, in which the speakers do not change their speech and non-verbal behaviours. Most research based on the framework of accommodation concentrates on convergence rather than on divergence and maintenance.

Giles' early research focused on interpersonal accent convergence in an interview situation. He was dissatisfied with Labov's criterion of *attention to speech* to explain the variation in his data. According to Labov (1972), the style of a speaker may be ordered along a spectrum, and it is measured by the attention the speaker paid to speech. The speaker pays maximum attention to speech in a formal context, and minimum attention in informal situations. However, Labov's idea was criticized by Giles because it neglected the psychological factors such as the speakers' attitude and their perception of the communicative situations. Giles (1973) found that casual speech in the interview may have been produced not because of the informality of the context, but the interpersonal influence, such as the interviewer shifted to use more accent or the introduction of certain motive topics. An informal style might be the result of interpersonal accommodation processes.

Bilingual accommodation investigated by Giles, Taylor and Bourhis (1973) in Quebec also provided supportive evidence for speech accommodation theory. In their research, a French Canadian (FC) stimulus speaker provided a message to a bilingual English Canadian (EC) in French (no accommodation), a mixture of English and French and English (full accommodation). The results demonstrated that the more English the FC spoke, or the more the speaker converged, the more favourable evaluation he gained from his EC interlocutor, and in return, those ECs who were spoken to in English converge the most to their FC interlocutor. In this case, mutual accommodation occurred.

Coupland has also contributed a great deal to the development of AT by investigating accommodation in accent. Coupland (1984) investigated a Cardiff travel agent's phonological convergence to her 51 clients of different social-economic and educational backgrounds, and he found accommodation was also related to identity. Sue was a native of Cardiff, working in a travel agent in central Cardiff. Her conversations with 51 native Cardiff clients were tape-recorded and investigated. All the 51 clients are classified into 6 occupational groups according to their socioeconomic status, and four phonological variables are investigated: they are "aitch-dropping", "intervoc. t" realized as voiced or a tap, "g-dropping" and "simplification of final consonant cluster". The results showed that as the percentage of less-standard variation of each variable rises in the clients' speech from occupational class I to V, the percentage of accent variation also rises in Sue's speech with the groups. Both the percentage of clients' variation and Sue's accent variation proves to be a reliable index of her interlocutor's socio-economic and educational background. Coupland interpreted the result as 1) Sue is attempting to match the linguistic features of her interlocutor and 2) Sue is attempting to show her *persona* (identity) is similar to her clients via the phonological variation. So accommodation does not mean simply copying the speech of the interlocutor, but to convey via variations, verbal or non-verbal, an identity which is similar to that which is conveyed by the interlocutor. This is the "interpretive" version of accommodation theory, because it involves a complex interpretive procedure between reception and perception (Coupland, 1984, p.65) .

Accommodation theory has continued developing as research with a wider scope was carried out. Giles and Smith (1979) found that full accommodation is not the best strategy to win the best impression and there is an optimal level of accommodation. In convergence, speech content and speech rate won the highest attractiveness score. Thakerar, Giles and Cheshire (1982) proposed that there is an optimal level for accommodation, and people may have different needs for approval and their motivation for accommodation may also differ from each other. So an increase in the number of motivations underlying accommodation changes was introduced into the theory.

In 1988, the Communicative Accommodation Theory (CAT) was proposed as the up-dated version of Speech Act Theory. One feature of CAT is that "it allows discourse studies to engage with recent theory in social psychology, in line with our attempts to provide a multidisciplinary analysis" (Coupland, Coupland, and Giles, 1991, p. 25). In the new version of Accommodation Theory, speech convergence and divergence and maintenance are termed *approximation strategies*, which is a subcategory of attuning strategies. According to Giles and Coupland (1991, p. 88), there are 4 types of *attuning strategies*: interpretability strategies, discourse management strategies, control strategies and approximation strategies.

Depending on the addressee focus, a speaker may choose an appropriate attuning strategy. Interpretability strategies can be used to modify complexity of speech and increase clarity. For example, simplified syntax and lower lexical diversity are used to decrease complexity. Variation in pitch, loudness, rate and methods such as repetition, comprehension check and explicit boundary devices are used to increase clarity. Discourse management strategies include a variety of discursive options chosen to facilitate the ongoing of the talk, such as offering topics, face-saving strategies, back channelling. Control strategies reflect the role option in talk, for example it concerns interruption and address in conversations.

The theory was first proposed to explain speech variation in interpersonal encounters and later on became a “robust paradigm” (Giles and Coupland, 1991, p.89) and focused on the dynamic communication processes of human beings. It adapted research in sociolinguistics and psychology and has become a more interdisciplinary theory. The accommodation model can be used to interpret the social consequences of interactions, ideological factors, intergroup variables and consequences, discursive practice in natural settings and so on. The change can be seen clearly from the new definition of convergence which is referred to as “ a strategy whereby individuals adapt to each other’s communicative behaviours in terms of a wide range of linguistic / prosodic / non-vocal features including speech rate, pausal phenomena and utterance length, phonological variants, smiling, gaze and so on” (Giles and Coupland, 1991, p.63).

According to Boves (1992), the new model of CAT shifted its focus from approximation strategies to discourse management. Another shift can be distinguished in the new model in different accommodation behaviours such as over-accommodation and under-accommodation. All the strategies occurred in inter-generation interactions, in particular in conversations between young generation and old people, usually with the misperception of young people viewing older people as a prototype of needy, ill and disabled. For example, age-related over-accommodation refers to the “overbearing, excessively directive and disciplinary talk to older people” (Ryan et al., 1986, cited in Coupland, Coupland, and Giles, 1991, p.32) which resembles baby talk.

Regarding the effects of accommodation, Giles and Coupland (1991, p.89) mentioned three aspects: first, it can adjust social distance; second, it can bring people together psychologically, and third, it can facilitate comprehension and keep conversation going smoothly and successfully. Giles and Coupland’s viewpoints on the effect of accommodation is closely related to Ellis’ (1994, p. 264) remark on the major purposes of foreigner talk or L2 adjustment: “to promote communication, to signal the speaker’s attitude towards the interlocutor and to teach the target language implicitly”, but without the function of implicit language teaching.

2.32 The studies of accommodation from different aspects

Since AT was first proposed, many studies have focused on pronunciation, especially the phonological features of an accent. For example, Giles and his colleagues have investigated phonetic variation and they found that a speaker may change the features of his or her accent to those of the interlocutor. Coupland’s famous research in the Cardiff travel agency focused on the less-standard variants of four phonological variables used by different occupational classes and the convergence the clerk used to her clients. Speech rate was also much studied. Webster (1970) and Giles and Smith (1979) both studied speech rate in accommodation, and Giles and Smith (1979) proposed that optimal accommodation of speech rate and content are the most important factors to win the best impression of the interlocutor.

Accommodation also occurs during communication of bilinguals. The bilingual interlocutors may converge to or diverge from each other, which is closely related to the speaker’s impression or attitude towards his or her interlocutors. In earlier mentioned research of Giles, Taylor and Bourhis (1973), it was found that the more the French Canadians and English Canadians converge on each other, the more favourable opinions they have for each other and more convergence takes place. On the other hand, there was also research showing that divergence appeared between bilingual interlocutors. In the experiments of Bourhis and Giles (1977), some Welsh learners of Welsh were asked questions by an RP-sounding speaker in the booths of a language laboratory. When the English speaker called Welsh a dying language without a future, the learners generally diverged by broadening their Welsh accent. The Welsh learners broadened their Welsh accent to show disapproval or even anger to the English speaker who showed a negative attitude towards the Welsh language.

More recently, there have been some new trends in AT research. For example, Gregory and Webster (1996) studied the pitch patterns of the talk show host Larry

King and his guests. It was found that Larry King changes the pitch of his sound according to the status of his guests, altering his normal pitch to accommodate to guests of higher status, such as President Clinton; on the other hand, guests of lower status such as the actor Sean Connery accommodate their pitch to Larry King. Malvern and Richards and their associates (Richards and Malvern, 2000; Malvern and Richards, 2002; Duran et al., 2004) have carried out a series of research projects on accommodation of lexical richness in oral interviews in the past 15 years. Since lexical richness and accommodation are also the focus of the present research, studies on accommodation in oral interviews, especially studies on accommodation at lexical levels, are reviewed in more detail in the next section.

2.33 The studies of accommodation in oral proficiency interviews

In oral examinations, the examiner varies his or her speech to adapt to the perceived level of the examinees, which is in a way similar to what happens in modified L2 input or foreigner talk. Accommodation occurs at different levels in oral interviews. In addition to the accommodative discourse of examiners, new interest has shifted to accommodation at the lexical level along with the return of attention to vocabulary in SLA research.

Oral proficiency interviews (OPI), which became popular as a new form of language proficiency test in the 1980s, have drawn much attention from linguists and researchers. According to Silverman (1976, cited in Ross and Berwick, 1992, p.161), an interview is defined as “a scheduled encounter between unequal participants in which one or more persons have vested rights to ask questions and organize the topic and talk”. The majority of studies on oral proficiency interviews in the 1980s and 1990s focused on the features of the interviewer and candidate discourse, and in particular, the authenticity of the interaction of the interview (Ross and Berwick, 1992, Young and Milanovic, 1992; Lazaraton, 1996; He and Young, 1998). The extent to which the interview resembles natural conversation is regarded as a vital issue related to the validity of oral interview as a means of assessing L2 proficiency.

Young and Milanovic (1992) conducted a quantitative study on the interview section of the Cambridge First Certificate in English examination. They examined variables such as dominance, contingency and goal orientation (i.e. quantity of talk, topic initiations, reactiveness and topic persistence) as well as contextual factors (interview theme and task, examiner gender), and finally came to the conclusion that the discourse and interaction between interlocutors in oral proficiency interviews is highly asymmetrical. The results also showed that while interviewers show much greater goal orientation than students, it is the students who show greater "reactiveness" or conversational contingency to the language of their interlocutor.

Among the research, one of the focuses is the examiners' accommodation to the interviewees or candidates. Accommodation occurs in natural conversations and it can be the result of desire for social approval or efficiency of communication.

Ross and Berwick (1992) carried out a study to investigate the discourse of accommodation in oral proficiency interviews conducted in Japanese corporate settings. The focus was on the way interviewers accommodate to their candidates, the features of the accommodation and also the role of accommodation as a potential alternative of rating criteria.

Sixty full length Oral Proficiency Interviews (OPI) chosen from each of the four different rating categories ranging from 1+ to 3 (intermediate-High through Superior) were conducted by 12 professionally trained interviewers. Each interview lasts from 15 to 30 minutes. Detailed analysis of the language used by the interviewer during the course of interview was carried out and the relationship between interviewer language and the final scores of the candidates was analyzed. The main research questions were first, to what extent does the use of accommodation reflect the interviewers' mobile perception of the candidates' proficiency; and second, how does the interviewer perception accord with the foreigner talk in non-natural settings. 14 conversational modification exponents categorized into two groups were analyzed: 4 control exponents and 10 accommodation exponents. Results showed that accommodation exponents discriminate among ratings and the most frequently used accommodation exponents of or-question, slowdown and display questions represented teacher talk, which is a variety of foreigner talk. Finally the researchers draw conclusions: first, the OPIs share qualities of both conversation and interviews; second, exponents of control play a functional role without any special sensitivity to the interviewee's ability when engaged in conversation, but exponents of accommodation are especially sensitive to the interviewer's perception of the interviewee's current level of oral proficiency; finally, the authors claimed that accommodation provides a potential useful source for rating criteria of proficiency, but over-accommodation should be avoided.

Ross (1992) narrowed down her research from accommodative discourse in interviews to accommodation questions raised by interviewers at key junctures in the interview process. She argued that the examiner's perceptions of the examinees' oral proficiency are reflected in the extent of accommodation in interviewer questioning, and that the extent of accommodation may provide an additional criterion for assessing proficiency.

In order to answer the question of what triggers interviewer accommodation, a detailed study of 16 oral proficiency interviews conducted by trained interviewers was carried out. The candidates were Japanese company employees enrolled in a training program. The 16 interviews were selected from four most common ratings from 1+ (high intermediate), 2, 2+ to 3(superior), four audio-recorded interviews were selected from each category. Seven types of accommodation strategies were used to classify interview questions:

1) Display question: the interviewer asks for information known to the interviewer or that the interviewee ought to know, as perceived by the interviewer.
2) Or-question: the interviewer asks a question and provides options for the interviewee.
3) Fronting: the interviewer foregrounds a topic or sets the stage for the interviewee's response.
4) Grammatical simplification: the interviewer simplifies the utterance to facilitate comprehension.
5) Slowdown: the interviewer reduces the speech rate.
6) Over-articulation: the interviewer exaggerates the stress and pronunciation of

words and phrases.

7) Lexical simplification: the interviewer chooses words or phrases which he or she believes to be simple rather than difficult (Ross, 1992, p.176) .

In addition, discourse features from the immediately preceding turn were also coded and examined. Factors considered as potential triggers of accommodation are:

1) interviewee response to previous question
2) structure of response to previous question
3) the foregrounding of the current discourse topic
4) the perceived level of the interviewee; whether or not the interviewee gave a comprehensible answer or statement
5) the last speaker in the previous turn/ the outcome of the interview
6) accommodation in the previous question (Ross, 1992, p.176).

Results showed that the most clearly influential factors contributing to the occurrence of accommodation were the interviewee's response to the previous interview question, the structure of the interviewee's response to the previous question, the perceived level of the interviewee and whether or not accommodation had been used in the previous question. Finally Ross proposed that in addition to abstract definitions of proficiency, the manner and quantity of interview accommodation should be considered in the assessment process, and the degree of necessary accommodation might provide a useful dimension for assessment.

Lazaraton (1996) presented a qualitative analysis of the types of linguistic and interactional support that the native examiner provides to the non-native speaker candidates in an oral interview. Results indicated that the native examiners applied at least eight types of interlocutor support in the corpus of 58 transcribed Cambridge Assessment of Spoken English (CASE) interviews. The support from the native examiner include: 1) priming topics; 2) supplying vocabulary or engaging in collaborative completions; 3) giving evaluative responses; 4) echoing and/or correcting responses; 5) repeating questions with slower speech, more pausing and over-articulation; 6) stating question prompts as statements that merely require confirmation; 7) drawing conclusions for candidates and 8) rephrasing questions.

It is also suggested the practice of examiner support is positive in the sense that conversational practices are present in this assessment context and there is a kind of naturalness in oral interviews. On the other hand, there are concerns about the inconsistent support from examiners and the impact of the support on candidate performance and assessment. Lazaraton also suggested that the role of the examiner in the test interaction should be taken into account in the rating procedure and also shed light on interviewer training.

Lazaraton concluded that there are still a lot of questions which remain unanswered in this area. She proposed that "it is critical that more studies on oral test interaction, whether they be statistical or discourse analytic or, ideally, both, be undertaken on other widely used proficiency examinations", so we will have a better understanding of the "validity of oral proficiency assessment itself" (Lazaraton, 1996, p.167).

Among the research focusing on the variation of the interviewers, Brown (2003) conducted research on how personal styles affect the performance of the candidates, and the rater's perception of the candidate's ability. In Brown's study, each of the two interviewers conducted an interview with the same candidate on the same day. The two interviewers had been regarded as the easiest and the most difficult interlocutors by independent raters. Conversational analysis was used to study their discourse in the interview. Results show that the raters are more positive about the candidate's performance when she was interviewed by *teacherly* interviewer than with the casual interviewer. *Teacherly* interviewer is the easier one who behaves more like a language teacher in class who adapts her language to the students all the time. The difference between the two types of interviewers lies in first, they apply accommodation strategies to the candidates, and second, they awarded different scores to the same candidate. So the author concluded that more emphasis should be laid on the importance of interviewer training. It raised questions regarding the appropriate level of accommodation and also questions regarding the level of accommodation to be specified during interviewer training to maximize the reliability and validity of oral interviews.

Nakatsuhara (2011) examined the influence of the interlocutor's extraversion level and oral proficiency level conversational styles in group oral tests between two group sizes: groups of three and groups of four. The research data were collected from 269 Japanese students, who took group oral tests either in groups of three or four. Both quantitative and qualitative analyses were conducted on the data. It was found that the extraversion level played a more important role in groups of four than in groups of three; there was an influence of the proficiency-level variables in both group sizes, but the size effect was greater in groups of three than in groups of four. Conversational analysis helped reveal reasons behind the difference and explain the relationship between these impacts and co-constructed group interactions.

O'Loughlin (2002) investigated the impact of the gender of IELTS interviewers; however, both the discourse and test score analyses indicated that gender did not have a significant impact on the interview.

2.34 Lexical accommodation in oral examinations

Richards and Malvern and colleagues (Richards and Malvern 2000; Malvern and Richards 2002; Duran et Al. 2004) have done much pioneering research on the accommodation of the non-native speakers. They were particularly interested in accommodation at the lexical level, which combines the research into lexical richness and accommodation in the field of SLA.

Richards and Malvern (2000) studied oral examinations in French. In the oral exam of French GCSE (General Certificate of Secondary Education), examiners are non-native speakers of French and the examiner is also the teacher of the examinees. Interviews with 34 16-year-old learners of French with 2 non-native teachers were analyzed. Three types of student variables and teacher variables were analyzed. The first type of student variables are objective measures taken from the transcripts with the assistance of CLAN software, including number of words, number of utterances, number of different words, vocabulary diversity (MSTTR-30), Mean Length of Utterance measured in words (MLU words), utterance per turn (MLT) and words per minute. The second type of student variables come from the final results of GCSE examination including the score for oral examination (out of 7) and total score for listening, reading & writing (out of 21), and the third type of student variables are 6 further measures obtained from the mean rating of the tape recording by 24 experienced teachers of French: range of vocabulary (0-7scale), fluency (0-7scale), complexity of structure (0-7scale), content (0-3scale), accuracy (0-3scale) and pronunciation (0-3scale). There are seven teacher variables obtained directly from the transcripts, including vocabulary diversity (MSTTR-100), MLU words, percentage of utterances which overlap with the student, percentage of utterances which are imitations of the student, percentage of utterances which are exact/expanded/reduced imitations of the student, percentage of utterances which are back channels and length of conversation (seconds). After careful analysis of the variables, the researchers concluded that measures of the students' French are related to indices of the teachers' language and accommodation to the proficiency of individual students does take place, but they also found that the teacher variable of vocabulary diversity is grossly tuned to the general proficiency of students in their class rather than finely tuned to individuals.

In a follow-up study Malvern and Richards (2002) used a new measure of lexical diversity *D* to investigate accommodation with the same data. This time they added "student *D*" and "teacher *D*" to the former student variables and teacher variables respectively. The authors claimed that the new measure of lexical diversity *D* and the software *vocd* to produce it is a valid tool for language data analysis. *D* is strongly correlated with other measures of vocabulary diversity and independent of the quantity of spoken discourse. Second, teachers are not fine tuned to the language level of individual students. They are controlling their lexical diversity and tuned to the level of the whole class.

In this study, Malvern and Richards proposed that the lack of appropriate accommodation might be a threat to the validity of the GCSE examination. They stated that "it does, however, introduce the very threats to validity identified by Ross and Berwick (1992), the first of which is the absence of appropriate accommodation"

(2002, p.101).

It is considered a problem facing all oral examiners: without appropriate accommodation, the students (especially students with low proficiency) will have difficulties in comprehension and will have difficulties in communication. Yet on the other hand, if the examiner gives unnecessary accommodation to students, especially students with medium or high proficiency, then he or she will not stretch the students to show their proficiency to the fullest extent. Both under-accommodation and over-accommodation will affect the performance of the students. All examiners of oral examinations are facing a dilemma: that is how to keep a balance between being reliable and fair to all candidates on the one hand and being valid and fine-tuning to individual candidates on the other.

Lorenzo-Dus and Meara (2005) focused on the relationship between examiner accommodation and the candidate vocabulary. They selected one key aspect of examiner performance (use of support strategies) and one crucial area of test-taker performance (vocabulary) and linked them to the relevant score outcome.

30 Spanish oral examinations were analyzed from qualitative and quantitative viewpoints. For each examination the following data were collected: test-taker's number of word types, lexical diversity (D); the examiner's number of word types, lexical diversity (D); the examiner's total accommodation strategies and vocabulary accommodation strategies.

By the total accommodation strategies of the examiner, the authors used strategies identified in the literature of Lazaraton (1996) and Brown (2003). The total accommodation strategies include:

Topic priming

Correct talk

Repeat question

Formulations

Asking for additional information to a previous question

Asking for additional information to an earlier prompt

Rapport building

Clear marker of topic change

Supply or complete vocabulary

Simplify a question or statement

Confirmation questions

(Lorenzo-Dus and Meara 2005, p.245)

Among the total accommodation strategies, there are three subsets of the strategies that had direct impact on test-taker's vocabulary output and were named vocabulary accommodation strategies, including Simplify Question or Statement (SQS), Supply or complete Vocabulary (SCV) and Confirmation Questions (CQ)

The results of the qualitative analysis indicate that the relationship between accommodation and test-taker vocabulary is not straightforward at all. It is found that the vocabulary output cannot explain some of the grades awarded. Only the number of *types* of the test takers discriminates between grades for vocabulary. The higher the number of types generated by the test takers, the higher the grade for vocabulary awarded. Second, there is more vocabulary variation in the high level grade bands than in the low level grade bands. The frequency analyses show that the total number of examiner accommodation strategies discriminate between grades for vocabulary, with more instances of strategy use in the examination that receives lower grades and vice versa. Among the subset of vocabulary strategies, only Confirmation Questions (CQ) fails to discriminate across the grade bands. The other two vocabulary strategies can discriminate between candidates of different grade bands. Finally the authors proposed that an integrative approach that combines both statistical and qualitative analysis is the optimal framework within which to conduct research on oral interviews.

What is in common in the above-mentioned research of Richards and Malvern and colleagues (Richards and Malvern, 2000; Malvern and Richards, 2002; Duran et al., 2004) and Lorenzo-Dus and Meara (2005) is that they both focus on the accommodation of examiners and lexical diversity measures of test-takers. Another common feature among the above mentioned studies is that the data the researcher collected is not big, with a couple of examiners and less than 50 candidates, and the data were obtained on the basis of availability. They were not random sampling data from a large-scale corpus. Aware of the limitations of previous studies, the present thesis collected data on a random sampling basis, involving 180 candidates and 21 examiners. It is hoped that a more extensive data can show a more complete picture of the lexical richness and accommodation in widely used oral interviews.

Another common point in Malvern and Richards (Richards and Malvern 2000; Malvern and Richards 2002; Duran et al. 2004) and Lorenzo-Dus and Meara (2005) is that lexical diversity can reflect the active vocabulary that can be used by the speaker, and it is an indicator of a learner's lexical richness. Research on the measurement of vocabulary richness is the focus of the next section.

2.4 Vocabulary and lexical richness

2.41 The important role of vocabulary in L2

Krashen (1989, p. 440) once argued that，"a large vocabulary is for mastery of a language. L2 learners know this … they carry dictionaries with them, not grammar books." Since the 1980s there has been a trend in L2 research fields, that is, a rediscovery of vocabulary in Meara's words (Meara, 2002) and an explosion of publications on vocabulary (e.g. Carter,1988; Carter&McCarthy,1988; Gairns & Redman, 1986; Hatch & Brown, 1995; Nation, 1990; Schmitt & McCarthy,1997). From then on, the attention of many researchers has turned from syntax to lexicon.

According to Singleton (1999), the above quotation implies that the major challenge of learning and using a L1 or L2 does not lie in the general syntactic rules but in the lexicon. "Lexical knowledge is now known to be an absolutely crucial factor across the whole spectrum of L2 activities" (pp.4-5).

More recently, from 1999 to 2010 there have also been a series of books on vocabulary. Table 2.3 is a list of books published in the past 15 years. Among those works, some are focused on L2 vocabulary acquisition and the more recent ones are on vocabulary assessment and lexical richness.

In addition to published books, there has also been a boom in vocabulary research published in journal articles in the past 20 years (Nation, 1990, 2001;

Laufer, 1995; Richards and Malvern, 2000; Malvern and Richards, 2002; Duran

Table 2.3 A list of books on vocabulary research

Author (s)	Year	Title	Main content	Publisher
Singleton, D.	1999	*Exploring the second language mental lexicon*	L2 vocabulary acquisition	Cambridge University Press
Read, J.	2000	*Assessing Vocabulary*	Vocabulary assessment	Cambridge University Press
Schmitt, N.	2000	*Vocabulary in language teaching*	L2 vocabulary acquisition	Cambridge University Press
Nation, I.S.P.	2001	*Learning vocabulary in another language*	L2 vocabulary acquisition	Cambridge University Press
Malvern, D., Richards, B., Chipere, N., and Duran, P.	2004	*Lexical diversity and language development: Quantification and assessment.*	Vocabulary Assessment	Palgrave Macmillan.
Daller, H., Milton, J. and Treffers-Daller	2007	*Modelling and assessing vocabulary knowledge.*	Vocabulary Assessment	Cambridge University Press
Richards, B., Daller, M.H., Malvern, D.D., Meara ,P., Milton, J. & Treffers-Daller, J. (eds.)	2009	*Vocabulary studies in first and second language acquisition the interface between theory and application.*	Vocabulary assessment and acquisition	Cambridge University Press.
Milton, J	2009	*Measuring Second Language Acquisition*	Vocabulary assessment and acquisition	Multilingual Matters
Schmitt, N.	2010	*Researching vocabulary: A vocabulary research manual*	Vocabulary research methodology	Palgrave Macmillan

et al., 2004; Laufer and Goldstein, 2004; Long and Richards, 2007; Daller, van Hout, and Treffers-Daller, 2003; Meara, 2005; Yu, 2009 etc.). Vocabulary knowledge is now regarded as a key component of L2 proficiency. Vocabulary has been found to have a high correlation with reading (Laufer, 1995; Albrechtsen, Haastrup, and Henrisken, 2008), and according to the results of Laufer and Goldstein (2004), knowing the form-meaning link of words accounts for 42.6% of the total variance in the subjects' language grade in class in a regression analysis. In addition, lexical knowledge is also widely accepted as the main prerequisite of bilingual children. (Daller, 1999; Schoonen, 1993, 1998; Vermeer, 1997).

Read (2000) argued that vocabulary learning is one of a range of goals that are important in the language classroom. He listed both the general and specific goals for language learning and believed that vocabulary is one of the four specific goals under the general goal of language item. Vocabulary learning is a very important aspect of language learning.

Table 2.4 General and specific goals for language learning

General goals	Specific goals
Language items	pronunciation, ***vocabulary***, grammatical and constructions
Ideas/ (content)	subject matters knowledge, cultural knowledge
Skill	accuracy, fluency, strategies, process skills or sub-skills
Text/(discourse)	conversational discourse rules, test schemata or topic scales

(adopted from Read 2000, p.1)

In order to communicate with an L2 speaker, a considerable amount of vocabulary is necessary for language use. According to Nation and Waring (1997, pp.7-10), knowledge of 3,000 to 5,000 word families could provide a very good comprehension repository for L2 learners. So it is assumed that 3,000 to 5,000 word families are needed for intermediate to more advanced L2 learners who need to communicate with native speakers and read or listen to original texts for study and work purposes. Hazenberg and Hulstijn (1996) suggest a minimal vocabulary size of 10,000 Dutch base words for university studies in Dutch. Schmitt (2010, p.6) argues that "a range of 16,000-20,000 word families seems a fair estimate of the vocabulary size for an educated native speaker". As a result, for the L2 who wishes to achieve near-native proficiency, he or she should also have a very large vocabulary that is close to a native speaker. However, since the research on vocabulary size counts vocabulary with different methods, for example, some researchers used word families and some used base words, there is some discrepancy in the literature. This thesis will give a detailed description on what is counted as a word in the present research in the chapter on methods.

In addition to the enormous size of vocabulary, depth of word knowledge is also required to master an L2. In each word, there are so many properties to grasp. Read (2000, p.25-28) presented different types of word knowledge proposed by different researchers. For example, Richards (1976) listed seven assumptions in an attempt to cover all the aspects of what is meant by knowing a word.

While researchers all recognize the importance of vocabulary knowledge, they have different ideas on how to assess vocabulary. According to Read (2000, p17), vocabulary ability not only involves merely knowing some lexical items, the learners "must have ready access to that knowledge and be able to draw on it effectively in performing language-use tasks". Before discussing how to assess vocabulary and considering reliable and valid assessment measures for vocabulary knowledge, two questions should be answered: first, what is considered as a word? Second, what is involved in knowing a word? In the next section, key terms on word knowledge are presented.

2.42 The definition of word

"Word is not an easy concept to define, either in theoretical terms or for various applied purposes" (Read, 2000 p.17). It is usually believed that "according to the context and need, researchers can consider *types, tokens, running words, lemmas* and *word families* as words (Daller, Milton and Treffers-Daller, 2007, p.3).

What is a word? A definition is not as easy as people often assume. *Word* is traditionally defined as the smallest element that may be uttered in isolation with semantic or pragmatic content (with literal or practical meaning). Bloomfield introduced the concept of "Minimal Free Forms" in 1926. Words are thought of as the smallest meaningful unit of speech that can stand by itself. However, some written words are not minimal free forms, for example, *the* and *of* , as they make no sense by themselves. In practice, word may refer to token, Type, lemma and word families, which is often confusing. In applied linguistics, it is unavoidable to meet the problem of what is being counting as a word in vocabulary research, thus it is necessary to explore the distinction between some basic terms.

2.421 Tokens vs. Types

Read (2000, p.18) stated that tokens and Types are both units that can be applied to the count of words in a text. Distinction should be made between them in different situations. The number of tokens is "the total number of word forms, which means that individual words occurring more than once in the text are counted each time they are used". The number of Types means "the total number of different words forms, so that a word which is repeated many times is counted only once".

In a text, individual words that occur more than once are counted each time they are used. For example, in the sentence *the boy was crying when the door bell rang*, there are totally 9 word forms , but the word *the* appears twice, so there are 8 different word forms. There are 9 tokens and 8 types in the above sentence.

2.422 Function word vs. content words

It is generally agreed by linguists that *articles, prepositions, pronouns, conjunctions, auxiliaries,* etc. might be regarded as function words, because they are regarded as having no notions of their own and are belong more to grammar than to vocabulary. Their chief function is to express the relation between notions, the relation between words as well as between sentences.

Yet words of nouns, full verbs, adjectives and adverbs are regarded as content words. Most of the English words are content words and are the focus in many vocabulary tests. For example, in the sentence *the book is extremely exciting, the* and *is* are functional words and *book, extremely* and *exciting* are content words.

2.423 Lemma Vs. Word family

As for content words, they may have different forms. For example, a noun may have a plural form and a verb may have third personal singular, past tense and participles. Should the different forms of a content word be regarded as one word or not? According to Read (2000, p.18), "the base and inflected form of a word are collectively known as a lemma". Usually all the items included under a lemma are the same part of speech. The English inflections consist of plural, third person singular, past tense, past participle, *-ing*, comparative, superlative and possessive (Bauer and Nation, 1993). For example, the verb of *help, helps, helping* and *helped* are normally regarded as the different forms of the same lemma *help*.

In addition to inflections, words may also have derivatives that may change the word class and change the meaning of base word. Word family is another term that is applied to define words with relations. Word family has a larger scope than lemma. For example, the words *happy* and its inflected forms and derivative forms such as *happier, unhappy, happiness* and *happily* belong to one *word family*. Although there is change in the word class and the meaning compared to the base word *happy,* all these words are still closely related in form and meaning. "Such a set of forms, sharing a common meaning, is known as a word family" (p.19).

The reason to distinguish between the above mentioned terms is that researchers count words differently for their research purposes. Thus those who conduct research on vocabulary acquisition and assessment or those who compile word lists for teaching or testing have to define what they mean by a word. In the present research, words are counted mainly by Type and token.

2.43 Different dimensions of vocabulary knowledge

Vocabulary or lexical knowledge can be described either from a global aspect (with one or two dimensions) or from many different aspects including all aspects of word knowledge (e.g. Nation, 1990; Richards, 1976). Richards listed seven dimensions of knowing a word (1976, p.83):

1) Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words we also know the sort of

words most likely to be found associated with the word.

2) Knowing a word implies knowing the limitations on the use of the word according to variations of function and situation.
3) Knowing a word means knowing the syntactic behaviour associated with the word.
4) Knowing a word entails knowledge of the underlying form of a word and derivations that can be made from it.
5) Knowing a word entails knowledge of the network of associations between the word and other words in the language
6) Knowing a word means knowing the semantic value of a word.
7) Knowing a word means knowing many of the different meanings associated with a word.

Nation (1990) divided word knowledge into 4 categories: form (spoken form and written form), position (grammatical pattern and collocation), function (frequency and appropriateness), and meaning (concept and association), each category comprising two subcategories. Learning a word means to grasp these eight properties.

Read (2000, p.7) also argued that “knowing a word is taken to include not only knowing the formal aspects of the word and knowing its meaning, but also being able to use the word”. Nine aspects are involved in knowing a word : “spoken form; written form; concept and referents; word parts; connecting form and meaning; associations; grammatical functions; collocations; constraints on use” (p.36-59).

From the above aspects of lexical knowledge given by different researchers, it can be seen clearly that vocabulary knowledge is a rather complicated system and learning a word is a complicated process rather than simply memorizing the basic meaning of the word.

However, Meara (1996, p.3) argued that it is impracticable to measure all the attributes of individual words although it is theoretically reasonable. He therefore proposed a model of lexical competence with only two dimensions: *size* and *organization,* which are believed to be independent of each other.

Wesche and Paribakht **(**1996) also argued that there were two aspects of lexical knowledge, but in addition to *breadth*, which is similar to Meara's *size* dimension, they have another dimension of *depth* versus *size (breadth)*. The existing measures of vocabulary *size (breadth)* are uninformative as to the quality of lexical knowledge (*depth*) that learners have about particular words.

Henriksen (1999) tried to balance between the global and the separate trait view and proposed three dimensions of vocabulary knowledge: (a) a partial-precise knowledge dimension (b) a "depth of knowledge" dimension, and c) a receptive-productive dimension. According to this author, the three dimensions proposed reflect the vocabulary development process, and the three dimensions are three continua of the development of a learner's vocabulary.

Meara (1996, pp.3-12) proposed that "lexical competence is probably not just the sum of speakers' knowledge of the items their lexicons contain", instead it might be described "in terms of a very small number of easily measurable dimensions" despite the complexities of the lexicon, the dimensions of *size* and *organization* which are "properties of the lexicon considered as a whole". According to him, size is the best measure of overall vocabulary knowledge up to a level of 5,000-6,000 words in the case of English. He developed a checklist test to measure *size*, which consists of a set of real words and a set of imaginary, non-existent words. The test-taker is asked to identify which of these words they actually know. There is a substantial number of non-words in the test, and if the test-taker claims to know non-words, it means the test-taker is over-estimating his or her vocabulary knowledge and the score is adjusted accordingly. The checklist test has proved an ability to estimate with some degree of accuracy how many of the real words a test-taker knows. Meara (p.13-14) also proposed a method by asking a test-taker to produce chains of associations to connect pairs of words chosen at random to infer the degree of connectivity in a lexicon. For example, the association chains between *sea* and *butterfly* might be *Sea ... blue ... sky ... fly ... butterfly*. For native speakers, there is a higher degree of interconnection than in an L2, and "those who have a more developed vocabulary knowledge have a more complex and highly structured network of associations among the words they know" (Read 2000, p.248). Size and Organization are expected to estimate two independent dimensions: how large is the size and how structured is the vocabulary. "They are characteristics of the system as a whole, rather than features of the individual words that make up the system."

Meara reminded us of a practical question in vocabulary assessment: how should we use a few dimensions of global properties to assess a learner's complex vocabulary knowledge in reality? Because of the many dimensions of vocabulary knowledge, it is not straightforward to define the meaning of "knowing a word".

Daller et al. (2007) tried to summarize vocabulary knowledge in a theoretical three-dimensional space that composed of *breadth* , *depth* and *fluency*. Details are absent in this model, but breadth and depth are the passive vocabulary knowledge dimensions and fluency, which reflects the ease and speed of accessing and using vocabulary, is the active vocabulary knowledge dimension.

Milton (2009) also remarked that it does not seem possible to have a measure that could evaluate every aspect of vocabulary knowledge. Researchers usually chose a workable method to measure one or more aspect of vocabulary knowledge.

2.44 Dimensions of vocabulary assessment

Read (2000) proposed three dimensions of vocabulary assessment, which may present the development of researchers' understanding of it and help bring in more vocabulary assessment measures.

The three dimensions of vocabulary assessment are 1) Discrete – Embedded dimension, 2) Selective – Comprehensive dimension and 3) Context-dependent – context-independent dimension. According to Read (2000), discrete tests assess vocabulary as a construct independent from other aspects of language ability, while an embedded measure of vocabulary "is the one that contributes to the assessment of a larger construct" (p.9). For example, vocabulary assessment is embedded in oral proficiency interviews, in the assessment of the performance of language tasks of the learner.

The selective test assesses a range of vocabulary chosen by the test writers, while the comprehensive test assesses the overall vocabulary use in a spoken or written test of a test taker.

If a word is presented “as an isolated element”, the test is context-independent. From a more contemporary perspective, a content-dependent test assesses vocabulary in discourse. The assessment of vocabulary is developing from traditional discrete, selective and context-independent measures to ones that are more embedded, comprehensive and context-dependent. Some comprehensive measures of vocabulary were proposed to judge candidates’ vocabulary abilities and the characteristics of input text both in written and spoken discourse.

Table 2.5 Three dimensions of vocabulary assessment

Dimensions of vocabulary assessment	
Discrete A measure of vocabulary knowledge or use as an independent construct	**Embedded** A measure of vocabulary which forms part of the assessment of some other, larger construct
Selective A measure in which specific vocabulary items are the focus of the assessment	**Comprehensive** A measure which takes account of the whole vocabulary of the input material or the test-taker’s response
Context-independent A vocabulary measure in which the test-taker can produce the expected response without referring to any context	**Context-dependent** A vocabulary measure which assesses the test-taker’s ability to take account of contextual information in order to produce the expected response

(Adopted from Read 2000, p.9)

Measuring lexical richness in oral proficiency interviews is a type of vocabulary assessment from a more embedded, comprehensive and context-dependent dimension.

2.45 Lexical richness

According to Read (2000), lexical richness is the general term for those lexical measures that are used for the characteristics of effective vocabulary use by the learner. Most research on lexical richness is concerned with evaluations of writing.

Linnarud (1986) applied measures of lexical richness in analyzing written compositions of native and non-native speakers: *lexical individuality, lexical density, lexical variation* and *lexical sophistication*. *Lexical individuality* means words used by one writer only; *lexical density* refers to the percentage of lexical (content) words in the text; *lexical variation* is the type/token ratio and *lexical sophistication* means the difficult words that are beyond the level of instruction in classroom settings.

Read (2000) proposed that there were 4 aspects of lexical richness in writing compositions: *lexical variation, lexical sophistication, lexical density* and *number of errors*. Good writing is assumed to have the following lexical features:

First, "a variety of different words rather than a limited number of words used repetitively". This is the aspect of "*lexical variation*", and it is often referred to as "range of expression" in assessment criteria (p. 200). Traditionally it was measured by type/token ratios, or the percentage of different words in the total number of words in a text.

Lexical sophistication or *rareness* is an indication of level of difficulty of the words. Good writing is assumed to have "a selection of low frequency words that are appropriate to the topic or style of the writing, rather than just general, everyday vocabulary and it allows writers to express their meanings in a precise and sophisticated manner" (Read, 2000, p. 203). It in fact refers to the number of low frequency words which are considered difficult.

Lexical density is a measure that distinguishes the lexical (content) words and grammatical (function) words. Good writing is expected to have a high percentage of lexical (content) words. Lexical density is usually considered as the characteristic that distinguishes written from spoken language. It is also the percentage of lexical (content) words in the text, which is similar to the term applied in Linnarud's research. *Number of errors* means the number of mistakes counted in the writing. Good writing composition is assumed to have few errors in the use of words. But in reality it is very difficult to define what a lexical error is, or to distinguish a lexical error from a grammatical error or pragmatic inappropriateness; it is seldom used in research of lexical richness nowadays.

It can be seen from the definitions of lexical richness of Linnarud (1986) and Read (2000) that the common components of lexical richness are lexical variation, sophistication and density. But as Yu (2009) mentioned, the term *lexical richness* is confusing sometimes, because it is frequently used interchangeably or overlaps with lexical diversity, vocabulary diversity, lexical sophistication or rareness, vocabulary richness and so on. Since researchers use different terms to address lexical richness, and they conducted their research based on different data, there is no consensus on the exact definition of lexical richness, nor consensus on the different aspects for lexical richness. In the present research, the definition of Read (2000) is adopted and lexical richness is used as a general term. Sub-terms of Lexical variation and sophistication are the focus for research.

Researchers have proposed different methods to measure lexical richness in the past 15 years or so and recent research turns more attention to lexical richness in spoken texts. (Vermeer, 2000; Read, 2000, 2001; Malvern & Richards, 2002, Malvern et al. 2004; Duran et al., 2004; Jarvis, 2002; Laufer & Nation, 1999; Daller, van Hout & Treffers-Daller, 2003; Meara, 2005; McCarthy and Javis, 2010; Lu, 2011). Daller and Xue (2007) divided the different measures of lexical richness into two types: word-list free and word list-based measures. All word-list based measures focus on lexical sophistication, whereas the word-list-free approaches focus on lexical variation or diversity. The most frequently used measures of lexical diversity such as the traditional TTR (the ratio between different words and the total number of words in a language sample) and other transformations of TTR are believed to be problematic. Therefore some researchers proposed new measures of lexical richness and claimed that theirs were more valid than others when measuring lexical richness in spoken or written discourse: for example the mathematical measure ***D*** (a parameter in the equation that relates TTR and text length N) proposed by Malvern and Richards (2002) and Guiraud Advanced (AG) proposed by Daller, van Hout and Treffers-Daller (2003). In the following section, recent studies on lexical variation or diversity and lexical sophistication are reviewed.

Lexical variation or diversity was measured by type/token ratio (TTR) and TTR is calculated by dividing the numbers of different words (types) by the total number of words (tokens). It was widely used in child language development research and later also used in L2 research in the 20th century. It can be seen from the earlier part of the chapter that most research on input and interaction in SLA in the 1980s and 90s used TTR as a measure to describe the modification or accommodation in interactions in the lexical aspect (eg. Ellis, 1994; Long, 1991).

TTR has been criticized by many researchers (Vermeer 2000; Malvern and Richards, 2002; Malvern et. al 2004 and Daller, van Hout and Treffers-Daller 2003) in the past 15 years for its instability. It is very sensitive to text length. As more words are introduced to the text, TTR will decline. Jarvis (2007) explained the instability of TTR by reference to Heaps's law, which predicts that the more words (tokens) a text has, the less likely it is that new words (types) will appear. Thus, "the diminishing returns of new types flaw the most commonly used LD metric, the TTR" (p.460). When TTR is used to compare two texts, the longer one generally gets a lower TTR, which means the longer one is lexically less diverse. So when comparing the lexical diversity of texts with different sample size, TTR will give very faulty results.

Many other measures based on TTR or the transformation of TTR, were proposed, but van Hout and Vermeer (1988) concluded on the empirical results that the index Guiraud seemed most stable for learner data. Guiraud (*G*) is also thetransformation of TTR and was expected to keep stable TTR over a longer sample. Guiraud = types/√ tokens and "the square root in the denominator leads to a higher G with a longer text with the same TTR as a shorter one" (Daller, van Hout and Treffers-Daller 2003, p.200).

Vermeer（2000）discusses the reliability and validity of 10 measures of lexical richness and examines their behaviours in spontaneous speech data. The measures of lexical richness discussed are *tokens, types, lemmas* (number of different dictionary entries), *hapaxes* (number of types occurring only once), *TTR, corrected TTR，Guiraud, log TTR, Uber* index and theoretical vocabulary.

In Vermeer's design of the study, the subjects were 70 Dutch natives and 76 ethnic minorities aged 4 to 7 years old. The minorities have Dutch as their second language. Both indirect (receptive vocabulary test and definition test) and spontaneous speech data (individual interview, story-telling and interview on topics) were collected. The results of the study indicated that only the Guiraud index gave a better indication of lexical richness, but in later stages of vocabulary acquisition (from 3,000 words on) it is not valid any more. Since all the traditional measures of lexical richness investigated in the study were unsatisfactory in spontaneous speech data, the author suggested that a more valid measure of lexical richness might be related to the difficulty of words, measured by their frequency in corpora rather than counting types and tokens in the data.

Malvern and Richards (1997) argued that even *Guiraud* is no better than TTR. They proposed a parameter *D* which was claimed to be an indicator of lexical richness, the bigger the *D*, the more diverse the text. A computer program *vocd* was designed to process transcribed data. According to Malvern and Richards (2009, p.166), *D* "is based on mathematically modelling how the TTR of any given language sample falls with increasing tokens". *D* was claimed to have many advantages and the most important one is that it is independent of sample size.

Malvern and Richards (2002) tried to prove the validity of the *D* as a measure of lexical diversity. Their research data were transcribed audio-tapes of 34 British students (12 in Class A and 22 in Class B) taking their oral exam in French for the General Certificate of Secondary Education (GCSE), which was conducted by their own teachers. Objective measures from CLAN software, results of GCSE, ratings of the GCSE exam given by 24 experienced teachers and *D* values of both teacher and student were obtained for analyses. The study results showed that *D* is correlated with another measure of lexical diversity, MSTTR, rather than with measures of general language proficiency. *D* is correlated with the number of different words rather than with the total number of words. *D* has no correlation with TTR. The results indicate that *D* is an effective measure of lexical richness and *vocd* is an effective tool for analyzing language data. By investigating the teacher variables of lexical diversity, it was found that each teacher was not finely tuned to the ability of individual students. Instead, they were adapting to the general level of the class they were teaching. The researchers proposed that lack of appropriate accommodation might be a threat to the validity of the public oral examination.

Yu (2010) used *D* as a measure of lexical richness on both spoken and written data of the same subjects. The main purpose was to investigate the relationship between lexical diversity and the holistic quality of written and spoken discourse. It was found from his research that *D* was an effective measure of lexical richness and it was correlated significantly and positively with the overall writing and speaking performance of the candidates as well as general language proficiency. It was also found that *D* was a better indicator of speaking than writing performance.

Jarvis (2002) compared the accuracy of five formulae in terms of their ability to model the type-token curves of written texts produced by adolescent learners in Finland and Sweden and by native English speakers in the United States. The data include written narrative descriptions of an eight-minute segment of a silent film and a self-report vocabulary test of 84 nouns and 74 verbs.

The results of the study seemed to indicate that the curve-fitting formulae of *D* provided accurate models of the type-token curves of short texts, and the results of the study also indicated a clear relationship between lexical diversity and amount of instruction and vocabulary knowledge. However, the author also pointed out that there are problems with the parameter *D*. First, he thought Malvern and Richards (1997, 2002) were wrong about the need for a random sample. "Random sampling treats texts as if they were composed of a vocabulary substance that has identifiable particles but no structure" (p. 62), so the curve might be different from the real curve. Another problem with *D* is that in the research conducted by Jarvis (2002), all but one of the texts are short texts, so the accuracy of *D* on longer texts needs further research.

McCarthy and Jarvis (2007) provided both theoretical and empirical evidence to prove that there are problems with the rationale for *vocd* as regards random sampling and curve-fitting. Based on data drawn from different corpora, McCarthy and Jarvis (2007) pointed out that although *D* seemed to be a reliable and valid indicator of lexical diversity in many earlier researches, its reliability was still in question because they found *D* is also significantly affected by text length when sample sizes are over certain ranges.

McCarthy and Javis (2010) also examined the validity of a new index measure of lexical diversity known as *textual lexical diversity* (*MTLD*), which "is calculated as the mean length of word strings that maintain a criterion level of lexical variation" (p.381). To validate *MTLD*, lexical diversity measures of *vocd-D, TTR, Maas, Yule's K*, and an alternative to *vocd-D* known as *HD-D* were compared. The results indicated that *MTLD* is a valid measure of lexical diversity and is the only measure of lexical diversity that was not found to vary as a function of text length. *HD-D* is an alternative of *vocd-D*, and finally the three index of lexical richness seemed to capture unique lexical information. So the authors suggested that a different measure of lexical diversity be used to get more lexical information under investigation.

In fact Malvern et al. (2004) also acknowledged that *D* could be affected by text length, but that the effects are not significant for the ranges of text lengths with which they are concerned. In other words, they realized that for longer texts of more than a few hundred words, *D* might also be affected by sample size.

In addition to the research on lexical diversity as an indication of lexical knowledge, many researchers (Laufer and Nation, 1995; Wesche & Paribakht, 1996; Vermeer, 2000; Wen, 1999; Daller, van Hout and Treffers-Daller, 2003) have argued that a more effective measure of lexical richness may involve lexical sophistication or frequency of words.

Laufer and Nation (1995) studied the reliability and validity of the Lexical Frequency Profile (LFP) as a measure of lexical richness in written compositions. LFP developed by Laufer and Nation (Laufer 1995; Laufer and Nation 1995) is "the percentage of words a learner uses at different vocabulary frequency levels in his or her writing" or "the relative proportion of words from different frequency levels" (p.310).

The subjects of Laufer and Nation's research were 65 EFL learners of three different proficiency levels in New Zealand and Israel. The subjects were asked to write two compositions in class in one week. The learners were also given the active version of the Vocabulary Levels Test (Nation, 1983) which elicited the use of target words of 5 frequency levels (the second 1,000 words, the third 1,000 words, the fifth 1,000, the University Word List, and the tenth 1,000) in given sentences.

The results of the study indicated the following. 1) The three proficiency groups were significantly different from each other in the percentage of words they used that belonged to the first 1,000. 2) As for the second 1,000 words, the three groups were not significantly different from each other, although there was a tendency for the less proficient group to use more of the second 1,000 words. 3) As for the University Word List (UWL), the three groups were different from each other for composition. 4) The three groups were also different from each other in the use of words that were not in any of the lists. These differences constituted significant evidence for the validity of the LFP. Therefore, a positive answer could be given to the first research question of the study: will there be a significant difference between the LFPs of learners of different language proficiency levels?

The researchers also compared the results of LFP and the Levels Test. They found that learners who received higher scores on the Levels Test used more sophisticated vocabulary (i.e. words from the UWL and words that were 'not-in-the-lists'). They also found a negative correlation between the Levels Tests and the first 1,000 words. The research proved that the LFP of the compositions correlate highly with the scores of the same learners on the active version of the Vocabulary Levels Test. The results showed that there was no difference between the two essays and the LFP was stable between compositions written by the same student.

Laufer and Nation (1995) conclude that the Lexical Frequency Profile has been shown to be a reliable and valid measure of lexical use in writing. The LFP has the advantage that it provides a more detailed picture of the different words of different frequencies. Thus it is a useful diagnostic tool as well as a sensitive research tool. It can show the extent to which learners are making the fullest use of their available vocabulary knowledge in writing.

Meara (2005) reported a set of Monte Carlo simulations designed to evaluate LFP. According to Meara, Monte Carlo Analysis is an approach which "relies on randomly generated data sets to model a complex process" (p.35), and it is widely used in psychology and engineering. He used simulated data to investigate whether LFP has the features claimed by Laufer and Nation (1995). However, the simulations results suggested that the LFP was not as sensitive as Laufer and Nation claimed. It is suggested that LFP does not distinguish between learners of different levels of proficiency and probably does not have a strong correlation between scores of different writings by the same learner. As a result, the real scores of the LFP profile may not be as consistent as reported in published research.

In response to Meara's (2005) critique of the LFP, Laufer (2005) stated that Meara has misinterpreted their work and stressed that there are differences between simulations and reality. But she also concluded that "we do not have perfect measures of vocabulary knowledge and use. Therefore revisiting and refining the existing tools is a legitimate and useful scholarly activity" (p.587).

Daller, van Hout and Treffers-Daller (2003) compared different measures of lexical richness used in the spontaneous speech of two groups of Turkish-German bilinguals. One group of subjects are more competent in Turkish while another group is more competent in German. After analysis of the indexes, it was concluded that the two measures of lexical richness Advanced TTR and Guiraud Advanced proposed by the researchers had advances over traditional measures, especially Guiraud Advanced, which adds the information about degrees of difficulty of the word to lexical variation. Guiraud Advanced is the ratio of advanced *types* shared by the square root of the total number of *tokens*. The definition of advanced types is normally based on frequency lists.

2.5 The gap in literature and research questions

This chapter has reviewed literature and research focused on input and interaction in SLA, the Accommodation Theory and its application in SLA and research on vocabulary knowledge and assessment and their application in oral proficiency interviews. However, the three areas of input and interaction, Accommodation Theory and vocabulary research seemed quite distinct inquiries in previous studies of applied linguistics. Very few studies have ever combined these three areas. In fact, in GESE, an oral proficiency interview, the interactions between the examiner and the candidate are composed of the output from one speaker, which is on the other hand, also the input for his or her interlocutor. Input from both sides is in constant modifications as the result of interaction. In oral proficiency interviews, examiner accommodation occurs during the interview, and accommodation in vocabulary is a very important aspect. This thesis is an attempt to link these three areas in an oral examination setting.

From the literature review it is found that researchers have different understanding of the vocabulary knowledge and proposed different dimensions of vocabulary knowledge as well as measures of lexical richness. However, up to now, as Laufer (2005) has mentioned, "we do not have perfect measures of vocabulary knowledge and use. Therefore, revisiting and refining the existing tools is a legitimate and useful scholarly activity" (p.587). In the present research, by exploring the relationship between the examiner's and the candidate's lexical variables, as well as their relations with candidate GESE scores and examiner accommodation, it is hoped to shed more light on the construct of vocabulary knowledge and lexical richness, and obtain a better understanding of the features of different lexical measures.

In the literature there is a lot of research on the relationship between writing and vocabulary, but much less is known about vocabulary in spoken texts. Researchers have conducted investigations on the validity of different statistical measures of vocabulary in oral interview settings, but the results are different. Not much is secured about the nature of the lexical richness measures or what Read called "the comprehensive lexical measures" (p.188), especially in oral examinations with participants of different levels of proficiency. Read (2000, p.188) pointed out that those comprehensive measures "are particularly suitable for assessment procedures in which vocabulary is embedded as one component of the measurement of a larger construct, such as communicative competence in speaking", but another issue arises here: what is the relationship of the comprehensive measures of vocabulary with other assessment criteria in oral interviews? There is still much to explore in this area.

Vocabulary acquisition and assessment and Accommodation Theory were quite separate fields in SLA. This present research combines the two fields by exploring the lexical accommodation that GESE examiners adopt towards candidates. Lexical accommodation is investigated by examining whether or not and if so to what extent the examiner's lexical variables correlated with candidates' lexical variables.

Another issue is that there are only a few studies on lexical accommodation in oral examinations. Among the few studies (Richards and Malvern 2000; Malvern and Richards 2002; Lorenzo-Dus and Meara 2005) of accommodation at the lexical level, most of them used data selected by the researcher on the basis of availability, and the number of subjects is comparatively small. No research has been conducted so far with a large scale random sampling data collected by the research. Large scale random sampling has advantages in that its larger size "is more representative of the population as a whole" (Selinger and Shohamy 1997, p.98) and random sampling ensures that "the data collected are truly representative of the natural behaviour of the group" (p.104).

Based on the literature review and the discussion about the gap in literature, five research questions are raised:

Research Question 1

Will the student variables, including lexical richness measures (number of types, number of tokens, D, Giraud (G), AG etc.) and MLU, differentiate candidates of three different grades of GESE?

Research Question 2

Will student variables, including lexical richness measures (number of Types, number of tokens, D, G, AG etc.) and MLU, differentiate between good performers and poor performers at the same grade of GESE?

Research Question 3

Will student lexical richness measures (number of Types, number of tokens, D, G, AG, etc.) and MLU, correlate with student GESE score variables?

Research Question 4

Do examiners accommodate lexically to candidates of different grades? If so, how and to what extent do they accommodate to the candidate in vocabulary?

Research Question 5

Do examiners accommodate lexically to good performers and poor performers of the same grade? If yes, how and to what extent do they accommodate to different performers in vocabulary?

Chapter 3 The Pilot study

3.1 Introduction

The research questions formulated in Chapter 2 need to be investigated and testified through a series of procedures: transcriptions of the recording; tidying up the data and changing the transcriptions into Codes for Human Analysis of Transcripts (*CHAT*) format for the Child Language Data Exchange System (*CHILDES*) (MacWhinney 2000); using *CLAN* commands to calculate teacher variables and student variables, and finally using Statistics Package for Social Sciences (*SPSS)* to investigate the relationship among those variables. Before the present research started, a pilot study was carried out to investigate small scale data to get preliminary results of the characteristics about the GESE data, to experiment with research instruments and research procedures, and to help form refined research questions for the main research.

The pilot study was conducted before the data collection of the main research of the present project. The major aims of the pilot research were to investigate: 1) how Chinese local GESE examiners rate vocabulary and what aspect(s) of lexical richness is (are) the indicator(s) of the examinee's vocabulary knowledge according to these examiners, and 2) the characteristics of the lexical richness measures of both the examiner and the examinee. First, a small scale survey was conducted and then both student and teacher variables were compared based on the data of seven GESE examinations. It was expected that the pilot study might provide insights to a more reasonable design of the main research and test the feasibility of those research methods to be used in it. The pilot study included two parts: 1) a survey and 2) the quantitative analysis of teacher and student variables in seven examinations.

3.2 Research design of the pilot study

3.21 Research questions

1) According to the survey, which aspect(s) of lexical richness is (are) the indicator(s) of the candidate's vocabulary knowledge?

2) According to the quantitative analyses, which lexical measure(s) is (are) more

likely to differentiate between candidates of different grades ?

3) According to the quantitative analyses, which lexical measure(s) is (are) more likely to differentiate between candidates of different scores at the same grade?

3.22 Subjects

In the survey research the subjects were all the Chinese local examiners of GESE who attended the standardization training organized by Trinity, London and Beijing Education Examinations Authority (BEEA) in January 2005 in Beijing, China. All subjects were experienced teachers who had taught English in universities in Beijing for at least eight years by then and had been a GESE examiner for at least three years. There were 11 women and only one man. According to the statistics of the Higher Education Development Centre of BEEA, the average GESE working time of Chinese examiners was around 60 hours in the year 2004.

In the first part of the pilot study, the survey, the researcher just took all the examiners who attended the training as the subjects. In the second part of the pilot, the subjects were one examiner and seven candidates, and the data were collected from seven GESE examinations of different grades conducted by the same examiner. Among the seven candidates, six of them were children and one of them was an adult.

3.23 Instruments

For the first part of the piloting, questionnaires were used to obtain the Chinese local GESE examiner's opinions on the assessment of vocabulary in GESE examinations. The survey is expected to provide important information that cannot be obtained from the quantitative research.

In the quantitative analysis, the transcribed data were transformed into *CHAT* format and *CLAN* of *CHILDES* project (MacWhinney 2000) was used to compute and obtain lexical variables of *Token, Type, TTR* and *D*. The variable of difficult words was decided by the researcher and another experienced GESE examiner and teacher of English. Daller, Van Hout and Traffers-Daller (2003) proposed that there is high reliability of the teachers' ratings on advanced lexical items. They chose to rely on teacher ratings rather than word list.

The reason why I chose to rely on examiner judgment in the pilot study is that a "difficult word" is in fact not very easy to define as there are different grades of GESE. For example, in Grade 3, the candidates are expected to know some basic words pertaining to their places of study and school subjects such as *library*, *mathematics* and *music*, but these words are counted as difficult words for a Grade 1 candidate who is only expected to know very limited words such as the parts of the body and a little personal information. Similarly, some basic vocabulary for Grade 7 topics would be difficult words for candidates of lower grades such as Grade 3 and Grade 5. In addition, candidates of the lower stage such as the Initial Stage (Grade 1-3) have very limited vocabulary, so it would be inappropriate to use the most frequently used 1,000 words to distinguish between basic and difficult words. There is no word list for each grade of GESE, so experienced examiners tend to give a more reliable judgment in this case. SPSS (17.0) is used to compute differences and correlations between variables and help to testify hypotheses.

3.24 Data collection and procedures

Data collection of the pilot study included two parts: collection of questionnaires and collection of recorded data of the GESE examination. First of all, a questionnaire (see Appendix 1) was designed to investigate examiners' viewpoints on the indicators of examinees' vocabulary knowledge in assessment of GESE. The contents of the questionnaire include examiner opinions of the relationship between vocabulary knowledge and four aspects of lexical richness (Read, 2000), understanding of the assessment criteria of the examination and some general question of the opinions on vocabulary assessment. After the questionnaire was formulated, I filled in the questionnaire myself and revised it to make it easier to complete. The questionnaire was written for subjects who are very fluent in English, so questions were written in English that would avoid possible confusion caused by the translation of some technical terms.

There are 10 items altogether in the questionnaire. The first nine items are multiple choices and the options are arranged on a five-point scale. For each item, the subject is asked to choose one option from 'strongly agree' (5) to 'strongly disagree' (1). The last item is an open question concerning indicators of effective vocabulary usage. The questionnaires were distributed on the second day of the 3-day training and were collected on the same day or the next day of the distribution by the researcher. All questionnaires (12) were collected and all of them were answered and could be used for analysis.

Questionnaires were administered face to face by the researcher, which was very convenient and ensured that all the subjects understood the items. Brief explanation was given after the subjects got the questionnaire and questions were answered and opinions exchanged if the subjects had any questions. The results were collected and analyzed by the researcher. Second, seven full length GESE examinations were collected from the Higher Education Development Centre of BEEA. The seven examinations cover four different grades: Grades 1 and 3 of the Initial Stage, Grade 5 of the Elementary Stage and Grade 9 from the Intermediate Stage. There are two examinations with candidates of different scores from Grades 1, 3 and 5 respectively and there is one examination from Grade 9.

Table 3.1 The grades and stages of the collected GESE data

Stage	Initial Stage (Grade 1-3)	Elementary Stage (Grade 4-6)	Intermediate Stage (Grade 7-9)
Phase(s) of each Stage	Conversation	Prepared topic Conversation	Prepared topic Interactive task Conversation
Time duration	5-7 minutes	10 minutes	15 minutes
Reference to the CEF*	The first common reference level (Basic User)	Between the first common reference level to the second common reference level (Basic User to Independent User)	Second common reference level (Independent User)

CEF* The Common European Framework for Language Learning, Teaching, Assessment (2001)

(Source: Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages from 2002)

Grades 1 and 3 are in the Initial Stage, which relates to the first common reference level (Basic User) of The Common European Framework for Language Learning, Teaching, Assessment (2001) (CEF); Grade 5 is in the Elementary Stage which is between the first common reference level to the second common reference level (Basic User to Independent User) of CEF, and Grade 9 is in the Intermediate Stage which relates to the second common reference level (Independent User) of CEF. The examination time of Grade 1 is five minutes, Grade 3 is seven minutes, Grade 5 ten minutes and Grade 9 fifteen minutes. **Table 3.1** shows the basic information of GESE Grades and Stages of the collected data.

The main purpose of analyzing examinations was to get some preliminary characteristics of the lexical measures of GESE data and to investigate whether the examiners really assess vocabulary in the way they described in the questionnaire. The seven examinations were conducted by a very experienced GESE examiner and they covered four grades in three stages. The reason why the examinations were conducted by the same examiner is to get the characteristics of the examiners and examinees' lexical variables clearly without being influenced by the examiner's personal style. Brown (2003) found from her research that examiners' personal styles can affect the behaviour of the candidates and their scores to a large extent.

The seven full-length examinations were transcribed into *CHAT* format of the *CHILDES* project (MacWhinney 2000) by the researcher. The researcher did not make much change of the transcriptions, just deleted the unintelligible words and redundancy such *as um, ha* and *ah* etc. *CLAN* of the *CHILDES* project (MacWhinney 2000) was then used to get both examiner and student variables. The Examiner (teacher) variables include examiner *D, Tokens, Types* and *TTR*. For examinee (student) variables, in addition to the *D, Tokens, Types* and *TTR*, the student's GESE scores were also obtained. The number of difficult words in each grade was decided by the researcher and another experienced GESE examiner according to their professional judgment. Finally, the results of the questionnaires, the student and teacher variables were analyzed.

3.3 Analyses and results

The results of the questionnaire survey and the quantitative measures are analyzed and presented in this part.

3.31 Analyses and results of the questionnaires

First of all, the questionnaires about the Chinese GESE examiners' viewpoints on how to assess vocabulary are presented and analyzed. The results are summarized in Table 3.2.

It is found that the Chinese local examiners of GESE gave very strong positive responses on Items 1, 2 and 6.

Item 1 is intended to investigate the examiners' general opinions on vocabulary as an essential part of L2 proficiency. From the results we can see that 91.7% of the examiners held a very firm belief (58.3% strongly agree and 33.3% agree) that the examinee's vocabulary knowledge is closely related to their overall L2 proficiency.

Table 3.2 Responses to item 1 to item 9 of the questionnaires

	5-point Scale		**Frequencies and percentage of responses of each item**				
Item	Mean	Std. Deviation	Strongly agree 5	Agree 4	3	Disagree 2	Strongly disagree 1
1	**4.50**	**.67**	**7 (58.3%)**	**4 (33.3%)**	**1 (8.3%)**	**0**	**0**
2	**4.42**	**.67**	**6 (50%)**	**5 (41.7%)**	**1 (8.3%)**	**0**	**0**
3	**2.25**	**1.06**	**0**	**2 (16.7%)**	**2 (16.7%)**	**5 (41.7%)**	**3 (25%)**
4	**2.17**	**.94**	**0**	**1 (8.3%)**	**3 (25%)**	**5 (41.7%)**	**3 (25%)**
5	**3.42**	**.90**	**2 (16.7%)**	**2 (16.7%)**	**7 (58.3%)**	**1 (8.3%)**	**0**
6	**4.33**	**.89**	**6 (50%)**	**5 (41.7%)**	**0**	**1 (8.3%)**	**0**
7	**2.75**	**.75**	**0**	**1 (8.3%)**	**8 (66.7%)**	**2 (16.7%)**	**1 (8.3%)**
8	**3.33**	**1.61**	**3 (25%)**	**5 (41.7%)**	**0**	**1 (8.3%)**	**3 (25%)**
9	**2.08**	**1.17**	**0**	**1 (8.3%)**	**5 (41.7%)**	**0**	**6 (50%)**

Examiners also gave a very positive answer to Item 2: "I mark vocabulary according to specific rules derived from assessment categories." 91.7% of the examiners believe (50% strongly agree and 41.7% agree) that they derive some rules or indicators from the assessment criteria to mark vocabulary. This is not difficult to understand. How to assess vocabulary is not described very clearly in the syllabus. From the Summary of assessment criteria of the GESE Syllabus (2002) used in the year 2005 by the Chinese examiners it is found that the criteria of vocabulary assessment which is included in the criterion of *Usage* is not described very clearly. For example in Grades 2 and 9, the candidates are expected to fulfill the requirements in the criterion of *Usage*:

Grade 2

Describe people and things

Use learnt phrases as necessary

Identify positions

Understand and respond to simple questions about activities

Days of the week and months

Numbers up to 50

Grade 9

Elicit information and opinion

Talk about prior experience

Express abstract ideas

Express regrets and wishes

Express hopes

(Source: Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages from 2002: pp.40-42).

It is found that some functions are listed for each grade, but the quantity and the quality of the vocabulary of the candidate of each grade is not specified. Since the assessment criteria for vocabulary is rather general in nature, each examiner might use his or her own rules or indicators of vocabulary instead of the criteria described in the syllabus. Examiners also hold a very positive opinion on Item 6 and 91.7% (50% strongly agree and 41.7% agree) of them are in strong agreement or in agreement with the statement, which indicates that the overwhelming majority believe that difficult words (lexical sophistication) is a very strong indicator of lexical richness.

In this questionnaire, Items 5 to 8 are focused on the role of different aspects of lexical richness in the rating of vocabulary knowledge. In these four items, Item 6 gets the highest score of 4.33 (Std. Deviation =.89) and the result shows that lexical sophistication (difficult words) is considered as a more important indicator of vocabulary knowledge than other aspects by those examiners who answered the questionnaire. They also have a rather positive opinion on Items 5 and 8 with the mean score of 3.42 (Std. Deviation=.90) and 3.33 (Std. Deviation=1.61) respectively. For Item 5: "I tend to give a high mark of vocabulary if the examinee uses synonyms or rephrasing to avoid repetition", 33.4% of the examiners agree with the statement, but the majority (58.3%) think it is hard to come to a conclusion about lexical richness just based on lexical variation. The responses to Item 8, "the more grammatical errors the examinee makes, the lower the mark of vocabulary" are very diverse, with 66.7% in strong agreement or in agreement and 33.3% in disagreement or in strong disagreement. Nobody chose 3 (hard to say). From responses to Item 7, "I tend to give a high mark of vocabulary if the examinee uses very complicated sentence structures", we can see that examiners hold a slightly negative belief in sentence structure (in a sense lexical density) as an indicator of vocabulary knowledge. 25% of the examiners disagree with the opinion that pronunciation is the indicator of lexical richness and more than 67% of the examiners chose "hard to say".

The Chinese local examiners of GESE give the most negative opinions on Items 4 and 9. These two items are designed to see whether fluency and pronunciation have significant influence on the rating of vocabulary. The results of questionnaire seem to indicate that neither the examinee's general level of fluency nor pronunciation has much influence on the rating of vocabulary. In later communication with them, many examiners expressed their ideas that it is often the case that an examinee who has a good proficiency is fluent and good at pronunciation, but as there is an assessment criterion of pronunciation that covers fluency, assessment of vocabulary is relatively independent from the fluency and pronunciation of a speaker.

It is also worth noticing that Chinese local examiners give a negative response on Item 3, “Experience and professional instinct are more important than assessment categories.” 66.7% disagreed and strongly disagreed. All the subjects were considered senior examiners and most of them (11 out of 12) had been a GESE examiner for more than 5 years by 2005, but they did not agree that experience and professional instinct are more important than assessment criteria. From this we can infer that 1) they put much emphasis on assessment criteria, since they have to give specific marks for each criterion; 2) they do not give a mark based only on their impression; 3) they are not very confident about their subjective judgment. The reason might be that because they are non-native speakers, they are not as confident as native speakers in judging the examinee’s language proficiency.

Item 10 is an open question. Since it is not obligatory, only 5 subjects out of 12 gave an answer and the responses are very scattered, but at least we can see that the factors are paid attention to and they may influence the examiners’ judgment in one way or another. Range of vocabulary and appropriateness are mentioned twice, communicative skills, subjects, register and difficult words are also mentioned once.

From the results of questionnaires we can get some very tentative information on the examiners’ opinions on the assessment of the candidates’ vocabulary knowledge. First of all, the examiners think that in the oral examination, they are not affected by other aspects of language proficiencies such as fluency and the pronunciation of the examinee when assessing vocabulary. But whether or not the real behaviour of the examiners is the same as what they expressed in the survey needs to be investigated.

Second, lexical sophistication seems to be regarded as the most important indicator of lexical richness by GESE examiners. This is also supported by the results of the quantitative measures obtained from teacher ratings. In Grade 5, the candidate who used more difficult words got a higher GESE score. Lexical diversity comes after lexical sophistication and is regarded as the second important indicator of the examinee's lexical knowledge. However, lexical diversity is difficult to assess when the examiner rater is engaged in both interviewing and rating.

There are many limitations of this survey. First of all, the design of the questionnaire is far from perfect, and some items in the questionnaire may over-simplify the real situation. Secondly the sample is very small. The results and the conclusions made in this research may not be applied to the whole population or be appropriate in different situations. Finally, it was found that Item 8, "the more grammatical errors the examinee makes, the lower the mark of vocabulary", is not very appropriate. Responses to the item are polarized. 67.5% agreed and strongly agreed while 32.5% disagreed and strongly disagreed. In Item 8, the meaning of "grammatical errors" is not defined very clearly, and it also shows that examiners may have different opinions on the relationship between vocabulary and grammar. These problems may affect the examiners' assessment of vocabulary.

3.32 Results of the quantitative measures

In the second part of the Results student variable and examiner variables were compared and analyzed.

Although only 7 examinations were analyzed, and the sample was small, we can still get some preliminary results:

First of all, both teacher/examiner and student/candidate lexical variables of *Type, Token* and *D* rise as grade goes up. As the examination goes on longer, more *Tokens* and *Types* enter the discourse and lexical diversity also goes up as the proficiency level rises. This is the same as expected. Examiner *TTRs* are rather stable, but Student *TTRs* goes in the opposite direction and there is a trend that the higher the Grade, the lower the *TTR*. As many researchers have claimed, TTR is problematic and it is greatly influenced by sample size. The longer the sample size, the lower the *TTR*.

Second, within a certain grade, particularly in Grade 1 and Grade 5, it seems that *Type* (number of different words) is a better indicator of vocabulary than any other variables. It can differentiate different Grades and also between scores in Grades 1 and 5. In Grades 1 and 5, the higher the number of *Types*, the higher the candidate's mark.

Table 3.3 The Examiner lexical variables (E) and Student variables (S)

ID. Grade	D		Type		Token		TTR		Number of Difficult words	score
	E	S	E	S	E	S	E	S		
1 (G1)	32.1	-	80	30	226	38	0.35	0.79	0	B
2(G1)	46.2	49.5	96	52	230	82	0.42	0.63	0	A
3(G3)	30.6	-	43	27	78	41	0.56	0.66	0	A
4(G3)	43.8	46.1	102	73	263	146	0.39	0.50	3	C
5(G5)	67.7	52.9	162	165	415	553	0.39	0.30	4	A
6(G5)	65.4	55. 9	193	153	561	337	0.34	0.41	3	B
7(G9)	90.41	88.3	200	281	455	885	0.44	0.32	0	B

There is an exception in Grade 3: the student who has a higher *D* and *Type* got a lower mark for vocabulary in GESE. Further analyses of the examination indicate that the main reason why No. 3 candidate has a lower *D* but got a higher score is that his answer is well-focused and more relevant to the question. In addition, he also has very good pronunciation. However, although Candidate No. 4 got a higher *D*, he did not understand some of the questions and we can assume that he had not mastered the basic vocabulary to fulfill the tasks of the Grade. Third, all the teacher lexical variables but *TTR* vary in the same trend as the student lexical variables across grades, and we may interpret this as that the examiner accommodates lexically to the candidates of different grades. Since there are only at most two candidates in each grade, we are not sure whether she accommodated to perhaps good performers and poor performers within a group on the lexical level. However, these very preliminary results need to be proved by large scale data in the main research.

3.4 Discussion

After obtaining and analyzing the results from both the questionnaires and quantitative analysis of both teacher and student lexical variables from the GESE examinations, it is not difficult to find that lexical diversity measure of *D* and the number of *Types* (number of different words) are strong indicators of the students' scores of vocabulary, especially across different grades. The number of difficult words is not a very effective indicator of an examinee's vocabulary knowledge in the pilot study although GESE examiners rank it as the most important indicator of vocabulary in the questionnaires. The main reason might be that there were only two samples in each grade, and the small sample is unable to show whether or not it is an effective indicator of lexical use of the candidates. The function of the difficult word needs to be further investigated in the main research with a large data.

In the analysis of the variables, difficult words are not easy to define or to count, and a lexical measure that can show vocabulary difficulty is greatly needed. Not long after the piloting was finished, a software Guiraud Advanced (*AG*) was developed by Daller (2006). The measure AG= Advanced *Types/√tokens*, which is the ratio of the advanced *Types* (the number of different words in a text) and the square root of the number of all the *tokens* (the number of all the words in a text). Advanced *Types* mean the *Types* beyond the most frequently used 1,000 words according to NBC Corpus. In the main research, the researcher decided to use *AG* instead of teacher rating as the index of lexical difficulty.

Table 3.3 shows that there are two students whose *D* could not be calculated by *VOCD* software because there are less than 35 Types in their speech, and 35 is the minimum for *D* calculation. After the pilot study, the research experimented on more data and found that many Grade 1 candidates have less than 35 *Types* in the whole examination. For Grade 2 and Grade 3 there are very few candidates whose Number of *Type* is less than 35. So in future data collection for the main research, the researcher will not chose Grade 1 as research data for analysis but Grade 2 as the lowest grade.

By doing the pilot study, the researcher experimented with the variables to be used in the main research, experimented with different research instruments, found out what were the reasonable procedures to help formulate the research questions of the main research.

Chapter 4 Research Methodology of the Main Study

The results of the pilot study in Chapter 3 indicate that there is some relationship between the candidate's grades and their lexical richness measures; and there are also relationships between the teacher/examiner's lexical richness measures and those of the candidate/student/examinee. It is also worthwhile to study the relationship of the student/examinee's score and the teacher/examiner lexical richness measures. In the light of the pilot study, elaborated research questions of the main study were established based on the research questions proposed in Chapter 2 of the literature reviews.

This chapter provides the overview of the research methods of the main study of the present research. The participants, instruments, data collection and research procedures are described in different sections.

4.1 Refined Research Questions

The research questions proposed in Chapter 2 are refined in this section. The answers to Questions 1 to 3 are presented in **Section 5.11** Lexical measures of the students/candidates of three different grades of **Chapter 5**.

Research Question 1

Will the student variables, including lexical richness measures (number of *Types*, number of *Tokens*, *D*, *Giraud*, *AG* etc.) and *MLU* differentiate candidates of different GESE grades?

1.1) Will the lexical variables rise as the grade goes up?

1.2) Are there any statistically significant differences between the lexical variables of different GESE grades?

1.3) What lexical variables can differentiate between different grades?

Research Question 2

Will student variables including lexical richness measures (number of *Type*, number of *Tokens*, *D, Giraud, AG* etc.) and MLU differentiate between good performers and poor performers at the same grade of GESE?

2.1) Will the qualified performers have higher lexical variables than the poor performers at the same grade?

2.2) Are there any statistically significant differences of the lexical variables between the qualified performers and poor performers of the same grade?

2.3) What lexical variables can differentiate between qualified performers and poor performers of the same grade?

Research Question 3

Will student lexical richness measures (number of *Types*, number of *Tokens, D*, *Giraud, AG* etc.) and *MLU* correlated with student variables obtained from GESE scores?

3.1) Will all the student lexical variables correlate with each other?

3.2) Will all the GESE score variables correlate with each other?

3.3) Will all the student lexical variables correlate with the GESE score variables?

The answers to Questions 4 to 5 are presented in **Section 5.12** Different measures of good performers and poor performers at the same grade of **Chapter 5**.

Research Question 4

Will examiners accommodate lexically to candidates of different GESE grades? If so, how and to what extent do they accommodate to a candidate in vocabulary?

4.1) Will examiners use more varied and sophisticated vocabulary with candidates of higher grades?

4.2) Is there any difference among the examiner lexical variables used to candidates of different grades?

4.3) Is there a significantly positive correlation between examiner lexical variables and candidate lexical variables?

Research Question 5

Will examiners accommodate lexically to good performers and poor performers at the same GESE grade? If yes, how and to what extent do they accommodate to different performers in vocabulary?

5.1) Will examiners use more diverse and sophisticated vocabulary with candidates of qualified performers than with candidates of poor performers at the same grade?

5.2) Is there any difference among the examiner lexical variables for qualified performers and poor performers at the same grade?

5.3) Is there any significantly positive correlation between examiner lexical variables and candidate lexical variables at the same grade?

5.4) Is there any significantly positive correlation between examiner lexical variables and the candidate's score?

4.2 Subjects

The subjects of the main research consisted of two groups of people. The first group are 180 examinees or candidates who took Grades 2, 5 and 7 of GESE examinations in Beijing, China in the year 2008. They were randomly chosen from the corpus of Beijing Education Examinations Authority (BEEA). The other group of subjects are the 23 examiners who conducted the examinations chosen for the present research.

4.21 The candidates

The first group of subjects of this research are candidates of Grades 2, 5 and 7 of GESE examinations. They are mainly primary school students and junior-middle school students studying in public schools in Beijing, but there are also a small number of adults in Grades 5 and 7. Most of the candidates have followed a three-month training course in a training school aiming at GESE examinations in addition to the English courses in their own schools. The information about the candidates' age, gender and pass rate is shown in Table 4.1.

Table 4.1 shows that the average age of Grade 2, Grade 5 and Grade 7 examinees is 9.1,11.9 and 15.8 years respectively. The majority of the examinees are children and adolescents. The difference in the pass rates among the three grades is also obvious. It can be seen very clearly that with the rise in grade, the pass rate decreases sharply.

Table 4.1 Basic information of the candidates

Grades	Age (year)			Gender		Pass rate
	maximum	minimum	mean	male	female	
Grade2 (n=59)	14.5	6	9.1	34	25	83%
Grade 5 (n=60)	30.6	10.1	11.9	31	29	55%
Grade7 (n=60)	45.9	8.9	15.8	28	32	25%

4.22 The examiners

23 examiners conducted GESE examinations in 2008, and all of them were involved in the data collected for the present study. All of them were experienced college teachers of English working in colleges and universities in Beijing and they had been GESE examiners for at least 6 years. Among them, there are 20 female examiners and 3 male examiners, which represents the reality that there are many more female rather than male English teachers in China.

The examiners do not know the candidates and should avoid conducting examinations with their own students or members of their own families. All the GESE examinations conducted by Chinese examiners were audio-recorded and supervised by examiner panels both in China and Trinity London in the UK. All Chinese examiners receive standardization training sponsored by Trinity London and Beijing Authorities of Education Examinations (BEEA) twice annually, once in January and another in July. The trainers are native speakers of English and senior examiners of GESE from Trinity London. The training includes two parts, language training and standardization of marking. Language training usually includes exercises in pronunciation, intonation, grammar and question forming; for standardization, examiners are required to remark some video-taped examinations and then the marks given by examiners are benchmarked. In addition, some issues that concern the examiners are discussed and feedback from Trinity College, London is also given.

The data used in the research involves 180 examinations conducted by 23 examiners. The examiner (rater) information is shown below:

Table 4.2 Examiner information

Code	Gender	Grade 2 examinations conducted	Grade 5 examinations conducted	Grade 7 examinations conducted	Total
T01	female	1	1	0	2
T02	female	1	1	0	2
T03	female	2	0	0	2
T04	female	1	2	1	4
T05	female	1	3	0	4
T06	female	3	2	0	5
T07	female	4	1	0	5
T08	female	5	0	0	5
T09	female	1	2	2	5
T10	female	4	2	0	6
T11	male	1	3	2	6
T12	female	7	0	0	7
T13	male	2	1	4	7
T14	female	0	2	5	7
T15	female	8	0	0	8
T16	female	4	4	0	8
T17	female	6	4	0	10
T18	female	2	1	8	11
T19	female	2	9	0	11
T20	female	1	5	8	14
T21	female	1	2	11	14
T22	male	0	11	5	16
T23	female	3	4	14	21
total		60	60	60	180

4.3 Instruments

Research instruments of the main research are presented and discussed in this part of the chapter.

4.31 The Corpus of 2008 GESE Examinations

The data used in this study were collected from the database of the 2008 GESE corpus of BEEA. The corpus contains all the GESE examinations conducted in the year 2008 under the guideline of Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages of 2007-2010. However, there are some differences between the UK examiners and the Chinese ones: the UK examiners started to apply a holistic assessment system in 2004, but their Chinese counterparts started to apply the holistic assessment system only in 2010. In 2008, all the GESE examinations conducted by Chinese examiners still followed the analytical assessment criteria based on Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages of 2002.

There are two reasons why the corpus of 2008 was chosen: first, 2008 is the second year of using the syllabus 2007-2010. It is expected that the examiners should have become familiar with the syllabus after using it for a year; and another reason is the 2008 corpus is the first one that has entered the new GESE corpora management system of BEEA. In the new management system, all data were uploaded on the computer and it is comparatively easier to run the random sampling program. According to the statistics of BEEA, about 20,000 examinations were conducted by the Chinese local examiners in 2008, and the average working time for each examiner is around 120 hours, or 18 working days.

4.32 The introduction of GESE in Beijing, China

GESE of Trinity College London was introduced to Beijing in 1999. As is shown in Table 4.3, there are 12 grades in four stages altogether, with three grades in each stage: Initial Stage (Grades 1 to 3), Elementary Stage (Grades 4 to 6), Intermediate Stage (Grades 7 to 9) and Advanced Stage (Grades 10 to 12). Chinese local examiners (non-native speakers of English) have been conducting GESE Examinations of the initial stage and elementary Stage (from Grades 1 to 6) since 1999. Five years later, some of the most experienced senior examiners who had been GESE examiners for at least 4 years then started to conduct some examinations of the Intermediate Stage (Grades 7-9) from September, 2004. The Advanced Stage (Grades 10 to 12) has always been conducted by examiners who are native speakers of English from the UK. The GESE examinations were arranged and conducted almost every weekend and occasionally on weekdays in 2008. According to BEEA statistics, about 20, 000 examinations were held in Beijing in 2008 and the majority (nearly 80%) of them are examinations of the initial and elementary stages, and most candidates are children and teenagers.

According to Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages 2007-1010, the GESE examinations have different procedures for each stage (see Table 4.4). In the Initial Stage (Grades 1 to 3), there is only one phase, the conversation phase. In this phase, the examiner gives the candidate opportunities to demonstrate through both speech and actions the range of language indicated for the stage by asking simple questions and asking them to do some actions according to the instructions. In this part the examiner controls the conversation and communication. In the Elementary Stage (Grades 4 to 6), there are two phases, the Topic Phase and the Conversation Phase. For the Topic Phase, the candidate is expected to present one topic he or she prepared beforehand, which is written on the topic form with a title and 4 to 6 points to talk about. The examiner may ask questions on at least 4 points. In the conversation phase, the examiner may choose two subject areas from all the subject areas listed in each grade and ask questions. The candidate in the elementary stage is also required to ask questions both in the topic and conversation phase. In the Intermediate Stage (Grades 7 to 9), in addition to the two sections of topic and conversation, a third Interactive Task Phase is added. Here the examiner provides an oral prompt to which the candidate has to respond by asking questions or making comments. In this phase, the candidate is responsible for keeping the conversation going and maintaining the interaction.

4.33 The Assessment of GESE

In GESE examinations, the examiners are also the raters. They give a score in each of the assessment criteria required for each grade and then convert them into a final score right after the examination. According to Trinity GESE 2004-2007 syllabus, a holistic rating system was applied by the native examiner of English from 2004. The UK examiner assesses the candidate's performance by selecting one of four levels of performance and awards a letter grade A, B, C or D. These levels can be classified as follows: A – Distinction; B – Merit; C – Pass; D – Fail. Examiners indicate areas which are in need of improvement by using the appropriate tick box provided on the report form.

However, the Chinese local GESE examiners still used the old way of analytical rating in 2008. In other words, although they were using the syllabus 2007-2010, the assessment system was still the old one stipulated in the syllabus from 2002. The consideration here might be that the Chinese local examiners were less experienced than their British peers and they also lack the instinct a native speaker has for their mother tongue. The main purpose of the analytical rating system was to help non-native examiners to award scores in a more reliable and valid way before moving to holistic rating in the future. (The Chinese local examiners started to adopt the holistic rating from March 2010).

For the Initial Stage (Grades 1-3), the examiner will rate a candidate from three criteria at each grade: *readiness, pronunciation and usage*. For the Elementary Stage (Grades 4-6) and Intermediate Stage (Grades 7-9), the examiner will rate a candidate from four criteria: *readiness, pronunciation, usage* and *focus*.

Initial Stage (Grades 1-3)

Readiness: the candidate's understanding of the examiner;
satisfying the requirements listed under Candidate Performance for each grade.

Pronunciation: At all grades, production of individual sounds to form words which are intelligible;
Additionally at Grade 2, the use of appropriate contract forms and the beginnings of the use of stress in short answers;

Additionally at Grade 3, extension of the use of stress and initial use of intonation.

Usage: Accuracy of grammatical items used;
use of appropriate vocabulary.

Elementary Stage (Grades 4-6)

Readiness: The candidate's understanding of the examiner;
maintain the flow of conversation through promptness of response, although a short pause will be allowed for candidates to formulate responses at Grade 4 and Grade 5;
satisfying the requirements listed under Candidate Performance for each grade and for all previous grades.

Pronunciation: Production of intelligible individual sounds, including weak forms in connected speech;
satisfactory use of stress, rhythm, intonation and linkage features, including unstressed forms, so that speech sounds natural at the sentence level.

Usage: Accuracy of grammatical items used;
choice of appropriate vocabulary and grammatical items;
range of vocabulary , grammatical items and functions used.

Focus: Communication of sufficient and relevant information required by the tasks set;
coherent organization of information and opinions;
ability to state communicative purpose.

Intermediate Stage (Grades 7-9)

Readiness: Understanding the speech of and points made by the examiner;
maintain the flow of conversation, displaying promptness of response and avoiding too much repetition;
taking the initiative or influencing the direction of the conversation as necessary;
satisfying the requirements listed under Candidate Performance

for each grade;

and for all previous grades.

Pronunciation: Production of combination of individual sounds and the use of stress, rhythm, intonation so as to produce intelligible and natural sounding speech;

competent variation of stress and intonation patterns to express attitudes and specific meanings.

Usage: Accuracy of grammatical items used;

choice of appropriate vocabulary and grammatical items;

range of vocabulary , grammatical items and functions used.

Focus: Communication of sufficient and relevant information required by the tasks set;

coherent organization of information and opinions communicated;

ability to state communicative purpose;

use of strategies, including rephrasing where necessary, in order to maintain the conversation and to emphasize particular points.

(Source: Graded Examinations in Spoken English for Speakers of Other Languages from 2002: pp. 40-44).

It is shown that among the three criteria in the Initial Stage and the four criteria in the Elementary Stage and Intermediate Stage, there is no specific criterion for vocabulary, and vocabulary is mixed with grammatical items in the criterion of *usage*. Since *usage* was the only criterion adopted by the local examiner to assess vocabulary, the score of *usage* in each grade is taken as the score of vocabulary in the present study for research purposes.

For the Initial Stage, there are three assessment criteria: Readiness, Pronunciation and Usage (lexicon and grammar), and the score of each of the three criteria can be classified as A (distinction), B (merit), C (satisfactory) D (almost satisfactory) and E (not satisfactory) Finally, the scores of the three criteria of a candidate are converted to a numeric score according to the conversion tables and classified into A = Pass with Distinction (100-85), B= Pass with Merit (84-75), C= Pass (74- 65) and D=Fail (< 65).

Regarding the Intermediate Stage, 12 scores are given to the 4 criteria in each of the three phases of topic, interactive task and conversation. Finally, the scores of the 12 criteria of a candidate are converted to a numeric score according to the conversion tables and classified into A, B, C and D, which means Pass with distinction (100-85), Pass with Merit (84-75), Pass (74- 65) and Fail (< 65) respectively.

Table 4.4: GESE examination phases and assessment criteria of three different stages

Stages	Phases	Assessment criteria			
Initial	conversation	Readiness	Pronunciation	Usage	
Elementary	topic conversation	Readiness	Pronunciation	Usage	Focus
Intermediate	topic interactive conversation	Readiness	Pronunciation	Usage	Focus

4.34 Lexical richness measures

In the pilot study, five lexical measures were applied: *Token, Type, TTR, D* and Number of difficult words by teacher ratings. *Token, Types* and *D* can to some extent differentiate between candidates of different grades, and they are still used in the main study. The indexes of *Token, Type* and *D* are obtained by using the software of *CLAN* of the *CHILDES* database.

TTR cannot differentiate between candidates of different grades, it was greatly influenced by sample size. As more words are introduced to a text, or as the text becomes longer, the chances that a new word will appear in the text will decrease, and TTR will drop. As a result, TTR cannot be used for texts of different lengths. Since the problem of TTR has been confirmed by many researchers, it is not used as a lexical variable in the main research. Difficult words are not used in the main research either. The main reason is that for different grades, the so-called difficult words are totally different. For example, the word ***smart*** might be a very difficult word for grade 2 candidates, but it is not a difficult word for Grade 5 or Grade 7 at all. The measure of difficult words is replaced by *Guiraud Advanced (AG)* in the main research. In addition to *Token, Type* and *D*, three lexical variables are applied in the main study. They are *Guiraud (G), Guiraud Advanced (AG)* and *MLU(words)*.

Guiraud's index (G) was first proposed by the French scholar Pierre Guiraud in 1954 as a transformation of *TTR*. Traditionally, *TTR* was the most widely used lexical richness index, but *TTR* decreases as the text goes longer and cannot compare texts of different lengths. The formula of *G* is Type/√Token = *G* and it was proposed to compensate for the declining *TTR*. *G* was proved by many researchers (Tidball and Treffers-Daller, 2007; Van Hout and Vermeer, 2007) to be a reliable and valid measure of lexical richness. All the studies mentioned showed that there was a high correlation between *G* and other measures of lexical richness. In the main study of the present research, *G* is added as a measure of lexical richness to replace *TTR*.

Similarly, *Giraud Advanced (AG)* was applied in the main study as a measure of lexical richness which combines lexical diversity and sophistication. *AG* was first proposed by Daller, van Hout and Treffers-Daller (2003) and the formula of *AG* is as follows:

AG= Types advanced/√Tokens. The advanced *Types* are the types beyond the first 1,000 most frequently used words according to NBC corpus. The index of *AG* in the present research were obtained by using the software *Guiraud Advanced* (Daller, 2010).

In addition to the lexical richness indexes, *MLU* is also applied in the main study as a measure that may indicate the gross language development of a learner. The measurement of *MLU* was developed by Brown (1973) and it is mainly used in the field of Child language research. *MLU* computes the mean length of utterance, which is the ratio of morphemes to utterances. Brown (1973) emphasized the value of *MLU* in terms of morphemes, rather than words, because he believed that the acquisition of grammatical morphemes reflected syntactic growth and that *MLU*m or mean length of utterance in morphemes would reflect this growth more accurately than *MLU*w. Brown stated that *MLU* in Morpheme is "an excellent simple index of grammatical development" (p.54). However, Brown also found from research that *MLU* is highly correlated with grammatical complexity until up to an average of *MLU*m, and if over the level, it is no longer considered as an accurate measure. Since Brown proposed *MLU*m it has been widely accepted and used in as an index for gross language development of child language and it has also been used in SLA research.

Many researchers (for example, Arlman-Rupp et al., Hichkey, 1991, Parker and Brorson, 2005) believe that *MLU* by words has advantages over *MLU* by morphemes because words are easier to identify and calculate than morphemes. Richards and Malvern (2000) and Malvern and Richards (2002) used *MLU*w as an index of both student variables and teacher variables in their investigation of accommodation in oral interviews. *MLU* can be obtained from *CLAN of CHILDES* project and the *MLU* is the mean length of utterance by words. The *MLU* used in the present study is *MLU* by words.

4.35 Quantitative analysis

In the main study of the present research, mainly two instruments of *SPSS* are applied: *ANOVA* and Correlation. First of all, one-way between-groups analysis of variance (*ANOVA*) is conducted to explore the differences between variables of different groups. In addition, correlation is computed between the examiner (teacher) and examinee (student) measures to investigate whether accommodation occurs in the oral interview.

4.36 Qualitative analysis

It can be seen from the quantitative analysis that there are some unexpected results about the relationship between the lexical variables and GESE grades.

Some of the results cannot be interpreted by the quantitative data. In order to get a further understanding of the relationship and a better understanding of the GESE examinations conducted by Chinese local examiners, interviews were conducted. After listening to each examination they conducted in 2008, the examiners were asked to do a second marking and then explained how they rated the candidate and why they gave such scores to the candidate. After the second marking, the examiners were shown the original marking and they were expected to give further explanations or comments. The qualitative data helped to get a better understanding on how the examiners rate the global performance of the candidates, and how they rate the vocabulary of the candidates; it also provided further explanations to some unexpected results of the quantitative data, especially their opinions on the reasons why the lexical variables of Grade 7 are surprisingly low compared with those of Grade 5.

4.4 Procedures

4.41 Data collection procedures

Since Chinese GESE local examiners only get involved in the first three stages of the GESE examinations, data of the first three stages were collected and data of the Advanced Stage were not chosen. A simple computer program of random sampling was used to collect data. Finally 60 examinations from Grade 2 of the Initial Stage, 60 examinations from Grade 5 of the Elementary Stage and 60 examinations from Grade 7 of the Intermediate Stage were chosen randomly from the corpus. Altogether there are 180 examination results. The reason for random sampling is that candidates of various backgrounds and scores and all the examiners involved in the examinations were chosen, which gives a more objective and comprehensive picture of the real situation. Previous research of the oral interviews only chose rather small data on the basis of availability, and no research has adopted random sampling. This is also one of the novelties of the present research.

For ethical reasons, all the names of the subjects were omitted and a code was assigned to each subject. The GESE examiners were coded as T1 to T23, and the candidates were coded with digits, the code of the grade and the code of the number. For example, if a candidate is from Grade 5, and the number is 35 out of the total 60 candidates, then the code is 535. The codes for Grade 2 candidates are 201 to 260, the codes for Grade 5 are 501 to 560 and the codes for Grade 7 are 701 to 760.

Table 4.5 Summary of each stage and data derived from three different stages

Stage		**Initial Stage (Grade 1-3)**	**Elemental Stage (Grades 4-6)**	**Intermediate Stage (Grades 7-9)**	**Advanced Stage (Grades 10-12)**
Phase(s) of each Stage		Conversation	1.Prepared topic 2. Conversation	1. Prepared topic 2. Interactive task 3. Conversation	1.Prepared topic 2. Prepared text 3.Listening comprehension 4.General conversation
Time duration		5-7 minutes	10 minutes	15 minutes	20 minutes
Data	***grade***	***Grade 2***	***Grade 5***	***Grade 7***	***None***
	time	***6 minutes***	***5 minutes***	***5 minutes***	
	phase	***Conversation***		***Interactive task***	

(Source: The International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages from 2002).

The examination of GESE Grade 2 lasts 6 minutes; Grade 5 lasts 10 minutes, with 5 minutes in each of the two phases of Prepared Topic and Conversation. The Grade 7 examination lasts 15 minutes, with 5 minutes in each of the three phases of Prepared Topic, Interactive Tasks and Conversation. In Grade 5, the first phase is based on a prepared topic, which may not reflect the real vocabulary knowledge of the candidates because they have prepared the topics when they take the examination. So the first 5 minutes of each examination is taken away and only the conversation part which is more impromptu in nature remains for analysis. For Grade 7, the researcher chose the Interactive task which is the most impromptu part as data for analysis. In summary, the whole examination of Grade 2 which lasts about 6 minutes, the Conversation phase of Grade 5 which lasts about 5 minutes, and the Interactive Task of Grade 7 that also lasts about 5 minutes were analyzed for the purpose of the current investigation. In Grades 5 and 7, the data starts from about 5:00 and ends at about 10:00, and each data starts from the beginning of an utterance and ends with a complete utterance, so there is a 10 to 15 seconds difference in time duration in each data. In addition, the data chosen from Grade 5 may contain a little bit of the Topic phase, and in Grade 7 it may contain a little bit of either one or two other phases.

Grades 2, 5 and 7 belong to three different stages of GESE and there are two grades between Grades 2 and 5 and there is one grade between Grades 5 and 7. The data collected from grades 2 and 5 are both from the Conversation phase, and the data of Grade 7 was chosen from the Interactive task phase. The reason why the conversation part in Grade 5 and Interactive task in Grade 7 were chosen was that these two phases are comparatively more impromptu than the prepared topic phase. In addition, the candidates /examinees are expected to take more control as the stage goes up and they are asked to take more initiatives as the grade rises. It is also assumed that these three grades of examinations are significantly different from each other. Grade 2 is in the Initial Stage, which relates to the first common reference level (Basic User) of The Common European Framework for Language Learning, Teaching, Assessment (2001) (CEF); Grade 5 is in the Elementary Stage which is between the first common reference level to the second common reference level (Basic User to Independent User) of CEF, and Grade 7 is in the Intermediate Stage which relates to the second common reference level (Independent User) of CEF. There are obviously very different conversation topics and requirements for candidates of each Grade. For example, topics for Grade 2 are daily topics for children, such as rooms of the house, family and friends, days of the week and months of the year, etc., and the conversation is mainly in the form of simple questions and answers. Topics in Grade 5 are festivals, means of transport and music, etc., which are more abstract and need explanation or clarification, and the candidate is expected to take more initiatives during the conversations. While in Grade 7, more formal and abstract topics such as education, national customs and products and recycling are discussed and opinions are expressed and exchanged, and the candidates are expected to take the responsibility to keep the conversation going. In Chapter 5 detailed differences among the three stages are discussed.

Data collection started after careful planning. First of all, a random-sampling computer program was used to draw data randomly from the corpus. 60 examinations were collected randomly from Grade 2, Grade 5 and Grade 7 respectively. There are 180 examinations together. Later after the transcription it was found that one examination in Grade 2 was unintelligible, so I just deleted it and there are 59 examinations in Grade 2 as a result, so altogether 179 examinations are used for analyses.

4.42 Data transcription

Those audio-recorded oral examination samples were recorded on a disc and transcribed into Microsoft Word files by my assistant, Chen Hui and Wang Xiaoqing, who were MA students of English. The word files were checked and corrected by the researcher who is an experienced teacher of English and GESE examiner in Beijing, China. After that, all the word files were changed into Codes for Human Analysis of Transcripts (*CHAT*) format for the Child Language Data Exchange System (*CHILDES*) (MacWhinney 2000). Unintelligible words were deleted from the data and ambiguities were solved with the help of other GESE examiners.

The transcriptions were cleaned up by the researcher: the program of *CLAN* was used to tidy up the data and analyze the transcribed data. All Word files were changed into *CHAT* format by using the *textin* command. Since the data is oral data, all the words were spelled into the correct form as long as it is intelligible but the grammatical mistakes such as “I go to school yesterday” were kept instead of changing it into the correct form, “ I went to school yesterday” so there was an endeavour to keep the original text of the examination. Fillers such as *umm*, *aha* were excluded by putting a ***&*** sign in front of those words which meant they would not be counted when *CLAN* program was used. The ***&*** sign was also used before any Chinese names or Chinese utterances. Some candidates, especially candidates of the Initial and Elementary stages, sometimes tended to use Chinese when they could not find a suitable word in English or when they did not understand what the examiner was talking about. Chinese words should not be considered as their vocabulary of English. Different spelling forms of the same word were standardized. For example *yes yeah, ye* were transcribed into *yes*. Numbers were transcribed as words. For example, the time *3:10* was transcribed as *three ten* and the number *120* transcribed as *one hundred and twenty***.** Next, the ***check*** command was used to check any errors that existed in the format. After that when there was no error in the transcription, the ***freq*** command was used to check the spelling errors in the transcription. Since this command will list the frequencies of each different word, we can easily find the error if a word is misspelled. The mistakes in the transcriptions were corrected for the second time.

Then a ***mor*** and a ***post*** command ran on all the transcripts before calculating *Types, Tokens* and *D*s. After running the ***mor*** command, the ***%mor*** line (morphological analysis with parts of speech) coding *tier*: gives the part of speech for each word, along with the morphological analysis of affixes, such as the past tense mark (-PAST) on the verb.

It should be noticed that there are may be ambiguities in the category of a word or the marks of affixes. For example, is *to* a preposition or an adverb? Is the suffix *ed* a mark of the past tense and participle? The ***mor*** command can also solve the problems of words that have the same spelling but a different meaning. For example the word *May* in the sentence *I was born in May* and the *may* in the sentence *May I ask you a question* would not be regarded as one word. The use of a ***post*** command can resolve such ambiguities. But if the question mark ***?*** appears in the lines after ***post*** is run, it is very probably a mistake or a word not in the English lexicon of *CLAN*. It will be corrected if there is a mistake.

Sometimes a question mark will appear when a word does not exist in the English lexicon of *CLAN* but it is accepted in daily communication, the researcher will decide whether to add it to the lexicon or not. For example, *T-shirt* does not exist in the English lexicon, but it is a very frequently used word in the transcription. However, in the lexicon of *CLAN*, the written form of *T-shirt* is *tee shirt*. In this case, the researcher can either replace all the words of *T-shirt* with *tee shirt* or add *T-shirt* to the lexicon. In the present research, many Grade 2 candidates talked about clothes, so *T-shirt* appears in the transcriptions many times, so it was added to the English lexicon of *CLAN*. Finally the ***freq*** command is used to check the word list again for any remaining inconsistencies. All these procedures help reduce the chances of error to the minimum level.

4.43 Analyses of the quantitative variables

Both examiner variables and candidate variables of *Type, Token, D, MLU* were obtained by using the program of *CLAN*, and *AG* is calculated by using the software of Guiraud Advanced developed by Daller (2010). *G* is calculated by using *SPSS*.

As mentioned earlier, the Examiner (teacher) variables include examiner *Type*, examiner *Token*, examiner *D*, mean length of utterance (*MLU*) and examiner *Guiraud* and examiner *AG*. For candidate (examinee) variables, in addition to the variables from the *CLAN* software which are the same as the examiner variables, *D, Type, Tokens, MLU , G,* and *AG*, the student's scores of Grade 2, the scores of Grade 5 in the phase of Conversation and scores of Grade 7 in and the Interactive phase were also obtained. The score of vocabulary (***Usage***) and other assessment categories of ***Readiness, Pronunciation* and *Focus*** as well as the final score of GESE, ***Overall mark*** were also obtained. In addition, another variable ***Overall mark 2***, the sum of the scores of all the assessment categories of the studied phase is computed by SPSS. All the data of the variables were input into a SPSS table for analysis.

For the first three questions that explore the candidates' measures of lexical richness of different grades, the descriptive statistics of the three groups are compared and One-way *ANOVA* is carried out to investigate whether there are significant differences among the three groups of candidates and find out what measures of lexical richness can differentiate students of different stages and also what measures of lexical richness can differentiate good performers from poor performers at the same stage .

The last two research questions are focused on the examiners' lexical accommodation with candidates of different performance. Spearman correlation is carried out to answer the two research questions: will examiners use different vocabulary to candidates of different grades and will examiners use more diverse and sophisticated vocabulary with good performers at the same grade?

4.44 Qualitative analysis

In addition to the quantitative analyses, the researcher also interviewed three of the 23 examiners for in-depth research. Their opinions on the performance of the candidates and on how they rated the examinations, especially how they rated vocabulary, were collected.

The three experienced Chinese local examiners of GESE who have examined all the three grades of Grade 2, Grade 5 and Grade 7 were interviewed. They were named Examiner A, Examiner B and Examiner C. First of all, each of them was asked to listen to several GESE examinations from the collected data, which they conducted in 2008, and they were asked to do a second marking according to the 2008 assessment criteria. They should explain why they gave such scores to the candidate. After the explanation, they were shown the scores they had awarded to the candidates originally. A further explanation was given by the examiner after the second marking of the candidates. After that, an assessment of vocabulary was asked if the examiner had not mentioned it. Finally the examiners were asked to talk about the vocabulary use of Grade 5 and Grade 7 candidates, which may shed further light on the question of "why there is no significant difference between the lexical variables of Grade 5 and Grade 7 candidates".

4.5 Summary

This chapter describes the research methods of the main study of the present research. First of all, refined sub-questions under the main research questions were formulated based on the insight gained from the pilot study. After that, the overview of the subjects was given; and then the data collection, transcription, main instruments of the GESE examination and the present research are described and finally, research procedures presented and discussed with reference to the research questions.

Chapter 5: Results of the Quantitative Analyses

In this chapter, both candidate (student/examinee) variables and teacher (examiner) variables are investigated based on the research questions. It is focused on three groups of relationships: first, the relationship between candidate (student/examinee) variables; second, the relationships between teacher/examiner variables and finally the relationships between candidate (student/examinee) variables and teacher/examiner variables.

5.1 The candidate (student /examinee) variables

In this section of the chapter, all the candidate variables obtained from the software of *CLAN* and *AG* and those obtained from GESE scores are investigated. It mainly focuses on the analysis of the relationship among these variables**.** Investigations were conducted to answer the first three research questions presented **Section 4.1**of **Chapter 4** Research Methodology of the main Study.

5.11 Lexical measures of the candidates of three different grades

In this part, the lexical measures of candidates/examinees of Grade 2, Grade 5 and Grade 7 are compared to investigate: first, if all the candidate/examinee lexical variables rise as the grade goes up; second, if there is any difference among the lexical variables of different grades, and finally what lexical variables can differentiate between different grades.

As mentioned in the previous chapter, the candidate lexical variables include *Token, Type, D, G, AG* and student *MLU*. The descriptive statistics are presented in Table 5.1.

It can be seen from Table 5.1 that with these variables, there is a trend that the measures rise as the grade goes up. Student *Token, Type, Ds* and *MLU* all go up as each grade rises. In other words, the higher the grade, the higher the measure of student Token*, Type, D* and *MLU*. However, the other two measures *G* and *AG* present another picture. For these two measures, the measures of Grade 5 are higher than those of Grade 2, but the measures of Grade 7 are lower than those in Grade 5.

For example, the mean *G* of grade 5 is 7.0, which is higher than the mean *G* of grade 2 of 4.4 and the mean *G* of Grade 7, which is 6.8. The mean AGs of Grade 5 is 198.5, which is higher than that of Grade 2 (161.9) and the mean AGs of Grade 7, which equals 142.6. The AG of Grade 7 is lower than that of Grade 2 and Grade 5.

Table 5.1 Comparison of the student/examinee lexical variables of three different grades

Variables	*Grade*	N	Mean	SD	ANOVA		
					F	*df*	Sig.
Token	*2*	58	166.9	42.4	37.6	2	<.000
	5	60	247.1	87.6			
	7	60	273.9	70.3			
Type	*2*	58	72.3	14.8	35.6	2	<.000
	5	60	100.4	25.8			
	7	60	104.2	24.6			
D	*2*	58	33.0	10.9	30.6	2	<.000
	5	60	47.1	11.9			
	7	60	49.3	13.7			
Guiraud	*2*	58	4.4	.98	34.0	2	<.000
	5	60	7.0	2.2			
	7	60	6.8	2.1			
AG	*2*	59	161.9	59.8	14.6	2	<.000
	5	60	198.5	58.4			
	7	60	142.6	54.3			
MLU	*2*	59	4.5	1.0	111.4	2	<.000
	5	60	11.3	4.6			
	7	60	15.6	5.3			

After obtaining the measures of different variables, the next step is to investigate whether the measures of different grades statistically differ significantly from one another. Analysis of variance (one-way *ANOVA*) shows that there is a statistically significant difference at the p=.05 level for the student variables: *Type, Token, D, G, AG* and *MLU* among Grade 2, Grade 5 and Grade 7. The statistics are also presented in T**able 5.1**. It indicates that the candidates of Grade 2, 5 and 7 may have very different lexical indexes.

In Table 5.2, the results of a mean comparison of the variables are presented. Post-hoc comparisons using the *Tukey HSD* tests indicate that the mean score of

Token in Grade 2 (Mean=166.9, SD=42.4) is significantly different from Token S in Grade 5 (Mean=247.1 SD=87.6) and in Grade 7 (Mean=307.2 SD=270.8.), but Grade 5 *Token* does not differ significantly from that of Grade 7. Similar results occur in the analysis of variance regarding *Type, D and G*. There are significant differences between the measures of Grade 2 and Grade 7, but there is no significant difference between measures in Grade 5 and grade 7 *AG* presents a different picture.

Table 5.2 p-values: multiple comparison of student/examinee lexical variables of different grades

Grades		*Token*	*Type*	*Ds*	*Gs*	*AGs*	*MLUs*
2	5	<.000	<.000	<.000	<.000	.002	<.000
2	7	<.000	<.000	<.000	<.000	.162	<.000
5	7	.090	.624	.569	.756	<.000	<.000

Tukey HSD tests indicates that the mean score of *AG* in Grade 2 (Mean=161.9, SD=59.8) is significantly different from *AG* in Grade 5 (Mean=198.5, SD=58.4), Grade 5 differs from Grade 7 (Mean=142.6, SD=54.3) , but Grade 2 does not differ significantly from Grade 7. This is unexpected, because the Grade 7 candidates should be much more proficient than those in Grade 2, and they are expected to use more advanced and more diverse vocabulary. For *MLU*, there is a significant difference among *MLUs* of the three Grades. The higher the Grade, the higher the *MLU*.

From the results it can be concluded that *MLU* is the only variable that can successfully distinguish between the three different Grades, and the index of *MLU* rises as the grade goes up. The results seem to confirm that *MLU* can be regarded as a general indicator of the language development for the Chinese learners of English from the beginner level to the intermediate level. For other student/examinee lexical variables of *Token, Type, D* and *G*, they can distinguish between Grade 2 and Grade 5, Grade 2 and Grade 7, but cannot distinguish between Grade 5 and Grade 7. The mean variable of Grade 5 is higher than those in Grade 2, but the mean variable of Grade 7 is lower than those of Grade 5. There is no significant difference between measures of Grade 5 and Grade 7. *AG* is a special measure among all the variables. The Grade 7 *AG* is lower than in both Grade 2 and Grade 5. It can distinguish between Grade 5 and Grade7, but it cannot distinguish between Grade 2 and Grade 7.

It is unexpected that there is no significant difference between Grade 7 and Grade 5 lexical measures and Grade 7 has lower indexes than Grade 5 in some cases of candidate/examinee lexical variables. The reasons might be first, the pass rate of Grade 7 is very low compared to Grades 2 and 5. In the present research, the pass rate of Grades 2, 5 and 7 are 83%, 55% and 25% respectively. Most students failed in Grade 7 examinations, which may distort the real situation. In the next step, only the data of the students who passed the examination are computed; the statistics are presented in Tables 5.3 and 5.4.

Table 5.3 Comparison of the lexical variables of the candidates who passed the examination.

Measures	Grade	N	Mean	SD	*ANOVA*		
					F	*df*	*Sig.*
Token	*2*	49	173.1	41.2	46.9	2	<.000
	5	33	288.7	82.4			
	7	15	303.4	62.1			
Type	*2*	49	75.6	13.3	64.4	2	<.000
	5	33	114.2	21.2			
	7	15	118.6	19.9			
D	*2*	49	34.8	10.4	39.7	2	<.000
	5	33	50.8	10.0			
	7	15	57.6	10.6			
Guiraud	*2*	49	4.6	.94	51.1	2	<.000
	5	33	8.1	2.1			
	7	15	7.5	2.1			
AG	*2*	49	53.3	7.6	6.3	2	<.003
	5	33	55.6	9.7			
	7	15	63.3	16.4			
MLU	*2*	49	4.6	.94	75.1	2	<.000
	5	33	13.2	4.8			
	7	15	15.5	6.3			

Table 5.3 shows that for all the candidates who passed the examination, all the

lexical variables go up as the grade rises. The mean value of all indexes (*Token, Type, D, G, AG* and *MLU*) of Grade 2 is the lowest among the three Grades. The indexes of Grade 5 are higher than those in Grade 2 and lower than those in Grade 7. One way *ANOVA* shows there is significant difference among the grades. However, multiple comparison in Table 5.4 shows that for all the indexes but *AG*, there is significant difference between Grade 2 and Grade 5, Grade 2 and Grade 7, but there is no significant difference between Grade 5 and 7. For *AG*, the only significant difference is between Grade 2 and Grade 5. There is no significant difference between Grade 2 and Grade7, Grade 5 and Grade 7.

Table 5.4 p-values: multiple comparison of the lexical variables of candidates who passed the examinations

Grade		*Token*	*Type*	*D*	*G*	*AG*	*MLU*
2	5	<.000	<.000	<.000	<.000	.002	<.000
2	7	<.000	<.000	<.000	<.000	.954	<.000
5	7	.722	.694	.088	.447	.072	.141

It can be concluded that when the examinees who failed in the examinations are excluded, all the candidate lexical variables can distinguish between examinees of Grade 2 and Grade 5; all the student/examinee lexical variables but *AG* can distinguish between examinees of Grade 2 and Grade 7; no lexical variables can distinguish between examinees of Grade 5 and Grade 7.

The fact that there is still no significant difference between Grade 5 and Grade 7 candidate lexical variables, even if only for qualified examinees, is unexpected. Grade 5 and Grade 7 examinations are prepared for learners of different proficiency, which can be seen clearly from the syllabus of each grade. But why is there no difference in quantitative measures? It is assumed that one of the reasons might be caused by different examination items. The data chosen for analyses in Grade 5 is conversation whereas the data chosen for analyses in Grade 7 is an interactive task. The interactive task is more impromptu and more challenging for Chinese candidates of GESE. Another reason might be that the level of the Grade 7 candidates is not higher than Grade 5 candidates, as would be expected. Further investigation needs to be carried out by looking at the results of the qualitative results. The qualitative analysis is presented in Chapter 6.

5.12 Different measures of good performers and poor performers at the same stage

In this section, the main purpose is to investigate whether there are differences between good performers and poor performers at the same grade. In the corpus of the present research, the pass rate of the Grade 2, Grade 5 and Grade 7 candidates are 83% , 55% and 25% respectively. Since the pass rate declines dramatically as the grade rises, I decided to make a very general distinction between good performers and poor performers: candidate who passed the test of a certain grade are regarded as good performers or qualified performers of the grade and those who failed the test are regarded as poor performers of the grade. An analysis of variance (one-way *ANOVA)* is conducted to analyze the variable difference between qualified performers and poor performers.

First of all, the variables of all subjects as a whole are investigated. The index of the Pass group and the Fail group are compared and the results are as follows: among all the variables, there is only a statistically significant difference between the Pass group and the Fail group in *AG* (p = <.000) and *MLU* (p =.003). That means only *AG* and *MLU* could distinguish the good performers and poor performers in the pooled data. There is no significant difference between the Pass group and the Fail group in *Token, Type, D* and *G*. Next, the Pass group and the Fail group are compared in each grade respectively to investigate whether the results are the same in separate grades.

The variables of Grade 2 candidates who passed the examination and those who failed the examination are compared. The main results are presented in Table 5.5.

It can be found from Table 5.5 that in Grade 2, the Pass group has a higher index in all variables than the Fail group. It also indicated that differences between the students who passed the Grade 2 test and those who failed in the test are significant. It can be inferred that all the measures of lexical richness and *MLU* can differentiate between Grade 2 candidates who passed the test and those who failed in the test.

Table 5.5 Differences between Grade 2 student variables of the Pass and Fail group

Measures	Pass group (n=49)	Failed group (n=10)	*F*	df	Sig.
Type	75.6	54.3	21.04	1	.000*
Token	173.1	132.9	7.62	1	.008*
D	34.8	22.7	.576	1	.002*
G	4.6	3.8	11.07	1	.001*
AG	172.7	109.2	10.95	1	.002*
MLU	4.6	3.7	6.43	1	.014*

**p*< .05

The same procedures are conducted for variables of Grade 5 and Grade 7. The main results are presented in Tables 5.6. and 5.7.

Table 5.6 Differences between Grade 5 student variables of Pass and Fail group

Measures	**Pass group (n=33)**	**Failed group (n=27)**	***F***	***df***	**Sig.**
Type	114.2	83.5	32.17	1	.000*
Token	288.7	196.1	22.64	1	.000*
D	50.8	42.5	8.19	1	.006*
G	8.1	5.7	2.28	1	.000*
AG	216.1	177.0	7.36	1	.009*
MLU	13.2	9.0	15.90	1	.000*

**p*< .05

Table 5.6 shows the differences of variables between Grade 5 Pass group and Fail group. Similar to the results obtained from Grade 2, it can be found that the Grade 5 Pass group has a higher index in all variables than the Fail group, and the differences between the students who passed the Grade 5 test and those who failed in the test are significant at p=.05 level. All the measures of lexical richness can differentiate between Grade 5 candidates who passed the test and those who failed it. The results of Table 5.7 present the variables of Grade 7, which shows a different picture from Grade 2 and Grade 5.

Table 5.7 Differences between Grade 7 student variables of Pass and Fail group

Measures	Pass (n=15)	Failed (n=45)	*F*	df	Sig.
Type	118.6	99.4	7.66	1	.008*
Token	303.4	264.1	3.68	1	.060
D	57.6	46.5	8.21	1	.006*
G	7.5	6.5	2.22	1	.141
AG	177.5	130.9	9.44	1	.003*
MLU	15.5	15.6	.003	1	.955

$*p< .05$

First of all, all the measures of the Pass group are higher than those in the Fail group, which is the same as that in Grade 2 and Grade 5. But for Grade 7, not all the differences between the Pass and Fail group are statistically significant. For the variable of *Token*, the Pass group has a higher mean (303.4) than the Fail group (264.1), but there is no statistically significant difference between the two groups. Similarly, there is no statistically significant difference between *G* although the pass group has higher indexes.

Regarding the difference between the Pass and Fail groups in Grade 7, only *Type, D and AG* in the Pass and Fail group differ significantly from each other, while *Token, MLU* and *G* cannot differentiate between them. It is worth noticing that as the grade rises to Intermediate Stage, not all the variables can distinguish between qualified and poor performers within the same grade, as they did in the lower stages. The results may suggest that *Type, D and AG* are more sensitive than these three measures in capturing the minute difference of a candidate's vocabulary use.

5.13 The relationship among the candidate variables

The differences of candidates' lexical variables among the three grades have been discussed in the previous section. In this part, the relationship between a candidate's lexical variables and GESE score variables is explored. First, the relationship among the candidate lexical variables is investigated, and then the relationship between candidate GESE score variables is explored, and finally the relationship between candidate lexical variables and GESE score variables is examined. *Two-tailed bivariate* correlation is conducted to investigate the relationship. As the variables are not parametric and normally distributed, the Spearmen's correlation coefficient (*rho*) is calculated.

5.131 The relationship among candidate lexical variables

The relationship between the candidate/student lexical variables are shown in Table 5.8. According to the guideline of Cohen (1988, pp. 79-81), the correlation coefficient is small when .10 < rho < .29 , is medium when .30 <rho< .49 and large when.50<rho< 1. It can be seen from Table 5.8 that the student variables of *Type* is significantly correlated with all other variables. Among them, *Type* is highly significantly correlated to *D* (rho=.739), *MLU* (rho=.670) and *GS* (rho=.924) and it is moderately correlated to *Token* (rho=.457) and *AG* (rho=.415). *Token* is moderately correlated with *Type* (rho=.457), *MLU* (rho=.367) and *G* (rho=.420), slightly correlated with *D* (rho=.264). The correlation between *Token* and *AG* is not significant and *rho* is only.141. *D* is slightly correlated with *AGs* (*rho*=.282) and moderately correlated with *MLU* (*rho*=.498) and it is highly correlated with G (*rho*=.714). *AG* is moderately correlated with *G* and it is not significantly correlated with *MLU. MLU* is highly correlated with *G. D* is claimed by the developer to have overcome the shortcomings of the measures based on *TTR*. But here the results show that it is highly correlated with *G*, which is a transformation of *TTR*, so the validity of *D* should be further investigated. It might not be totally independent of sample size as the creator of *D* claimed.

All the measures of lexical variables are to some extent correlated with each other except *AG*. *Type* is highly correlated with variables of *Token, D, G, and MLU***,** but only moderately correlated with *AG.* Similarly, *Token, D, G and MLU* are highly correlated with all other variables but *AG. AG* is slightly to moderately correlated with all variables but *MLU* and there is no statistically significant relationship between *AG* and *MLU. AG* is special in a sense that all other variables have a high or near correlation, but it only has a low to moderate correlation with other variables. The reason might be that it is the only index that can show both the diversity and the sophistication of a candidate's vocabulary use, while *Type, D, G* are all indexes of lexical diversity. *AG* is different from all other measures in function and as a result, the correlation between *AG* and other variables is low.

Table 5.8 The (2-tailed) Spearman correlations between the variables

Measures	1. Type n=178	2. Token n=178	3. Ds n=178	4. AG n=179	5. MLU n=179	6. Guiraud n=178
1. Type	1	.897**	.766**	.374**	.725**	.934**
2. Token		1	.485**	.264**	.814**	.840**
3. D			1	.282**	.498**	.714**
4. AG				1	.043	.337**
5. MLU					1	.762**
6. G						1

**. Correlation is significant at the .001 level (2-tailed)

*. Correlation is significant at the .005 level (2-tailed)

5.132 The relationship among GESE score variables

In this section, the relationship among the variables obtained from GESE scores is investigated. In addition to the four assessment categories of R*eadiness*, *Pronunciation, Usage* and *Focus*, another two indexes are added: *overall mark* and *overall mark 2*. *Overall mark* is the final mark of the whole test; it was obtained from GESE scores and it includes four levels: Pass with distinction (4), Pass with merits (3), Pass (2) and Fail (1). The *overall mark 2* is the mark of the dialogue/interactive phase of the test, from which phase the data of this present research is collected. It is calculated by SPSS by adding up the marks of all the assessment criteria of the dialogue or interactive phase. In Grade 2, overall mark 2 means the sum of the mark of *Readiness, Pronunciation* and *Usage* in the Dialogue Phase. In Grades 5 and 7, it means the sum of the marks of *Readiness, Pronunciation, Usage* and *Focus* of the Phase. It is found from the results of the pooled data that the four aspects of GESE scores of *Readiness, Pronunciation, Usage* and *Focus* and the final score of the whole test together with the final score of the Phase are highly and significantly correlated to each other (.580 <rho<.839, significant at the *p* = 0.01 level)

In the next step, the correlation coefficients of all GESE score variables of each grade are calculated to explore the relationship among them of each grade.

In Grade 2, all the variables of GESE scores are highly correlated (significant at the *p*= 0.01 level, 2-tailed) with each other. The highest correlation is rho=.961 between the final mark of the test (*overall mark*) and the mark of the sum of the three assessment categories (*overall mark 2*), and the lowest is rho=.671 between *Readiness* and *Pronunciation*.

Similar results are obtained from Grade 5. All the variables of GESE scores at Grade 5 are highly correlated (significant at the 0.01 level, 2-tailed) with each other, but the correlation coefficients are generally lower than those in Grade 2. The highest is .878 between the *Readiness* and the mark of the sum of the four assessment categories (*overall mark 2*), and the lowest is .514 between *Readiness* and *Usage*.

Results from Grade 7 show that all the variables of GESE scores are highly or moderately correlated (significant at the 0.01 level, 2-tailed) with each other. Different from Grades 2 and 5, not all the Grade 7 variables are highly correlated with each other. The highest correlation rho= .851 appears between *Readiness* and *overall mark 2* (the score of interaction part), and the lowest is rho=.431 between *Readiness* and *Pronunciation*. Another moderate correlation is between *Focus* and *Pronunciation*.

The results of both the pooled data and data in each grade show that all the variables of GESE scores are highly or nearly correlated (significant at the *p*= 0.01 level, 2-tailed) with each other. The results seem to confirm the conclusion of Malvern and Richard (2002) that there is a halo effect of subjective rating of oral examination. The idea is confirmed in the qualitative research in Chapter 6.

5.133 The relationship between lexical variables and GESE score variables

In the previous section, results indicate that all the GESE score variables are highly or moderately correlated with each other, and all the lexical variables but *AG* are also highly significantly correlated with each other. *AG* correlated slightly to moderately with all other lexical variables but *MLU*. In this section, the focus is to investigate the relationship between lexical variables and GESE score variables.

First of all, correlation *coefficients* are investigated for the pooled data. Table 5.9 shows the *Spearman* correlation between lexical variables and GESE score variables of the pooled data.

Among the lexical variables of *Type, Token, G, AG* and *D*, only *AG* is correlated slightly with *Usage* (vocabulary), and other lexical measures are not correlated with the assessment of vocabulary at all. The five lexical variables except *AG* have no correlation or have no positive correlation with pronunciation and overall mark at all, but they have moderate correlation with overall mark 2 and *Focus*.

AG is the only lexical variable that correlates moderately to slightly with all GESE score variables. *MLU* is correlated negatively with the score of Readiness, Pronunciation, Usage and Overall mark, and the correlation is low.

Table 5.9 The (2-tailed) Spearman's rank order correlations between lexical variables (LV) and GESE score variables (*GSV *)of the pooled data

LV GSV	**Type**	**Token**	**Ds**	**AG**	**G**	**MLU**
Readiness	. 183**	.099	.062	.309**	.117	-.182**
Pronunciation	.127	.035	.056	.197**	.061	-.156**
Usage (vocabulary)	.069	.006	-.015	.226**	-.006	-.253**
Focus	.410**	.297**	.303**	.314**	393**	.025
Overall mark	.080	.008	-.030	.316**	.001	-.298**
Overall mark2	.433**	.327**	.303.**	.316**	.371**	.116

$^{**.}$ Correlation is significant at the .001 level (2-tailed)

$^{*.}$ Correlation is significant at the .005 level (2-tailed)

The results indicate that first, among the all the examinee lexical variables, only *AG* is correlated with the score of *Usage* (vocabulary). *AG* is the only measure that is corresponsive to the GESE vocabulary score of the candidates and *AG* is the only measure that is correlated moderately or slightly to all GESE score variables. However, the index of *AG* is correlated with all the GESE scores, which may indicate that when engaging in the conversation and rating of the candidate, the examiner rater must apply some "economical marking strategy" (Daller and Phelan, 2007 p.235) rather than calculating all the assessment categories. The use of difficult words might be one of such strategies. This is also in accordance with the results of the pilot study. According to the results of the questionnaires in the pilot study, when asked how to assess vocabulary in oral interviews, the examiner raters choose lexical difficulty as the most important aspect to consider.

Second, *Focu*s rather than *Usage* is the only GESE assessment category that is moderately or nearly moderately correlated to all the lexical measures of the candidate. This may appear strange at first, but it seems reasonable after further analysis. This is maybe another economical marking strategy applied by the GESE examiners: if a candidate can give relevant answers to questions or give responses relevant to the conversation, he or she has mastered the vocabulary of a certain grade. This is also confirmed in the qualitative research in **Chapter 6.**

Since each grade has a different level of vocabulary use, putting them together may blur the real situation. In the next part, the lexical measures and GESE score measures of each grade are compared to investigate whether there is any relationship among them.

Table 5.10 The (2-tailed) Spearman's rank order correlations between lexical variables (LV) and GESE score variables (GSV) in Grade 2.

LV GSV(n=59)	Type	Token	D	AG	G	MLU
	(n=58)	(n=58)	(n=58)	(n=59)	(n=58)	(n=59)
Readiness	. 615**	.497**	.303**	.196	.502**	.280**
Pronunciation	.536**	.362**	.294**	.045	.448**	.335**
Usage (vocabulary)	.624**	.455**	.801**	.290**	.477**	.321**
Overall mark	.651**	.514**	.924**	.222	.505**	.342**
Overall mark2	.661**	.499**	.936**	.209	.535**	.329**

$^{**.}$ Correlation is significant at the .001 level (2-tailed)

In Grade 2, there are three assessment criteria, they are *Readiness, Usage* and *Pronunciation*. As indicated in the previous section, these three variables are highly correlated to each other. Table 5.10 shows the *(2-tailed) Spearman's rank order* correlations between lexical variables and GESE score variables in Grade 2.

All the lexical measures of Grade 2 are significantly correlated with GESE score variables but *AG*. As a measure of lexical richness, *AG* is only significantly correlated with the GESE variable of *Usage*, but not correlated with other assessment categories of *Readiness* and *Pronunciation*. The GESE score measure that is worth mentioning is *Usage*. As stated earlier in this research, *Usage* mainly includes the candidate's use of grammar and vocabulary and is regarded as the score of vocabulary in this research. *Usage* is the only GESE variable that correlates with all the student/examinee lexical measures in Grade 2. It shows high correlation with *D* (rho=.801$^{**)}$ and *Type* (rho= .624**), moderate correlation with *Token* (rho=.455**) , *G* (rho=.477**) and *MLU* (rho=.321**) and nearly moderate correlation with *AG*.(rho=.299**). So the results may indicate that the GESE score of *Usage* (vocabulary) is correlated with all the lexical measures in Grade 2.

Table 5.11 The (2-tailed) Spearman's rank order correlations between lexical variables (LV)and GESE Score variables(GSV) in Grade 5.

LV (n=60) GSV (n=60)	Type	Token	D	AG	G	MLU
Readiness	.516**	.423**	.302**	.365**	.405**	.309**
Pronunciation	.643**	.567**	.387**	.349**	.446**	.375**
Usage (vocabulary)	.591**	.576**	.262*	.333**	.478**	.396**
Focus	.637**	.568**	.405**	.373**	.536**	.413**
Overall mark	.668**	.622**	.331**	.369**	.548**	.445**
Overall mark2	.686**	.611**	.392**	.410**	.535**	.428**

$^{**.}$ Correlation is significant at the .001 level (2-tailed)

$^{*.}$ Correlation is significant at the .005 level (2-tailed)

Table 5.11 shows that in Grade 5, all the lexical measures and *MLU* are significantly correlated with GESE score variables. *Type* is highly correlated with all the GESE score variables (from .516** to .686**), *Token* and *Gs* are highly or moderately correlated with all the GESE score variables (.622** to .405**) , and *D* and *AG* are moderately to slightly correlated with all the GESE score variables (.410** to .262**). It is noticed that *AG* only correlates with *Usage* (vocabulary) of GESE score variables in Grade 2, but it correlates moderately with all GESE score variables in Grade 5. The correlation between *D* and GESE score variables is the lowest among lexical variables, and the correlation between *D* and GESE score variable of *Usage* (vocabulary) is only .262**, which is the lowest.

So the results may indicate that not only the GESE score of *Usage* (vocabulary) but all the other assessment categories are correlated with all the lexical measures and *MLU* in Grade 5. In Grade 5, there might be a heavier halo effect than in Grade 2.

Table 5.12 shows that in Grade 7, only a few lexical measures and GESE score variables are significantly correlated with each other. *Type* is slightly correlated with *Readiness* and *Usage* and moderately correlated with *Overall mark2*. *Token* is only slightly correlated to *Overall mark2* and *D* is slightly correlated with *Usage* (vocabulary) and *Overall mark2*. *AG* is slightly correlated with *Readiness* and moderately correlated with *Overall mark2*. *G* is only highly correlated with the *Overall mark2* and *MLU* has no significant correlation with GESE score variables at all. It is noticed that among GESE score variables, only *Readiness* and *Usage* (vocabulary) have significant correlations with lexical variables. The lexical variables of *Type, Token, D* and *AG* are moderately to slightly highly with *Overall mark*, the final score of the whole examination instead of *Overallmark2,* the score of the studied phase.

So the results indicate that the GESE score of *Usage* (vocabulary) is only slightly correlated with the lexical measures of *Type* and *D* in Grade 7. In Grade 7, the relationship between GESE score of vocabulary (*Usage*) and lexical measures are not straightforward at all in Grade 7.

Table 5.12 The (2-tailed) Spearman's rank order correlations between lexical variables and GESE score variables in Grade 7.

LV (n=60) GSV (n=60)	Type	Token	D	AG	G	MLU
Readiness	.285*	.207	.180	.276*	.100	-.124
Pronunciation	.090	-.017	.094	.123	-.003	-.139
Usage (vocabulary)	· 256*	· 134	260*	.191	· 110	-0.95
Focus	· 196	· 110	· 166	· 179	· 078	¨206
Overall mark	345**	.272**	. 283**	.339**	.192	-.074
Overall mark2	. 246	.129	.208	.228	.622**	¨169

**. Correlation is significant at the .001 level (2-tailed)

*. Correlation is significant at the .005 level (2-tailed)

5.14 Summary

In summary, the results presented in this section of the chapter seem to suggest that candidates of three different grades of GESE have shown different levels of vocabulary use.

There is a trend that the higher the grade, the higher indexes of lexical measures and *MLU*. In other words, the higher the grade, the more varied and more difficult vocabulary the candidate tents to use. The differences between Grade 2 and Grade 5, Grade 2 and Grade 7 candidate lexical variables are positively significant. However, the differences between Grade 5 and Grade 7 candidate lexical variables are generally not significant. All the GESE scores are highly correlated with each other, which suggest the holistic rating of GESE examiners. Regarding the relationship between the lexical measures and the GESE score variables, in Grades 2 and 5, most lexical variables are significantly correlated with GESE score variables. However in Grade 7, only a few lexical variables and GESE score variables significantly correlate with each other and the correlation coefficients are not high. The relationship between the score of *Usage* (candidate's vocabulary use) and candidate lexical variables is not clear or obvious at all.

5.2 Analyses of Examiner / Teacher Variables

In this part of the chapter, the examiner/ teacher variables are investigated. The focus of this section is to investigate whether accommodation occurs at lexical level. The analyses are based on question 4 and question 5 presented in **Section 4.1** of **Chapter 4 Research Methodology** of the Main Study.

Research Question 4

Will examiners accommodate lexically to candidates of different GESE grades? If so, how and to what extent do they accommodate to the candidate in vocabulary?

Research Question 5

Will examiners accommodate lexically to good performers and poor performers at the same GESE grade? If so, how and to what extent do they accommodate to different performers in vocabulary?

5.21. The teacher/examiner lexical variables of three different grades

First of all, one-way between-groups analysis of variance *(ANOVA)* is conducted to explore the differences among teacher lexical measures of three grades. The descriptive measures of the lexical measures of Grades 2, 5 and 7 are presented in Table 5.13.

It can be seen from Table 5.13 that among the teacher lexical measures, there is a trend that the measures rise as the grade goes up. Teacher *Type , D, G* and *MLU* all go up as the grade rises. In other words, the higher the grade, the higher the measure of teacher *Type, D, G* and *MLU*. This is exactly the same as that for student/examinee lexical *variables*. The *Type, D, G* and *MLU* of student/examinee also go up as the grade increases. However, the other two measures *Token* and *AG* present different situations. The *AG* of Grade 5 is higher than that of Grade 2, but the *AG* of Grade 7 is lower than that of Grade 5. The teacher *Token* is very special among the measures. Grade 2, the lowest grade, has the highest measure of teacher *Token*, but the teacher *Token* measure of Grade 5 is lower than that of Grade 7. The highest number of teacher *Token*s in Grade 2 is easy to understand. In Grade 2, the data chosen for research lasts for 6 minutes, which is one minute longer than the data chosen from Grade 5 and Grade 7. In Grade 2, the candidate's contribution to the conversation is very limited. They can only give very short answers to the question or perform some actions following the instruction of the examiner. On the other hand, the examiner has to speak almost all the time asking questions or giving instructions. As the grade goes higher, the examiner will speak less and the candidate will contribute more in each phase.

Regarding the language proficiency, as the candidate's proficiency level rises so he or she can contribute more in the interaction. The turns become longer and the candidate has more control over the conversation. In a fixed time of period, the more the examinee talks with more tokens, the less the examiner talks with less tokens. In the present research, there is a trend that the candidate *Token* goes up as the grade rises and the teacher/examiner *Token* declines in the same direction.

Table 5.13 The descriptive statistics of teacher lexical measures of three different grades

Measures	**Grade**	**Mean**	**Std. Dev.**	**Minimum**	**Maximum**	**ANOVA**		
						F	***df***	**Sig.**
Type	2	94.3	13.6	68.	144			
	5	99.1	18.5	58.	153			
	7	119.1	18.1	67.	159	36.0	2	<.000
Token	2	272.1	53.0	184.	456			
	5	218.6	61.4	87.	489			
	7	249.8	55.5	145	363	13.3	2	<.000
D	2	36.6	6.7	24.0	58.5			
	5	57.6	15.0	31.6	106.5			
	7	69.4	12.7	40.0	97.0	113.7	2	<.000
AG	2	110.3	35.0	33.6	201.4			
	5	160.5	55.0	53.6	321.7			
	7	137.5	48.9	62.7	301.9	16.92		<.000
Guiraud	2	5.7	.51	4.8	7.0			
	5	6.7	.64	5.4	8.4			
	7	7.6	.62	5.4	8.6	139.4	2	<.000
MLU	2	5.8	.50	4.9	7.1			
	5	8.9	1.3	6.4	14			
	7	14.8	3.3	8.37	22.8	283.8	2	<.000

In Grade 7, the examiner uses some prompts to start the topics in the Interactive

Phase. The prompts are a few sentences prepared to introduce a situation for discussion, and all the prompts are expressed in a rather fixed pattern by different examiners. In the Interactive Phase of Grade 7, it is the candidate's responsibility to keep the conversation going according to the syllabus. The examiner will not rephrase or explain in an elaborate way to the candidate as they do in Grade 5. That might be one of the reasons why the *AG* of Grade 7 is lower than that of Grade 5. Another reason is that in Grade 7, the examiners do not contribute new ideas as much as they do in Grade 5. In Grade 5 the examiners ask a lot of questions and thus bring in a lot of new words, while the examiners in Grade 7 mainly present and repeat the prompts; they do not contribute a lot in the Interactive Phase.

Post Hoc tests indicate that there are significant differences among the three groups of teacher/examiner lexical measures. Multiple comparisons show that all variables are statistically significantly different from each other apart from two pairs: the differences between Grade 2 and Grade 5 *Type* (*p.* = .284) and Grade 2 and Grade 7 *Token* (*sig.* =.084). It seems that the GESE Examiners use different vocabulary with candidates of different grades. Generally speaking, they tend to use more varied and more difficult vocabulary to candidates of higher levels than with those of lower levels.

The next step is to investigate whether the teacher/examiner lexical measures are correlated with student/examinee lexical measures. As Malvern and Richards (2002) mentioned, if there is a positive correlation between the examiner lexical measures and student lexical measures, it indicates that the examiners have applied accommodation strategies. Table 5.14 shows the correlation between teacher lexical measures and student lexical measures of the pooled data.

It shows in Table 5.14 that all teacher /examiner variables are correlated significantly with the student/examinee lexical measures. Among them, teacher variables of *D* and *MLU* correlate highly significantly with student *D* and *MLU*; teacher *AG* correlate moderately with student *AG*; Student *Type* has a slightly positive correlation with teacher *Type*. The results are very similar to those of the student/examinee lexical variables. The only significant negative correlation (*rho*=-.231 *p*<.001) appears between *TokenT* and *TokenS*. The negative correlation suggests that the more the student talks, the less the teacher talks, which is just the case in the interview of a certain grade which has a fixed time period. This is also easy to understand for examinations of different grades: for candidates of lower level, they usually answer each question with very limited words. For example, the mean *MLU* of Grade 2 is only 5.8 words, while the mean MLU for Grade 5 and 7 is 8.9 and 14.8 words respectively. In a limited period of time, since the student/examinee gives short answers, the examiner has to take more initiative and ask a lot of questions to keep the conversation going. As the level rises, the candidates are generally more proficient in speaking and have the ability to speak more during the interaction and take more initiative. In other words, in a fixed period of time, the higher the candidate's grade, the more the examinee speaks, and the less the examiner speaks. So the correlation between examiner and examinee token is negative.

Table 5.14 The (2-tailed) Spearman's rank order correlations between lexical measures of the examiner/teacher (T) and candidate/student (S) of the pooled data

Teacher/Examiner lexical variables	**Student/Candidate lexical variables**	***rho***
TypeT	TypeS	.220**
TokenT	TokenS	-.231**
Dt	Ds	.510**
AGt	AGs	.364**
Gt	Gs	.414**
MLUt	MLUs	.838**

**. Correlation is significant at the .001 level (2-tailed)

It can be concluded that in the pooled data, all the lexical measures of examiners are significantly correlated with examinee lexical measures, and examiner/ teacher measures of *Type, D, G, AG* and *MLU* are correlated positively with the student/ examinee measures. All teacher/examiner lexical measures except *Token* correlate with student/examinee lexical measures. Teacher/examiner *D* and *MLU* are correlated high with student *D* and *MLU*. Teacher *AG* and *G* correlated moderately with candidate *AG* and *G*. Accommodation does exist at lexical level in the interactions between the examiner and the examinee.

5.22. The examiner lexical variables for qualified and poor performers

The next question to be answered is whether the examiners use different vocabulary with good performers and poor performers at the same grade and whether there is lexical accommodation to different performers at the same grade. In order to answer this question, first of all a one-way between-groups analysis of variance *(ANOVA)* is conducted to explore the differences between the teacher lexical measures of the qualified group and the failed group of each grade. After that, correlation between the examiner (teacher) and candidate (examinee / student) measures of each grade is computed to investigate whether accommodation occurs within a grade.

Descriptive measures of the lexical measures are presented in Table 5.15. Table 5.15 shows that in Grade 2, although the examiner (teacher) measures of the Pass group are slightly higher than the Fail group, there is no statistically significant difference between the two groups. Similar results are obtained from Grade 7 examiner (teacher) variables.

Table 5.15: Grade 2 examiner/ teacher lexical measures of the Pass and Fail group; differences of the two groups (one-way Anova)

Measures	Qualified or failed group;	N	Mean	Std Dev.	**F**	***df***	**Sig.**
Token	Pass group Fail group	50 9	276 251	53.34 48.61	1.74	1	.19
Type	Pass group Fail group	50 9	95.62 87.22	13.73 11.07	3.00	1	.89
D	Pass group Fail group	50 9	36.9 35.1	5.79 10.43	.576	1	.45
G	Pass group Fail group	50 9	5.78 5.50	.49 .59	2.53	1	.12
AG	Pass group Fail group	50 9	110.4 109.5	36.7 26.6	.006	1	.94
MLU	Pass group Fail group	50 9	5.77 5.70	.50 .50	.158	1	.69

The Pass group has higher indexes in all the lexical variables, but there is no significant difference between them.

Table5.16 Grade 5 examiner (teacher) lexical measures of the Pass and Fail group; differences of the two groups (one-way ANOVA)

Measure	Group	N	Mean	Std. Dev.	F	*df*	Sig.
Token	Pass group	33	213.27	51.68	.552	1	.461
	Fail group	27	225.15	71.99			
Type	Pass group	33	99.00	16.34	.001	1	.982
	Fail group	27	99.11	21.20			
D	Pass group	33	60.18	15.99	2.21	1	.142
	Fail group	27	54.46	13.25			
AG	Pass group	33	160.72	54.00	.002	1	.967
	Fail group	27	160.12	57.20			
G	Pass group	33	6.82	.61	1.24	1	.27
	Fail group	27	6.63	.68			
MLU	Pass group	33	8.95	1.45	.35	1	.56
	Fail group	27	8.75	1.17			

Table 5.16 shows the results of Grade 5.

The results of Grade 5 are different from those in Grade 2. In Grade 5, the Fail group has a higher mean of *Type* and *Token* than the Pass group, but for other variables of *D, G, AG* and *MLU*, the Pass group has a higher mean than the Fail group. Again, there is no significant difference among any variable. The measures of Grade 7 are presented in Table 5.17.

Table 5.17: Grade 7 examiner (teacher) lexical measures of the Pass and Fail group; differences of the two groups (one-way Anova)

Measures	Groups	N	Mean	Std. Dev.	F	*df*	Sig.
Token	Pass	15	267	60.11	2.18	1	.145
	Fail	45	243	53.19			
Type	Pass	15	124.53	16.82	1.85	1	.179
	Fail	45	117.24	18.34			
D	Pass	15	71.43	12.40	.510	1	.478
	Fail	45	68.72	12.90			
AG	Pass	15	155.83	62.40	2.89	1	.094
	Fail	45	131.42	42.61			
G	Pass	15	7.64	.40	.383	1	.539
	Fail	45	7.53	.68			
MLU	Pass	15	15.12	4.24	.215	1	.645
	Fail	45	14.66	3.03			

It is shown in Table 5.17 that in Grade 7, although the examiner (teacher) lexical measures of the Pass group are slightly higher than those of the Fail group, there is no statistically significant difference between the two groups. The results obtained from Grade 7 examiner (teacher) variables are similar to those obtained from Grade 2.

It is found that there is no significant difference between teacher/examiner lexical variables when the data of different grades are analyzed separately, although the Pass group generally has a slightly high mean than the Fail group. Next, the correlations between examiner and examinee lexical variables of each grade are presented in Table 5.18.

Table 5.18 The (2-tailed) Spearman's rank order correlations between lexical measures of the examiners/teacher (T) and candidate/student (S) of different grades .

Measures		***rho***		
		Grade 2	**Grade 5**	**Grade 7**
Type T	**TypeS**	.221	-.064	-.084
Token T	**TokenS**	.072	-.210	-.186
Dt	**Ds**	.231	.212	.132
AGt	**AGs**	.346**	.353**	.178
Gt	**Gs**	.222	-.012	-.096
MLUt	**MLUs**	.190	.290*	.624**

**. Correlation is significant at the .001 level (2-tailed)

*Correlation is significant at the .005 level (2-tailed)

In Grade 2, the only statistically significantly correlation between examiner lexical variables and candidate variables is *AG* (*rho*=.346**) and the correlation is moderate. In Grade 5, there are two statistically significantly correlations between examiner lexical variables and examinee variables: *MLU* (*rho*=.290*), and *AG* (*rho*=.346**) The correlations are moderate to nearly moderate. In Grade 7, the only statistically significant correlation is between examiner and examinee *MLU*, and the correlation is high (.624**) So in the separated data, only the measures of *AG* and *MLU* have a positively significant correlation: the correlation between examiner and examinee *AG* is moderately significant in Grade 2 and Grade 5, and the correlation between examiner and examinee *MLU* is moderately significant in Grade 5 and Grade 7.

It is shown that although in the pooled data the teacher/examiner lexical variables are significantly correlated with respective student/examinee lexical variables, when data are separated into three grades, there are very few significant correlations. It indicates that the teachers/examiners do accommodate to different grades lexically, but within each grade, they are not finely tuned to any individual. They are tuned to the whole grade instead of individuals. This conclusion is similar to that of Malvern and Richards (2002). However, the difference between the results of the present research and those of Malvern and Richards is that in the present research, teacher *D* is not correlated with student *D*; however, teacher *AG* is significantly correlated with student *AG* in Grade 2 and Grade 5 and teacher *MLU* is significantly correlated with student *MLU* in Grade 5 and Grade 7.

In the next section, the relationship between the teacher lexical variables and the score of the studied phase (*Overall mark 2*) and the final mark of the examination (*Overall mark*) is investigated.

Table 5.19 shows again that there are a very few significant correlations between examiner lexical variables and the final score of the whole examination or the score of the studies phase. For Grade 2, the overall mark 2 is the transformation of the Overall mark, and there are two significant correlations: the correlation efficient between Overall mark 2 (*overall mark*) and *Type* and *G*. There is no significant correlation between teacher/examiner lexical variables and the scores of the students in Grade 5. Finally in Grade 7, the only significant correlation is between *Overall mark 2* and the *G* of Grade 7.

Table 5.19 Correlation between students' over-all mark and examiner variables of each grade

Scores	Grade	Type	Token	AG	D	MLU	Guiraud
Over all mark2***	G2	.351**	.242	-.040	.177	.027	.353**
Over all mark	G5	.067	-.022	.085	.147	.001	.156
Over all mark2		.128	.077	.149	.054	.091	.167
Over all mark	G7	.200	.170	.127	.055	.014	.088
Over all mark2		.214	.073	.103	.253	-.114	.324*

**. Correlation is significant at the 0.01 level (2-tailed)
*. Correlation is significant at the 0.05 level (2-tailed)

Regarding the research question 5, the results indicate that there is no significant difference between the teacher/examiner lexical variables applied to the students of the Pass and Fail group. The teachers/examiners are using the same level of lexical measures with qualified performers of the grade and the poor performers of the grade. Both correlations between the teacher/examiner lexical variables and the student/candidate lexical variables of each grade and the correlation between examiner lexical variables and candidate scores suggest that the examiners are not finely tuned to individuals. They did not use more diverse or more difficult vocabulary to the good performers of a certain grade than with poor performers of the grade. They are just generally tuned to the level of the whole grade.

Table 5.20 Correlation between students' GESE scores and examiner lexical variables of the pooled data

Scores	Type	Token	AG	D	MLU	G
Usage	-.050	.193**	-.060	-.288**	-.355**	-.208**
Readiness	.006	.161*	.037	-.168*	-.313**	-.124
Focus	-.043	-.073	.161	.046	-.262**	.034
pronunciation	-.011	.169*	-.062	-.237**	-.305**	-.162*

**. Correlation is significant at the 0.01 level (2-tailed)
*. Correlation is significant at the 0.05 level (2-tailed)

Finally the relationship between student GESE score variable of *Usag*e (vocabulary) and teacher lexical variables is investigated. First the relationship between *Usage* and teacher lexical variables of the pooled data is investigated. The results are presented in Table 5.20.

From the results of the pooled data that among the teacher/examiner lexical variables, *D, MLU* and *G* are negatively correlated with *Usage*, and the correlation is slight to moderate. *Token* is slightly correlated with *Usage*, and *Type* and *G* are negatively correlated with Usage, but the correlations are not significant. The negative correlation between *Usage* and some of the teacher/examiner lexical variables means the higher the candidate's score in vocabulary, the lower the lexical variables and *MLU* of the teacher. This result is just the opposite to what Richards and Malvern (2000) and Malvern and Richards (2002) claimed: that Teacher D is the only variable that is responsive to the student variables of the pooled data. This result may indicate that the examiners did not change their way of using vocabulary (lexical diversity or sophistication) according to their perception of the candidate's lexical use and level.

Another fact worth noticing is that the correlations between teacher/examiner lexical variables and GESE score of pronunciation and focus are very similar to that between teacher/examiner lexical variables and *Usage*, which indicate again the halo effect of subjective rating of oral interviews. This phenomenon is also founded in the research of Richards and Malvern (2000), Malvern and Richards (2002) and Malvern et al. (2004). Next, the correlation is calculated in each grade to investigate whether the examiners behave the same as shown in the pooled data.

Table 5.21 Correlation between students' GESE scores and examiner lexical variables of each Grade

Grade	Scores	Type	Token	AG	D	MLU	G
G2	Usage	.365**	.205	-.048	.220	-.053	.358**
	Pronunciation	.320*	.168	.073	.267*	.050	.367**
	Readiness	.233	.158	-.275*	.103	.028	.213
G5	Usage	-.011	-.026	.217	-.016	.235	.013
	Pronunciation	.276*	.259*	.227	.009	.018	.192
	Focus	.116	-.012	.097	.201	.093	.281*
	Readiness	.082	.060	.073	.032	.030	.092
G7	Usage	.226	.142	.024	.174	-.016	.274*
	Pronunciation	.088	-.008	-.039	.066	-.079	.201
	Focus	.095	-.009	.116	.174	-.209	.227
	Readiness	.316*	.179	.201	.252	-.103	.325*

**. Correlation is significant at the 0.01 level (2-tailed)

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

The separated data show a very different situation. It can be seen from Table 4.20 that for Grade 2, only *Type* and *G* are moderately correlated with *Usage*. While for Grade 5 the teacher lexical variables have no relationship with *Usage* and for Grade 7, only *G* has a slightly positive correlation with *Usage*. It can be summarized as only in Grade 2 do the examiners accommodate to candidates in a way that the more *Type* they use, the higher the student's score of *Usage*. But for Grade 5 and 7, there is almost no relationship between teacher lexical variables and candidate GESE scores of *Usage* (vocabulary). In Grade 7, teacher lexical variables of *Type* and *G* have a moderate relationship with readiness.

5.3 Summary of the results of quantitative analyses

It was found from the answers to the research questions that there are some characteristics of the teacher/examiner and student/candidate/examinee lexical variables at in different GESE grades.

Regarding the first research question of the present research, the results can be summarized as follows. It was found that in the pooled data there is a trend for all the student/candidate/examinee lexical variables to go up as the grade rises, and there are statistically significant differences among different grades. All the lexical measures and *MLU* can differentiate between candidates of Grade 2 and Grade 5 and can differentiate between candidates of Grade 2 and Grade 7 as well. However, the lexical measures and *MLU* cannot distinguish between Grade 5 and Grade 7 candidates. When only the candidates who passed the examinations are calculated, the results are very similar: first, the higher the grade, the higher the student/ candidate/examinee lexical measures and *MLU*; second, the student lexical variables can differentiate between candidates of Grade 2 and Grade 5 and candidates of Grade 2 and Grade 7(except *AG*) but cannot distinguish between Grade 5 and Grade 7 candidates.

Regarding the second research question of the present research, the results can be summarized as the following. At first when the pooled data was analyzed, it was found that only the student/candidate/examinee *AG* and *MLU* can distinguish between the qualified performers and the poor performers. Then data was analyzed for each grade. In Grade 2 and Grade 5, the results are the same: all the student/candidate/examinee variables and *MLU* can distinguish between the candidates who passed the examination and the poor performers who failed in the same grade of examination. The lexical variables and *MLU* of the candidates who passed the examination are higher than those who failed in the examination. In Grade 7, only the student lexical variables of *Type, D* and *AG* can differentiate between candidates who passed the examination and the poor performers who failed it although the candidates who passed the examination have higher lexical variables and MLU of than those who failed it.

Regarding the third research question of the present research, the results are summarized as the following. Firstly, *MLU* and all the student/candidate/examinee lexical variables except *AG* are highly correlated with each other. *AG* is moderately correlated with *Type* and *G*, slightly correlated with *D* and *Token* but it is not correlated with *MLU*. Secondly, all the student/candidate/examinee GESE score variables are highly correlated with each other, which show a halo effect of the rating. Thirdly, the relationship between the student/candidate/examinee lexical variables and GESE score variables are not straightforward. In the pooled data, the score of *Usage* (vocabulary) is only slightly positively correlated with *AGs*. In the separated data, the GESE score of *Usage* (vocabulary) is significantly correlated with all student lexical measures and *MLU* in Grade 2. The highest correlation (rho=.801 p=.001) is between *Usage* and *D* and the lowest (rho=290 p=.001) is between *Usage* and *AG*. In Grade 5, the GESE score of *Usage* (vocabulary) is significantly correlated with all student lexical measures and *MLU*. The highest correlation (rho=.591 p=.001 is between *Usa*ge and *D* and the lowest (rho=.262 p=.05) is between *Usage* and *AG*. In Grade 7, the GESE score of *Usage* (vocabulary) is only slightly correlated with *Type* and *D*.

Regarding the fourth research question of the present research, the results are summarized as the following:

Concerning the teacher/examiner variables, there is also a trend for the teachers to use higher indexes of lexical measures to candidates of higher grade. The higher the candidate's grade, the higher the indexes of examiner lexical measures and MLU. Teacher/examiner *Token* is an exception, the lowest grade tends to have the highest teacher/examiner Token. All the teacher/examiner variables can differentiate between candidates of different grades; however they cannot differentiate between qualified performers and poor performers at the same grade.

The examiner lexical variables and *MLU* are significantly correlated with candidate lexical variables and *MLU* in the pooled data. The only negative correlation is between candidate and *Token* and examiner *Token*. However, such correlations are absent in the separated data of each grade.

Regarding the fifth research question of the present research, the results can be summarized as follows. Firstly, it is found that the examiners generally use higher lexical and *MLU* to the qualified performers (with the exception of *Token* and *Type* in Grade 5), however the differences are not statistically significant at all. It suggests that the examiners did not use more diverse and sophisticated vocabulary to better performers. Secondly, within the same grade, only very few examiner lexical variables show a correlation with candidate variables. Examiner *AG* is correlated with candidate *AG* in Grade 2 and Grade 5. Teacher / examiner *MLU* is correlated with candidate *MLU* in Grade 5 and Grade 7. There is hardly any relationship between examiner lexical variables and the GESE score of *Usage* (vocabulary). It is found that for Grade 2, only *Type* and *G* are moderately correlated with *Usage*. While for Grade 5 the teacher lexical variables have no relationship with *Usage* and for Grade 7, only *G* has a slightly positive correlation with *Usage*. The results from this research show that only in Grade 2 do the examiners accommodate to candidates in a way that the more *Type* they uses, the higher the student's score of *Usage*. But for Grade 5, there is almost no relationship between teacher lexical variables and candidate GESE score of vocabulary. In Grade 7, teacher lexical variables of *Type* and *G* have a moderate relationship with *Readiness* rather than with *Usage* (vocabulary) of the candidate.

Chapter 6 The Qualitative Analyses

In addition to quantitative analyses, qualitative research was also conducted for the present project. It was found that some of the results of the quantitative analyses could not be interpreted by the quantitative data, as: 1) there is no direct relationship between lexical measures of the candidates and their GESE scores for vocabulary, and 2) some lexical variables of Grade 5 candidates are higher than those in Grade 7, and there is no statistically significant difference between the lexical measures of Grade 5 and Grade 7.

The qualitative research was conducted based on the above-mentioned questions and it was expected to get more insights and interpretation of the quantitative results. The main research questions of this qualitative research are: 1) what are the factors that influence the examiners' rating of vocabulary? 2) What are the factors that lead to the comparatively low lexical indexes of Grade 7 candidates?

6.1 The participants

The participants who were interviewed for the qualitative research were three experienced Chinese local examiners of GESE who have been examining all the grades from the Initial St*ag*e to the Intermediate Stage (Grade 1 to 9). Among all the 23 examiners involved in the quantitative research of the present research, there were 12 examiners who have only conducted the first 2 Stages (Grade 1 to 6) and there are 11 senior examiners who have conduct the first 3 Stages (Grade 1- to 9). The three subjects were chosen because they were the top three who had conducted more examinations than others in the year of 2008. (See **Table 4.2** Examiner information). There are two female examiners and one male examiner and they are coded as Examiner A, Examiner B and Examiner C by the researcher. For Examiner A, she conducted 21 examinations among the total 180 examinations collected for the present project. Examiners B and C conducted 16 and 14 examinations respectively.

For Examiner A, four examinations were chosen for the second marking and interview, one from Grade 2, one from Grade 5 and two from Grade 7. For Examiner B, two examinations were chosen for the second marking and interview, one from Grade 5 and another from Grade 7. Three examinations conducted by Examiner C were chosen for qualitative research, and with one examination from each of the three grades of Grade 2, Grade 5 and Grade 7. Totally nine examinations were collected for a second marking and interview, two from Grade 2, three from Grade 5 and four from Grade 7. The cases were chosen based on the final score and the score of vocabulary (*Usage*) of the original markings. I tried to choose cases with different levels of scores: for the final score, three levels of scores B, C and D were included. With regard to the score of vocabulary (*Usage*), A, B, C, D and E were all included. The detailed information of the original marking and the second marking is presented in **Table 6.2** in **Section 6.5.**

6.2 Data collection and procedures

Each interview was conducted according to the interview plan, which is presented in **Appendix 8**. The data collection procedures mainly include two parts: the re-marking or the second marking of an examination and the interview with the examiner after each re-marking. The process of each re-marking and interview is presented in Diagram 6.1

First, the re-marking or the second marking of the GESE examination was conducted. Each of the three senior GESE examiners was asked to remark 2 to 4 examinations he or she had conducted in the year 2008. Altogether 9 examinations were re-marked by the 3 examiners. The main purpose of the second marking is to provide the GESE examiners with the chance to talk about how they rated the candidates, especially to elicit their comments on the vocabulary use of the candidates and how they rated vocabulary in GESE.

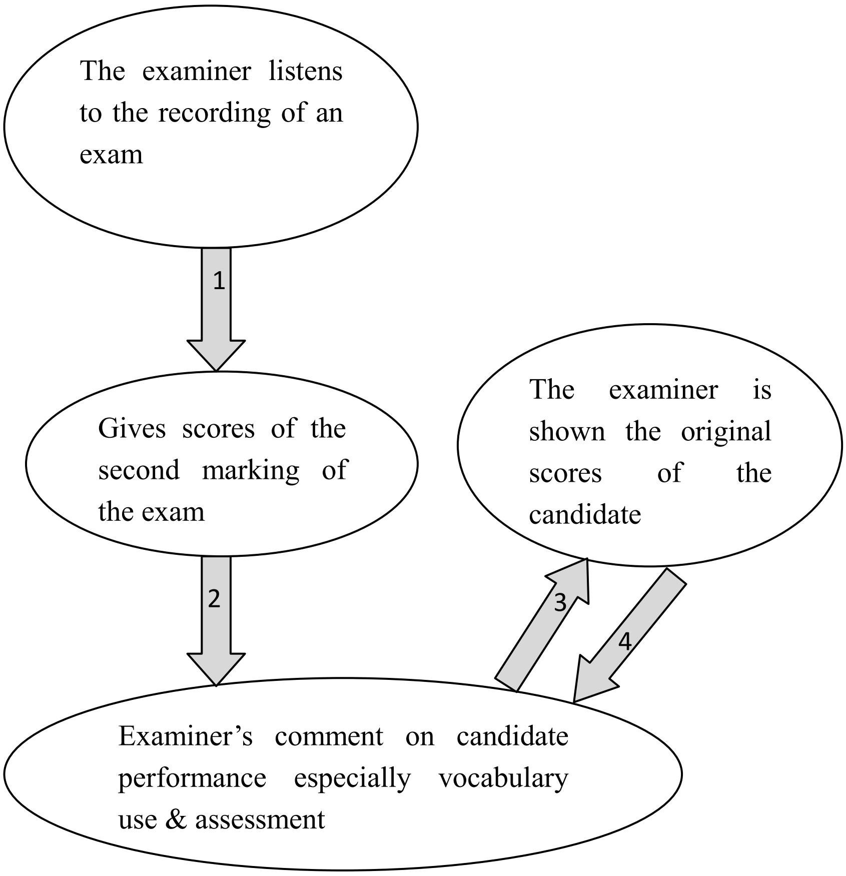
For each interview, the examiner was asked to listen to a recorded examination, but only the studied phase of each grade was replayed. To be more specific, the whole examination of Grade 2 (6 minutes), the Conversation phase of Grade 5 (5 minutes) and the Interactive task phase of Grade 7 (5 minutes) were replayed for re-marking and follow-on interview. After listening to the recording of the studied phase of each examination once, the examiner was asked to rate the phase according to the 2008 assessment criteria listed in the GESE syllabus. So both an overall score and scores for different assessment criteria were collected.

Following this, the examiners were asked to talk me through the reasons why they gave such scores to a candidate right after the re-marking. This was kept as a very open question and the examiners were invited to talk about all the issues they perceived as relevant. The examiners were also invited to talk about the candidate's vocabulary use if they had not mentioned it earlier in the interview.

After the second marking and the examiner's narratives on how they rated the candidate, the original scores were shown to the examiner and the examiner was invited to provide any reaction to differences or similarities.

During the first two steps of the interview, the researcher spoke only when it was necessary and tried to give very brief responses during interactions. The researcher tried to draw information from GESE examiners without leading the interviewees.

Figure 6.1 The data collection procedures of each re-marking and interview


The examiner listens to the recording of an exam
1
Gives scores of the second marking of the exam
2
The examiner is shown the original scores of the candidate
3
4
Examiner's comment on candidate performance especially vocabulary use & assessment

Finally, after each examiner had finished the second marking of all the examinations, in addition to the procedures mentioned above, the researcher revealed to the examiner the results regarding lexical variables of Grade 5 and Grade 7 candidates. The Examiners' opinions were collected.

All the interviews were audio-recorded with the permission of the three examiners. The recordings were transcribed by the researcher. The transcription was checked three times on three different days to make sure that all the transcribed information was complete and faithful to the original interview.

6.3 Coding and analyses of the qualitative research data

The transcription of the interview was carefully read and reviewed by the researcher in an iterative way. After that, the coding and analyses was conducted by the researcher. Attride-Sterling (2001, p. 390) proposed that "the full process of thematic analysis can be split into three broad st*ag*es: (a) the reduction or breakdown of the text; (b) the exploration of the text; and (c) the integration of the exploration". The analysis process of the qualitative data also follows the three general steps. However, the detailed procedures are described in the following paragraphs.

The researcher organized the qualitative data in two stages:

In the first stage, the data was segmented according to the two themes that are derived from the two research questions: factors that affect rating assessment and the examiners' opinions on the vocabulary use of Grade 5 and Grade 7 candidates. Segments are selected and sorted according to the two themes.

In the second stage, for each theme, a set of categories was derived from the segments of the text. In this present research, the steps of qualitative analysis were based on Selinger and Shohamy (1989, pp.205-207): The interviews with Examiner A were carefully reviewed, and notes about Examiner A's ideas and opinions were made.

1. A list of viewpoints that were derived from the data was compiled. presents the summary of Examiner A's information.

2. The list was analyzed in an attempt to collapse and combine certain categories of opinions.
3. A finite group of patterns or sub-patterns was formulated.
4. The pattern and categories of opinions were applied to the rest of the 5 interviews for further refinement.
5. A definitive group of patterns and categories of opinions was formulated.
6. To examine the reliability of the data, the whole process was revisited by the researcher three months after the analysis. The patterns on which the two analyses agreed were applied in the coding and analysis.

Table 6.1 The display of the summarized data of Examiner A

Examiner	Summaries, paraphrases	Direct quotes from candidate (A)
	Opinions on the overall performance of the candidates	
Examiner A	Giving a holistic score; Overall feeling or intuition; Rating categories such as relevance, vocabulary and grammar are also considered:	*Generally speaking he (C1) is good. *She (C3) should have passed the grade. *He (C2) did not provide much information…not very relevant to the topic.
	The assessment of the candidate's vocabulary use;	
Examiner A	Relevancy; Active use of vocabulary; Enough vocabulary to keep the conversation going; Range of vocabulary.	*For a Grade 5candiate, he (C2) can use his limited vocabulary to manage the topic. That's all right. * If a candidate doesn't have the Grade 7 vocabulary, he or she can't talk about the topics in Grade 7. * Many candidates of the same grade (Grade 5) prepared the same things, the same topic and similar answers on a certain topic. However, an important point that I to pay attention to is whether they can understand the questions, whether they can answer the questions and whether their answers are relevant to the questions.
	The vocabulary use of Grade 5 and Grade 7 candidates	

Examiner A	There should be differences between Grades 5 & 7; Difficulty of the interactive part of Grade 7;	*There must be a difference between the vocabulary use of a C candidate in Grade 5 and a C candidate in Grade 7. There must be something wrong if there is no difference. * If a Grade 7 candidate doesn’t have the vocabulary needed for different topics in Grade 7, he will definitely fail the grade, because he can’t talk about the topics at all. If a Grade 7 candidate only has the vocabulary of Grade 5, he will definitely fail the grade. * The interactive part of Grade 7 is very difficult.
	Grade 7 candidates	
Examiner A	Lack of understanding of the syllabus; Not enough feedback from the examination centre in 2008; Aiming at higher grade than the candidate’s real level; Exam-oriented inappropriate training; Cultural factors.	* I met a candidate that can be called the “craziest” candidate I’ve had ever had. The candidate got a D (Fail) for Grade 4, but he took Grade 7 very soon. How could such a candidate pass? I think the candidate and their parents were just trying their luck. Many Grade 7 candidates took a wrong grade. *I met some Grade 7 students from a training school. The candidates always answered “if you ask me the question, I will…” and “due to the fact that…” no matter what question you asked. It’s silly. *Chinese students are reluctant to ask questions, either in class or in lectures *If a candidate has a Grade 7 certificate at that time (2008), he or she will be accepted by The Affiliated Middle School in Beijing…The parents are also very keen on it. The children who wanted to enter a key middle school all take Grade 7. *The Chinese students are more self-conscious and they are afraid of asking questions. They are afraid of losing face if they can’t speak in a

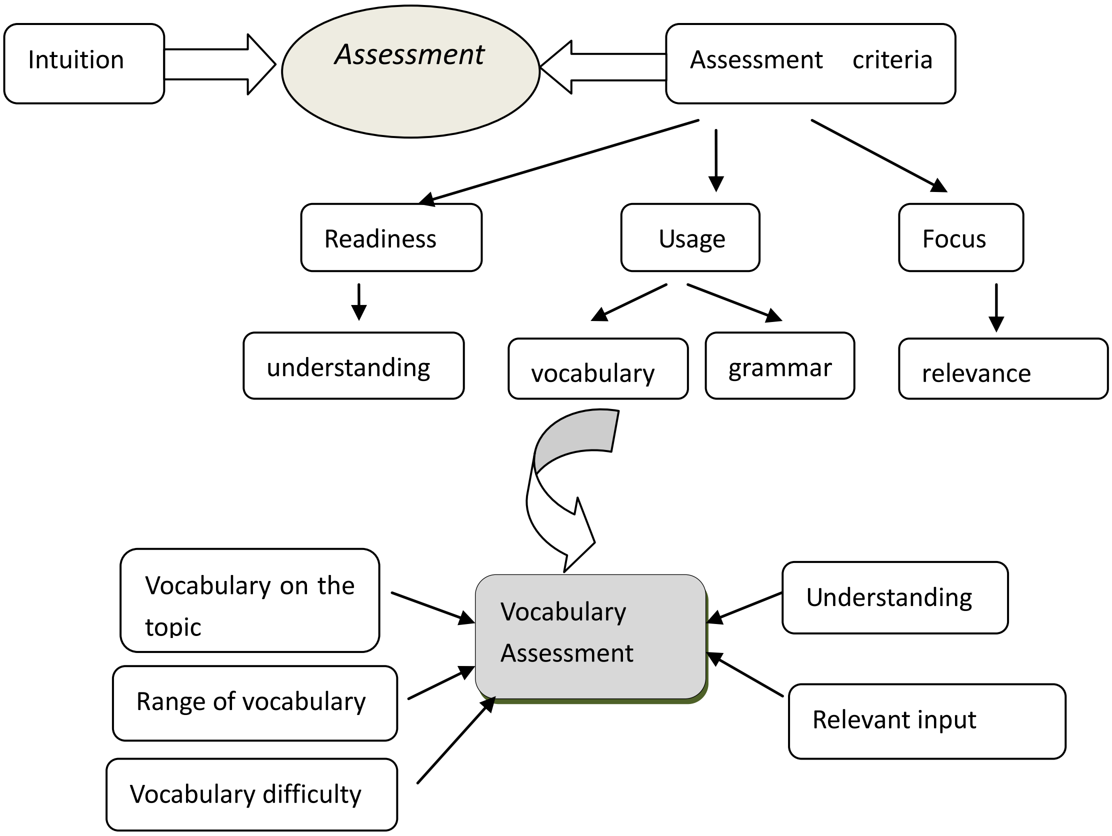
		perfect way; they are afraid they will offend the teacher if they ask a wrong question.

6.4 Results of the research on the first theme: factors that affect rating of GESE

With regard to the first theme, which emerged from the first interview question, the results are presented in this section of the chapter. Categories emerged during the coding and analysis of the qualitative data and the coding map is presented in Figure 6.2.

During the interview, when the examiners talked about factors that affect the rating of the candidates, two rather contrasting categories emerged, namely: intuition and assessment criteria. Under the assessment criteria, four sub-categories emerged: understanding, vocabulary and grammar and relevance

Figure 6.2 Factors affecting assessment



When vocabulary assessment was discussed, the opinions of the examiners on vocabulary use of the candidates were also elicited. Five sub-categories emerged that are associated with the assessment of vocabulary. Among the five sub-categories, two are related to vocabulary richness: *range of vocabulary* and *vocabulary difficulty*. The other three sub-categories are: vocabulary on the topic, understanding of the examiner and relevant input of the candidate during the interaction. In the following sections, the factors that the Chinese local examiners associated with GESE assessment are fully discussed based on the coding map, and some of the interpretations are provided. The assessment of vocabulary is also discussed with regard to the quantitative results obtained in Chapter 5.

6.41 Intuition vs. analytical assessment based on assessment criteria

The categories of *intuition* and *assessment criteria* emerged from the text.

It seemed that the examiners make a holistic assessment mainly based on their professional intuition and their understanding of the assessment criteria listed in the syllabus.

After listening to each examination, all the examiners tended to give a holistic *score according to their professional intuition, and this is reflected in their use of words* like *feel*, *feeling* and *intuition*. On the other hand, they also referred to the assessment criteria. When they explained how they rated the candidates, they mentioned some of the GESE rating criteria, but the criteria of *Readiness* and *Focus* are more stressed than those of *Usage* and *Pronunciation*.

From the interview it was found that the factors the examiners stressed are understanding and relevant input during the interactions. *Understanding* means the candidate should understand the examiner and eng*ag*e in meaningful communication, which refers to the assessment criterion of *Readiness* according to the syllabus of GESE. Vocabulary and grammar on the other hand, mainly refers to the criterion of *Usage*, and finally *Focus*, is related to relevant input or information from the candidate in the Elementary and Intermediate Stages. The criterion of *Pronunciation* is seldom mentioned in the interview. (Refer to **4.33** of **Chapter 4** for the full description criteria for different stages of GESE).

Holistic rating based on intuition seems to function as a dominant factor in rating, and criteria seem to be rather auxiliary factors. For example, all the examiners stressed intuition when they gave the holistic mark of the candidate. After listening to Examination 2, Teacher A said immediately, "I feel this is C." And Examiner B said after listening to Examination 5, “I feel it’s C according to the assessment standard we are using now.”

Examiner B’s explanation on how he rated a candidate may well represent the situation of GESE assessment:

> Even in 2008 when we were using the analytical system, I usually gave the candidate a general assessment like ‘this is an A candidate’ or ‘that is a D candidate’. If I feel there is a C candidate, after the general assessment, I’ll mark a phase according to the three or four assessment categories. But if I found the analytical scores were not consistent with my general judgment, I’d make some minor change of the analytical scores. The analytical scores should be consistent with my general judge of the candidate.

Since the Chinese examiners based their rating mainly on feelings or intuition, the rating style is rather holistic although they were asked to do analytical rating. The results of the qualitative research are also in line with the quantitative result presented in Chapter 5. According to Chapter 5, both in the pooled data and in the separated data of each grade, all the GESE Score variables are highly or moderately correlated with each other, which showed a heavy halo effect of rating. Holistic rating is dominant in GESE rating. Many researchers (Malvern and Richards 2002; Malvern et al. 2004) also noticed the problem of holistic rating in proficiency interviews. Malvern and Richards (2002) found that the Range of vocabulary correlates extremely highly with the other rating scales and all the inter-correlations in the matrix are above .900, among which the highest is between Range of vocabulary and Content at .996. They believed that “the rating of range of vocabulary is likely to be heavily contaminated by halo effects” (p.95).

Holistic rating may also explain why the second marking of the examinations in the qualitative research is very consistent with the scores of the original marking. The examiners might adopt a holistic rating even when GESE required analytical rating in 2008. This may also explain why the holistic rating system was used by Trinity to replace the analytical rating. The Chinese local examiners of GESE in Beijing began to adopt the holistic rating system, which was granted by Trinity, London in 2010. The research result provides evidence to support the system change.

6.42 Assessment of vocabulary

The qualitative analysis provided more information on how GESE examiners assess vocabulary and also provided interpretation of some of the results of the quantitative research.

When GESE examiners assess vocabulary, they do not just look at the lexical performance of the candidate exclusively. Five factors that may affect the examiners' rating of vocabulary emerged through the analysis. Three out of five are related to the candidate's use of vocabulary and the other two factors are related to the communicative ability of the candidate.

It seems that the most important factors that examiners consider in assessment are whether the candidate can understand the examiner, and whether the candidate can engage in meaningful interaction with relevant input. The examiners expressed very similar ideas on the assessment of vocabulary: if a candidate is able to manage to talk about the topics appropriate to a certain grade with relevant responses, he or she understands the examiner well and the vocabulary is more or less satisfactory for that level. On the contrary, if the candidate cannot communicate with the examiner even with help, the candidate's use of vocabulary is not satisfactory even if he or she uses some less-frequent words. So in the assessment of vocabulary, what the examiners stressed is not the different aspects of lexical variety or difficulty, but the general communicative ability. Aspects of lexical richness are considered by the examiners, however with less importance.

The result of the qualitative analysis indicates that there is a more complete and complicated picture of vocabulary assessment than that presented in the pilot study. According to the results of the questionnaires in the pilot study, it seemed that difficult words and range of vocabulary are important indicators of vocabulary use, and appropriateness, communicative skills, subjects, register are also mentioned as other important factors which affect lexical assessment in the open question of the questionnaire. However, the qualitative analysis indicated that in addition to the vocabulary range and vocabulary difficulty, factors such as understanding of the examiner and relevant responses or input are more closely related to communicative abilities rather than vocabulary use:

Examiner A remarked that for the Grade 2 candidate in Examination 1, although the candidate has a good range of vocabulary for Grade 2, she only gave the candidate B for *Usage* (vocabulary) and for the overall mark because the candidate did not get the meaning of several sentences and gave irrelevant answers to the questions.

Examiner B also expressed a similar idea when he commented on the vocabulary of a Grade 7 candidate in Examination 6.

> This Grade 7 candidate doesn't have an impressively large vocabulary, nor did he use big words, but he has a good understanding of the less-frequently used words for his grade such as ***substantial, worthy, trivial***. In addition, he understood the prompts well and gave relevant responses, so I marked him as a C candidate.

From the examiners' comments it can be found that they believe understanding and *relevant responses at each grade* is the baseline category to assess vocabulary. All three senior examiners stressed it when talking about the assessment of vocabulary.

The results regarding the factors that affect the assessment of vocabulary also explain some of the results in the quantitative results presented in Chapter 5. In the pooled data, the only GESE score variable that is significantly correlated with all candidate lexical variables is *Focus*. *Focus* mainly means sufficient and relevant information required by the task set, coherent organization of the information and opinions and abilities to maintain the conversation according to the GESE syllabus used in 2008. Relevant responses or input is the essence of *Focus***.** The qualitative data explained why Focus is correlated with all lexical variables in the quantitative data, and the reasons behind the examiner's behavior are revealed.

The qualitative analysis also indicated that a candidate’s use of vocabulary is closely related to the GESE rating criterion of *Readiness*. In the quantitative analysis, *Readiness* is significantly correlated with a candidate's lexical variables of *Type* and *AG* in the pooled data, which may suggest *Readiness* is correlated with *Type* and *AG*, the measures of vocabulary diversity and difficulty respectively. According to the GESE syllabus used in 2008, *Readiness* mainly includes: the candidate’s understanding of the examiner; maintaining the flow of conversation and satisfying the requirements for each grade and for all previous grades. Understanding of the examiner and the task set is the essence of *Readiness*. After the qualitative interview, we may get a better understanding of the relationship between the lexical variables of the candidates and *Readines*s.

The outcome of the qualitative result may provide an interpretation of the quantitative results regarding the relationship between candidate lexical measures and GESE score variables, and it may also suggest that Chinese examiners rated vocabulary the same way as expressed in the qualitative data. This is significant for the triangulation of the data.

The qualitative data also suggest that relevant responses on topics appropriate for a certain grade might be another *economical marking strategy* in vocabulary assessment of GESE. Although *relevant response* is not a measure of lexical richness, it shows the general language proficiency of the candidate, which embodies the use of vocabulary. It seems that the examiners were using relevant responses as what Daller and Phelan (2007, p.235) had proposed: an *economical marking strategy* in rating vocabulary in GESE. *Economical marking strategy* here means reliable holistic rating. This assumption is easy to understand: when it is not practical for examiners to compute different aspects of lexical richness during marking of writing or spoken text, they tend to use some *economical marking strategy or highly reliable overall rating* to assess the candidate’s language proficiency. As has been discussed in the previous section, all examiners seem to agree that if a candidate can give relevant responses during the interaction, he or she at least understands the examiner and the task set, his or her vocabulary will basically meet the requirements of a certain grade. However, if a candidate cannot give relevant responses during communication, he or she does not understand the examiner and the task, and his or her vocabulary is not satisfactory even if the candidate can use varied and sophisticated vocabulary.

6.5 Factors contributing to Grade 7 candidates' poor performance

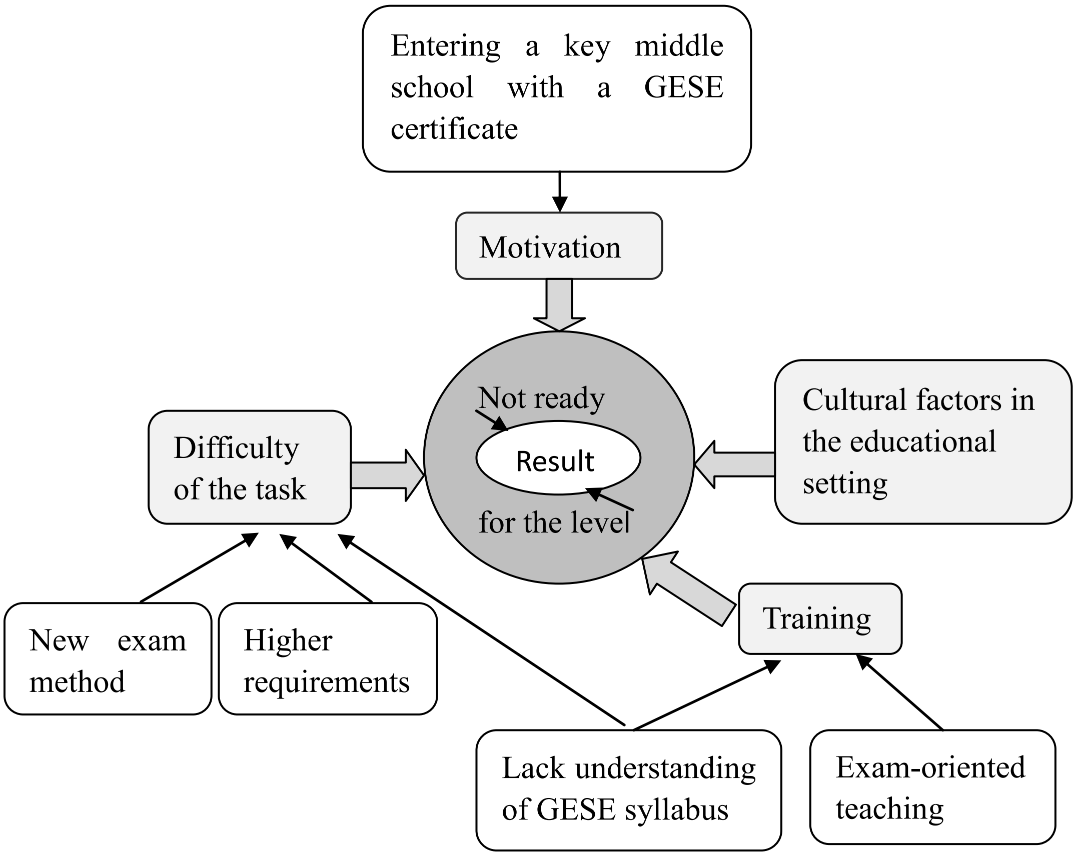
The results of the quantitative research indicate that there is no significant difference between most lexical measures of Grade 5 and Grade 7 candidates, and some Grade 5 candidates use more varied and difficult vocabulary than Grade 7 candidates. When the unexpected result concerning the vocabulary use of Grade 5 and Grade 7 candidates was reported to the examiners interviewed, all of them suggested that the Grade 7 candidates should have had a larger vocabulary and they should have applied more difficult or less-frequent words than Grade 5 candidates, because the topics in Grade 7 are less-frequently talked about and they are more abstract and difficult. The grammatical functions listed in Grade 7 are more complicated than those in Grade 5. There must be reasons that caused the unexpected results.

Concerning the unexpected outcome obtained from the quantitative results, some factors that may emerged from the data provided by the examiners. By coding and analyzing the segments of the text, four refined categories emerged as the factors that the examiners associate with the Grade 7 candidates: the motivation of taking GESE, the difficulty of the interactive task, improper training and cultural factors in the educational setting. Sub-categories are discussed and interpretations are provided below as well.

The coding map of the second theme is presented in Figure 3.

The three senior examiners agreed that in addition to the vocabulary use, there were various problems with the Grade 7 candidates of 2008. In fact, most Grade 7 candidates of 2008 did not really meet the requirements or the level of Grade 7.

Figure 6.3 Factors examiners associate with the result of Grade 7 candidates


Entering a key middle school with a GESE certificate
Motivation
Not ready
Result
for the level
Cultural factors in the educational setting
Difficulty of the task
Training
New exam method
Higher requirements
Lack understanding of GESE syllabus
Exam-oriented teaching

The examiners associated four factors with the result. According to the data collected for the present research, about 75% of Grade 7 candidates failed in the examination, which shows clearly that most of the Grade 7 candidates may have not been ready for the examination.

6.51 The motivation of taking Grade 7 GESE

One of the factors that contribute to the result is the candidate's motivation for taking GESE. According to the interview data, most Grade 7 candidates in 2008 wanted to pass the exam and get a certificate to enter a key school. Examiner A explained that:

> If a candidate has a Grade 7 certificate at that time (2008), he or she will be accepted by The Affiliated Middle School of Renmin University (one of the top middle schools in Beijing). The parents are also very keen on it. The children who wanted to enter a key middle school all take Grade 7.

What Examiner A said is also *ag*reed by other two examiners. Most GESE candidates in China were children or adolescents, and a GESE certificate, especially the certificate of an Elementary Stage (Grade 7-9) is a very important qualification that may help the candidate to enter a so-called key middle school with much better educational resources than ordinary middle schools. In such situations, many candidates prefer to take the first grade in each of the four stages instead of taking a higher grade after taking all the grades before that. For example, Grade 4 and Grade 7, which is the first grade of Elementary Stage and Intermediate Stage respectively, are the most frequently taken grades. After passing Grade 4, the first grade in the Elementary Stage, many candidates just take Grade 7 without taking Grade 5 or Grade 6. The potential benefit of a certificate of the Intermediate Stage and being admitted to a key middle-school attracted many candidates and their parents to try their luck. As a result, the candidates usually take a grade which is higher than their real proficiency level. That might also be one of the reasons why the pass rate of Grade 7 was only 25% in 2008. Examiner A described a candidate she met in the course of her examinations:

> I met a candidate that can be called the 'craziest' candidate I’ve ever had. The candidate got a D (Fail) for Grade 4, but he took Grade 7 very soon afterwards. How could such a candidate pass? I think the candidate and his parents were just trying their luck. Many Grade 7 candidates have taken a wrong grade.

This is not an extreme case. All the examiners mentioned that most Grade 7 candidate were not ready for the grade. They took the examination because it is useful, and they believed there was no harm in doing so even if they failed it. Examiner A remarked that, “They have nothing to lose if they fail the exam, but if they pass, they may benefit from it.”

6.52 The difficulty of the Interactive Phase for the candidates of Grade 7

All the examiners remarked that the Interactive Task Phase of the Intermediate Stage (Grade 7-9) was very challenging to Chinese candidates in 2008. Most candidates could not discuss anything with the examiner based on the prompts given, nor could they maintain the flow of communication by making comments or asking questions. According to the examiners, most candidates who failed the phase did not understand the examiner and did not produce much output.

Examiners A, B and C all stated that in 2008 many candidates looked confused after the examiner had given the prompts in the interactive phrase. There was very often a long pause after the prompts, and many candidates just tried to answer a question they themselves imagined and had prepared. As result, there were a lot of irrelevant responses from the candidates and the communication was very ineffective. There were more failures or breakdowns of communication in this phase than in other phases of the examination.

The difficulty of the Interactive Task Phase might be caused by three factors according to the interview:

First of all, the Interactive Task Phase is a new item to all the candidates and their teachers. The Interactive Phase was introduced into China in 2007 and it is very different from most oral English tests in China, in which the candidate is in a rather passive position and his or her responsibility is to answer questions. The Interactive Phase was very strange to most Chinese students and teachers, in which the candidate rather than the examiner has to take the initiative.

When the new Interactive task was introduced to China, it took a rather long time for the Chinese learners and training schools to get familiar with it and get used to it. Unfamiliarity added difficulty to the Interactive tasks.

It was difficult also because the candidates did not understand the syllabus very well. Examiner A commented on a Grade 7 Candidate in Examination 3: “It seemed that she didn’t understand what to do in this part (interactive tasks). She just wanted to express her own opinions about money. She didn’t know what she was supposed to do in this part. She didn’t know the requirements of Grade 7 at all.” As a result, the candidate did not perform well in the phase.

However, all the examiners agreed that the candidates perform better in more recent exams (2013) in the Interactive Phase at the Intermediate Stage. After more than 4 years, the Interactive task has become familiar to the Chinese learners and most candidates know that they should ask questions or make comments to keep the conversation going.

Second, higher requirements also contribute to the difficulty of the Interactive tasks.

In the Interactive Task Phase, the requirements on the communicative abilities of the candidate are much higher than those in the Conversation Phase. In the Conversation Phase, the candidates are expected to display their abilities to use the language of the grade, while in the Interactive Phase, the candidates are expected to take the responsibility to take control over and maintain the interaction while expressing the language function of the grade (GESE Syllabus from 2002, 2004, 2010). The candidates in the Interactive Phase have more tasks to perform and take a more active role than in the Conversation Phase.

In addition, the candidates know what topics there are in each grade in the Conversation Phase, and they can prepare and predict the questions that may be asked by the examiner on each topic. However in the Interactive Phase, the candidates are faced with prompts that are not known to them before the examination. Impromptu and unprepared interaction also adds difficulty to the Interactive tasks.

6.53 The improper training

Both examiners A and B mentioned that the problems of Grade 7 candidates were partly caused by the improper training given to them in 2007 and 2008. It was revealed from the interview that the problems of training mainly include the lack of comprehension of the GESE Syllabus and the exam-oriented teaching methods.

Firstly, as has been mentioned in the previous section, the Interactive Phase is a comparatively new GESE item, and even teachers in many training schools did not understand very clearly what the new interactive task was like, so they prepared the Interactive Phase according to their own understanding, which was still based on the traditional question-answer pattern. The young candidates of Grade 7 just prepared what the teachers had told them to do, and what seemed to be very obvious to the participant examiners, they did not understand the purposes of the Interactive Phase at all. After the second marking of the Grade 7 candidate, who did not know what to do in the Interactive Phase, Examiner A commented: "I wonder who trained the children like that! Did the teacher understand the exam at all?" The lack of understanding of the syllabus led to the poor performance of most Grade 7 candidates.

Secondly, another factor that contributes to the result is the exam-oriented teaching methods. The examiners mentioned that many students from the same training school seemed to have exactly the same background and same ideas for everything. Obviously it was the result of training. The examiners found that many candidates from the training schools just pay attention to the so-called *standard answer* to the mock examinations and exam *strategies* instead of real communicative abilities. Many candidates hold handouts of questions and answers when then enter the examination room and try to recite from them whenever possible. Examiner B described the candidates he met from the same training school:

> Most candidates were not quite there. They gave almost the same answer to my question as if it had been prepared. If they didn't understand my question; they would say, 'well, that is a good question. As a matter of fact…', then they continued to recite what they had prepared. Everybody did that and it drove me crazy.

It is common for a language training class which aims at an examination to be not communication-oriented. Alderson (2011) studied the backwash effects of TOEFL and found that the teaching style varies from teacher to teacher. It is found that some teachers did not teach communicatively in TOEFL-preparation classes in the United States, but the situation in China, according to the examiners, has gone to the extreme. More recently, the inequality in education resources has made the competition among students even fiercer and the situation even worse.

Zhang (2011) conducted a preliminary research on how Grade 6 candidates prepare GESE in another research. It was asked how the candidates prepared for

GESE when discussing the Grade 6 topic of *learning a foreign language.* All of the students mentioned reciting the prepared topics and memorizing the prepared answers to some questions given by their teacher or written with the help of the teacher or parents.

There is a long history of exam-oriented teaching and learning in China. This situation is still reported to be very common in China now, especially in English classes (Liu and Dai, 2003). The candidates aim to pass an examination in the shortest time possible and as a result, they are reluctant to spend time getting involved in communicative activities, which is time-consuming. However, during the exam, as Examiner B has remarked,

> Many Grade 7 candidates are eager to present to the examiner what they have prepared to leave a good impression on the examiner. When they were given the chance to maintain the conversation, they could not take control over the interactions and some of them would just wait for questions or recite what they had prepared.

The fact that the students only learn to take exams in training schools but ignore the real communicative abilities, or how to talk with people in real interactions might be one of the reasons they performed badly in the Interactive phase of Grade 7.

6.54 Cultural factors in the educational setting

Both Examiners A and C mentioned that cultural factors in the educational setting may also contribute to the result.

Examiner A explained that “Chinese students are reluctant to ask questions, either in class or in lectures”. She believed that the reason for this is that “the Chinese students are more self-conscious and they are afraid of asking questions. They are afraid of losing face if they can’t speak in a perfect way; they are afraid they will offend the teacher if they ask a wrong question.” Explaining interaction in cultural terms has been a common model (Zhu Hua, 2011) which has been criticized for creating and perpetuating static representations of learners. It is, however, widely accepted that different educational systems have different priorities or orientations.

In Chinese educational settings, the teacher is the authority and the students' task is to listen to the teacher and memorize what the teachers has said in class. A well-educated young student should show his or her respect to the teacher, who is the authority in the class by listening to them patiently and behaving politely both in and out of class. This rather normative representation of a Chinese class is still widely accepted. In addition, in the exam-oriented class, most students are considered as authority-dependent learners (Willing,1987; Gieve and Clark, 2005), whose learning style is rather passive and dependent on authority or teachers' instructions. With such a cultural background, Chinese students are not used to or encouraged to take initiatives during any conversation with their teachers. Asking questions and any expression of personal ideas in class, especially when the teacher is lecturing, is not accepted and is very often regarded as interrupting the teacher and wasting classmates' time, which is considered impolite and even rude behaviour.

Adding further to this, in oral interviews like GESE, the role of the examiner and the candidate is not equal at all. The examiner who takes control over the topics and the progress of the interview is much more powerful than the candidate. Even in GESE Grade 7 and above, although the candidate is expected to take control of the interaction and maintain the communication in the Interactive Phase, he or she is under the control of the examiner in other phases of the examination. Under such circumstances, it is very difficult for a young candidate to change his or her communicative style abruptly in one of the examination phases to lead the interaction by asking question or making comments.

Zhu Hua (2011) remarked that language is key to understanding culture, and culture is an essential part of studying language. Cultural factors in the educational setting may also help interpret the results of Grade 7 examinations. Although this has not been the focus of this work, the qualitative data open possible interpretations for further research.

6.6 The consistency of the rating

The second marking of the three examiners is very consistent with the original marking. In the following Table 6.2, the second marking and the original marks of the overall score of the studied phase are presented.

Table 6.2 The original and the second marking scores of the studied phases of 9 examinations

Number	*Examiner*	*Gender*	*Grade*	*Original*	*Second markin g*
1	A	F	Grade 2	B	B
2	A	F	Grade 5	C	C
3	A	F	Grade 7	C	C
4	A	F	Grade 7	D	D
5	B	M	Grade 5	C	C
6	B	M	Grade 7	C	C
7	C	F	Grade 2	C	C
8	C	F	Grade 5	C	C
9	C	F	Grade 7	***B***	***D***

Table 6.2 shows that the overall scores of the original marking and the second marking are consistent. All the second marking scores and the original ones are exactly the same except those for Examination 9. For Examination 9, the original score is B but the second marking is D. The interview with Examiner C provided some explanation of the difference between the second marking and the original marking.

When Examiner C was asked to do the second marking of the Interactive phase of Examination 9, she marked it D, and she explained that the candidate has the ability to talk about her own experiences, and her content and pronunciation is pretty good. But her problem is that although the examiner had given her every indication that it is the candidate's turn to ask questions, the candidate did not take the chance, or she did not catch the information to discuss with the examiner further about the issue. According to Examiner C, the candidate focused on her own ideas and did not pay much attention to what her interlocutor was talking about. Examiner C believed that this is not real interaction or communication.

When Examiner C was shown the original scores she awarded to the candidate, she explained some of the reasons that may account for the difference between the scores of the original marking and the second marking.

First, from her standard and understanding of the interactive phase now, the candidate's performance was not satisfactory. The candidate did not finish the interactive task satisfactorily. She did not take control of the interaction by asking questions or making comments although the examiner gave her several chances to do so.

Examiner C also mentioned an example to support her ideas. She was expecting the candidate to ask her a question by saying "I've been thinking about why I was so shy when I was a child", but the candidate took it as a question and spent a lot of time trying to answer it. After Examiner C failed to make the candidate ask questions and further discuss why she was so shy when she was young, she tried again by saying "I realized one of the reasons later on", but there was still no input of question or discussion from the candidate. Instead, the candidate still took this prompt as a question and answered "one of the reasons is that…", so Examiner C believed that after she saw the original scores, she still insisted that the candidate's performance in the interactive part was D.

Second, the reason why there is such a difference in marking may be related to the training and standardization of Trinity, London. According to Examiner C,

> Now we are using the new 2010 syllabus and Trinity is paying growing attention to the candidates' interactive abilities. The examiners are trained to speak less than five years ago. After each prompt, the examiners are trained to stop speaking and leave it to the candidate to catch the information and keep the conversation going.

Accordingly, there might be some change in marking standards. "I feel we are tougher now (than in 2008) in assessing the Interactive phase. It is greatly influenced by the examiner training and standardization. " Examiners A and B also mentioned that they feel the assessment of the Interactive phase might be tougher than in 2008, but it was not reflected in their second marking scores.

In summary, the second marking scores of the examinations conducted in 2012 and the original marking scores conducted in 2008 are highly consistent with each other except for one examination. For the nine examinations, eight second markings are the same as the original scores and the only discrepancy is between B and D for Examination 9. Although all the three examiners mentioned the change of assessment and performance of the candidates in more than four years, there is very little dramatic difference in the marking. The second marking and the original marking of the eight out of nine examinations are very consistent. It shows that the marking of the Chinese local examiners are very stable, and this may indicate that the GESE examinations conducted by these Chinese local examiners are highly reliable in this respect.

6.7 Conclusion

By coding and analyzing the qualitative data, themes and categories emerged from the data and some of the information that cannot be obtained from the quantitative research was collected. The results of the qualitative research not only provide interpretation of some quantitative results, but also provide useful information and shed light on the assessment of GESE, the assessment of vocabulary in GESE and problems with candidates and training in China.

It is found from the qualitative data that the Chinese GESE examiners adopted a holistic rating even when the analytical assessment system was applied. The second rating and the original rating are rather consistent.

The results of the qualitative data also indicate that relevant input is the most important factor that contributes to the holistic score of the candidate and it is also the economic rating strategy applied by Chinese examiners in vocabulary assessment. This is a result that was obtained from the qualitative data. The holistic rating of the vocabulary performance of a candidate is mainly related to whether a candidate can communicate with the examiner on the topics listed for a grade, and whether they can understand their interlocutor and give relevant responses during the interactions.

So the assessment of vocabulary is not isolated from the assessment of the candidate's overall performance. It is also associated with relevance and the success of communication. This result is also consistent with the quantitative result that the GESE assessment criterion of Focus is correlated with all the lexical variables and the criterion of *Readiness* is correlated with lexical variables of *Type* and *AG*.

It was found in the qualitative research that there is no significant difference between the lexical use of Grade 5 and Grade 7 candidates. The quantitative research provided little information as to what caused the result, while the qualitative data indicated several factors that led to the unexpected results: most Grade 7 candidates in 2008 chose the grade which was higher than their real level, attracted by the potential benefit of a GESE certificate; they were unfamiliar with the Interactive phrase and the training did not help them much in communicative abilities. The Interactive Task Phrase, as a new exam item then, was not easy for children and adolescents and Chinese cultural factors may also prevent them from taking initiatives during conversation with an examiner who represents authority and power.

Chapter 7　Conclusion

In this concluding chapter, the research results presented in Chapter 5 and Chapter 6 are summarized and synthesized. By revisiting the research questions, answers to the questions as well as interpretations are provided. Following that, the implications of the findings and contribution of the research to the field are also discussed. Finally, the limitations of the present research and suggestions for future work are also provided.

7.1 Conclusion of the quantitative results

7.11 Summary of the results of the first three research questions

The first three research questions are focused on candidate/student variables. The first research question is designed to compare the candidate/student lexical variables (*Type, Token, D, Giraud, AG and MLU*) of 3 different level of GESE.

It is found from the results presented in Chapter 5 that: 1) the lexical indexes of candidates go up as the level rises. All the Grade 5 lexical variables are higher than those of Grade 2, and all the lexical variables in Grade 7 but *AG* and *G* are higher than those in Grade 5. 2) All the differences between the Grade 2 and Grade 5 are statistically significant at the p=.05 level. All the differences between Grade 2 and Grade 7 lexical variables but that of *AG* are significant at the *p*=.05 level. 3) Concerning the differences between Grade 5 and Grade 7, there is no significant difference between Grade 5 and Grade 7 lexical variables in most cases.

It can be concluded that the all student/candidate lexical variables including *Types, Tokens, D, Giraud, AG* and *MLU* can differentiate between candidates of Grade 2 and Grade 5; all the student/candidate lexical variables can also differentiate between candidates of Grade 2 and Grade 7, but only the lexical variables of *AG* and *MLU* can differentiate between candidates of Grade 5 and Grade 7. The results seem to suggest that the lexical measures of Grade 7 are comparatively lower than expected.

The second question is designed to compare the candidate/student lexical variables (*Type, Token, D, Giraud, AG and MLU)* of the good performers and poor performers at the same grade of GESE.

It is found from the quantitative research that the qualified performers tend to have a higher index of *Type, Token, D, Giraud*, and *MLU* than those of the poor performers in all the three grades of Grade 2, 5 and 7. However, the lexical measures can differentiate between good performers and poor performers at the same grade at Grade 2 and Grade 5, but as the level of the candidates rise to Grade 7 (Intermediate Stage), only student/candidate/examinee *Type, D* and *AG* can differentiate between qualified performers and poor performers.

The results may indicate that the lexical measures performed differently at different levels. When all the lexical measures of *Type, Token, D, G, AG* and *MLU* can be both indicators of langu*ag*e proficiency level across the grades and within the same grade in Grade 2 and Grade 5 (the Initial Stage and Elementary Stage of GESE), *AG* is perhaps an effective indicator of langu*ag*e proficiency level across the grades in the Elementary and Intermediate Stage of GESE. T*ype, D and AG* might be effective elementary indicators of good performers and poor performers at the same grade in the Intermediate stage (Grades 7-9).

One of the major contributions of the research is that it proves that *AG* is the only measure that can not only differentiate between three proficiency levels of candidates, but also distinguish between qualified and bad performers of the same grade, which supports the argument proposed by many researchers (Laufer and Nation 1995;Wesche & Paribakht 1996; Vermeer 2000; Wen 1999; Daller, van Hout and Treffers-Daller 2003) that a more effective measure of lexical richness may involve lexical sophistication or frequency of words, and *AG* is such a measure in the present research. Further research on *AG* may promote our understanding of the global indicator of lexical richness and help revisit and refine the existing tools of vocabulary assessment when there is no single perfect measure in use (Laufer, 2005).

The third research Question (*Will student lexical richness measures correlated with student variables obtained from GESE scores?*) is focused on the relationship between student/candidate lexical variables and GESE score variables.

It is found from the results that: firstly, except for *AG*, all the student/candidate lexical variables of *Type, Token, D,G* and *MLU* are highly or moderately correlated with each other. *AG* has a slight correlation with *Type, Token, D* and *G*, but has no relationship with *MLU*. Secondly, all the GESE score variables in the pooled data are highly correlated with each other, and the correlation efficient is significant at the .001 level. There is a heavy halo effect.

The separated data show a similar situation: all the GESE score variables are highly correlated with each other in Grade 2 and Grade 5, while the correlation among the variables in Grade 7 is significant but not high. It can be indicated from the difference that in lower stages (such as Grade 2 and Grade 5) the halo effect of rating is very great, but as the grade goes up to Intermediate level, the halo effect of rating is not as heavy as that in the Initial and Elementary Stages. Thirdly, with regard to the relations between student/candidate lexical variables and GESE score variables, in the pooled data, the only GESE score variable that is correlated with all the lexical variables is Focus. As to the relationship between lexical variables and GESE score variable of *Usage* (vocabulary), the separated data show a clearer picture. All the lexical variables are significantly correlated with the GESE score of *Usage* (vocabulary) in Grade 2, with the highest correlation with D (*rho*=.801 *p*=.001) and the lowest with *AG* (*rho*=.290 *p*=.05). All the lexical variables are significantly correlated with the GESE score of *Usage* (vocabulary) in Grade 5, the highest correlation is with *Type* (*rho*=.591 *p*=.001) and the lowest with *D* (*rho*=.262 *p*=.001). However, in Grade 7 only *Type* and *D* correlated with *Usage* and there is only a weak correlation.

The results point very important values in understanding how Chinese local examiners of GESE rate vocabulary and the GESE examinations. The results may indicate that the Chinese examiners' rating of GESE in a rather holistic way although analytical assessment system was applied in 2008. It also can be inferred that the examiners' rating of vocabulary is affected by the sophistication of words. However, they did not rate vocabulary only based on the vocabulary use of candidates, they put more focus on the communicative abilities of the candidates. So it is assumed that lexical sophistication and relevant and meaningful input might be economical strategies or the reliable overall rating of vocabulary (Daller and Phelan, 2007) applied by Chinese GESE examiners. The results were confirmed by the results of the qualitative research.

7.12 Summary of the results of the research questions 4-5

Research questions 4 to 5 are focused on the examiners' lexical accommodation with candidates.

Research Question 4 is mainly designed to investigate whether or not the GESE examiners accommodate lexically to candidates of three different grades of GESE.

The results of the pooled data show the following:

1) Concerning the teacher/examiner variables, there is also a trend that the teachers use higher indexes of lexical measures with candidates of higher Grade. The higher the candidate's grade, the higher the index of examiner lexical measures and *MLU*. Teacher/examiner *Token* is an exception: the lowest grade tends to have highest Teacher/examiner *Token*.

2) The teacher/examiner lexical variables and *MLU* are significantly correlated with student/candidate/examinee lexical variables and *MLU* in the pooled data. The only negative correlation is between candidate/examinee and *Tokens* and teacher/examiner *Token*. However, such a correlation is absent in the separated data of each grade.

The results of Research Question 4 seem to indicate that all the teacher/examiner variables can differentiate between candidates of different grades; however, they cannot differentiate between qualified performers and poor performers at the same GESE grade.

The last research question concerns the accommodation the GESE examiners practise with the good and poor performers of the same grade of GESE.

The results indicate that 1) the teacher/examiner uses higher indexes of lexical variables to the qualified performers than to the poor performers, but the mean differences cannot differentiate between qualified performers and poor performers of the same grade. 2) The teacher/examiner lexical variables are correlated with student/candidate lexical variables in the pooled data. 3) The teacher/examiner lexical variables, however, are not correlated with student/candidate lexical variables in the separated data of each grade. 4) The lexical variables that show the positive correlation between teacher/examiner variables and candidate variables are *AG* and *MLU*. 5). There is very scarce relationship between teacher/examiner lexical variables and the GESE score of *Usage* (vocabulary).

It can be inferred from the results of Research questions 4 and 5 that the teachers/examiners accommodate to the level of the candidates of a certain grade as a whole, but not to individuals in the grade. They tend to use more varied and more difficult words to candidates of higher grades, but at the same grade, they don't change their use of vocabulary greatly. The reasons for such a phenomenon might be first, the GESE examiners are well-trained and there are a lot of prepared utterances in the conversation - for example the simple questions and directions in Grade 2, the utterances marking the transition of the phase in Grade 5 and the prompts of Grade 7- which are rather fixed patterns. Another reason is for the sake of reliability. The examiner who is talking with the candidate of a certain grade usually controls his or her language to an appropriate level. The fact that the examiners are not finely tuned to the candidates were also observed by Ross and Berwick (1992) and Malvern and Richards (2002). Although the lack of accommodation might be caused by the requirement of equality or fairness of public examinations, it may "introduce threat to validity" of the examination. The results also point out the practical problem in oral examinations. How to accommodate with candidates and to what extent should examiners accommodate with candidates need further exploration.

In conclusion, the lexical variables of students generally go up as the grade rises. All the lexical variables can differentiate between qualified performers and poor performers of Grade 2 candidates and Grade 5 candidates. Only *Type, D and AG* can differentiate between qualified performers and poor performers of Grade 7 candidates. All the student/candidate GESE score variables are highly correlated with each other, which show a halo effect of the rating. The relationship of the student/candidate lexical variables and GESE score variables are not straightforward.

Concerning the teacher/examiner variables, generally speaking they can also differentiate between candidates of three different grades, yet they cannot differentiate between qualified performers and poor performers of the same grade. There is a trend for the teachers to use higher indexes of lexical measures with candidates of higher grade and the good performers of a certain grade. It is also found that the teacher/examiner lexical variables are correlated with student/candidate lexical variables in the pooled data. However, such correlations are absent in the separated data of each grade. The lexical variables that show the correlation between teacher/examiner variables and student/candidate variables are *AG* and *MLU*. There is very scarce relationship between teacher/examiner lexical variables and the GESE score of *Usage* (vocabulary). The results may indicate that the teacher/examiners accommodate lexically to candidates of a certain grade as a whole, but they do not accommodate to different performers of the same grade. In other words, they do not adjust their use of vocabulary to individuals at the same grade.

7.2 Summary of the qualitative results

It is found from the second marking of the 9 GESE examinations that the scores of the second marking are highly correlated with the original scores. It is also found from the interviews that the examiners adopted a holistic rating method even when the analytical rating system was required. That reflected the reality of GESE assessment in 2008 and also provided evidence supporting the change in the GESE assessment system that started in 2010. A holistic rating system based on both intuition and assessment criteria was adopted by the examiners. With regard to assessment criteria, *Focus* and *Readiness* are stressed by the examiners.

When the three GESE examiners reflected on how they assessed vocabulary, they expressed the idea that they usually do not just look at the lexical performance of the candidate - for example, whether they apply a large size of vocabulary or difficult words - they also look at others aspects such as whether the student can communicate with the examiner successfully, whether the candidate can give relevant responses to the question or whether they can get involved in meaningful communication. The three examiners all hold the opinion that whether a candidate can give meaningful and relevant input during interactions is the basic category of vocabulary assessment. This may suggest that according to the Chinese local examiners of GESE, first, the meaningful and relevant input might be the *economical vocabulary assessment strategy* and second, the assessment of vocabulary is also related to general communicative abilities. Both the quantitative and the qualitative results suggest the assessment of vocabulary in GESE is not related to pronunciation at all. A similar result was also revealed in the pilot study (Chapter 3) of the present research. This may to some extent show the validity of GESE conducted by the Chinese local examiners.

The qualitative research also provides possible answers to the question of why there is no significant difference between the lexical measures of Grade 5 and Grade 7 candidates. The interviews with the GESE examiners indicate that many reasons have brought about the situation and the candidates of Grade 7 performed worse than expected. The factors that may contribute to the result are as follows: 1) Grade 7 candidates usually chose a grade that is higher than their real level; 2) candidates' unfamiliarity with the Interactive phrase in 2008; 3) a lack of understanding of the GESE syllabus; 4) the higher requirements of the Interactive phrase than the Conversation Phase; 4) the training methods in China and 5) Chinese cultural influence in educational settings may all contribute to the unexpected results shown by the quantitative research.

7.3 Implications of the present research

The present research may have implications and contributions in the areas of vocabulary assessment and international oral English testing in overseas settings. This section mainly discusses the implications of the present research in the theoretical, methodological and practical perspectives.

First, it is found from the results that the lexical measures function differently with the oral production of learners of different proficiency levels. In the Initial and

Intermediate St*ag*e of GESE, all the lexical measures can differentiate between oral productions of candidates of different grades, but as the level rises to Elementary St*ag*e (Grade 7), they are not as effective as in the lower st*ag*e. Fewer lexical measures can distinguish between oral productions of candidates of the same grade. *AG* is outstanding among the lexical measures because it may distinguish between candidates with different grades and distinguish between good and poor performers at the same grade. It is also the only lexical measure that reflects examiner lexical accommodation in Grade 2 and Grade 5. In the setting of GESE, *D* does not perform as well as *AG* as an effective lexical indicator of vocabulary usage. This may provide insights into a better understanding of the construct of the widely accepted measure of lexical diversity. It is also found that the lexical measure of *AG*, which combines lexical diversity and lexical sophistication or difficulty, performs better than other measures in capturing the subtle difference between good and poor performers of the same grade. The rather unexpected results may enhance our understanding of both the construct of vocabulary knowledge and lexical richness. We may expect some more effective global indicators of lexical knowledge that those based on *AG* or other similar measures, which may have captured more aspect of lexical richness. This effort is really “a legitimate and useful scholarly activity” (Laufer 2005, P.587).

Second, the research results indicate that there is no very direct relationship between the lexical variables and examiner lexical accommodation, which is very similar to the research results of Malvern and Richards (2000), Richards and Malvern (2002) and Lorenzo-Dus and Meara (2005). All the research suggested that the relationship between candidates’ use of vocabulary and examiner accommodation is not simple and straightforward. However, this research may also indicate that for different grades of GESE examinations and candidates, the examiners are found to show certain accommodation in certain aspects. The present research provides more evidence to the question under investigation. The results also require that more studies are necessary to explore the relationship before accommodation is taken into assessment criteria as Ross (1992) proposed.

Thirdly, in the methodological perspective, the present research applied random-sample data from a large corpus to investigate lexical measures and lexical accommodation in oral proficiency interviews. It has the advantages over previous research, which mainly used small data collected on the basis of availability. 180 transcriptions of the GESE examinations and nine interviews with Chinese examiners of GESE in 2008 were involved in the research. The data set is considered as a large scale one by comparison to others in the field. In addition, quantitative research and qualitative research are combined together to present more complete and reasonable results in the field of language testing and vocabulary research. Interviews with the examiners provided additional information that quantitative data could not explain.

Finally, the results may help examiners of non-native speakers of English become more conscious of their accommodation strategies. According to the results, although Chinese examiners of GESE accommodate to candidates of different grades at the lexical level, they did not adjust their speech to the candidates of a certain grade on the level of vocabulary. This may shed light on the administration of both GESE and other oral examinations. Based on the nature of the examination they are conducting, the examiners of oral examinations may need to learn to adopt appropriate accommodation strategies to candidates of different levels. This practical implication of the research also raises the issue of examiner training. It is found from the interview with GESE examiners that examiner training has a great influence on the performance and assessment of examiners, so all the results of the research may shed light on examiner training, and the training, in return, will have great effects on the way the examiner talks to the candidates. How to accommodate to candidates of different levels and to what extent should examiners accommodate to candidates might be standardized through examiner training.

7.4 Limitations of the present research

The present research has investigated the characteristics of both the examiners' and the candidates' lexical use in Grade 2, Grade 5 and Grade 7 GESE examinations as well as the relationship between the examiner lexical variables, candidate lexical variables and GESE score variables of the candidates. It is found that the lexical variables have their own characteristics in different grades, so the results obtained from the present study may not apply to other grades of GESE or other oral proficiency interviews for English.

This study has limitations: first, only investigates the examinations conducted by Chinese local examiners of GESE. Examinations conducted by their British peers were not collected due to the fact that the examinations conducted by native speaker examiners were not recorded and there was no data on the British examiners in the corpus. The contrast between the lexical variables and lexical accommodation of the native speakers and non-native speakers examiners would be of interest and shed more light on the GESE rater reliability. There is not much research on the differences between the native speakers and non-native speaker examiners in accommodation and assessment of vocabulary in the context of the same international oral English examination. Future research on analyses of both British and Chinese examiners may not only help us have a better understanding of the effects of an examiner's lexical use and accommodation on candidates and assessment but also help improve the validity and reliability of GESE conducted by both Chinese and British examiners.

Another limitation of the present research is that only six lexical variables are applied for the study. The six variables represent the most important aspects of lexical richness: lexical diversity and lexical sophistication. In addition to the lexical variables, *MLU*, an indicator of the general langu*ag*e development is also applied. However, there are still other lexical richness measures that were not applied in this research due to the limit of time and energy. Most of the measures applied in the present research were obtained by using *CLAN* software of *CHILDES.* There newly appeared some new measures that cannot be obtained from CLAN such as the lexical diversity measure of *MTLD* and *H-DD* proposed by McCarthy and Jarvis (2010). The application of more lexical measures may provide insights into further understanding of the construct of vocabulary knowledge and the features of different measures of lexical richness. Researchers generally agree that there is no single best measure of lexical richness and different measures may furnish us with different information of the learner's vocabulary use. So the use of more lexical measures would provide richer information of the learner's vocabulary use. It is suggested that more lexical variables are compared and investigated in future research to enhance our understanding of the characteristics and their effects of different measures of lexical richness.

Regarding the lack of a significant difference between Grade 5 and Grade 7 candidate lexical variables and the reasons why Grade 7 candidates did not perform as well as expected, two factors may have functioned in the process of research. Firstly, the problems with Grade 7 candidates and training methods, another factor that might be the effect of different task types in the examination. The data chosen from Grade 5 is the Conversation Phase while the data chosen from Grade 7 is Interactive Phase. It was stated earlier that the reason why different phases of Grade 5 and Grade 7 were chosen was to investigate whether task type has influence on lexical richness measures. The results of the quantitative and qualitative data seem to suggest that different examination items or different task types in oral examinations might have some influence on the candidates' use of vocabulary. However, which factor places a more important role is not clear. In future research, studies on the effects of different task types of the same grade as well as the same task type for different grades of candidates should be carried out, so that the effects of task types would be investigated from different dimensions. For example, the lexical variables of Grade 7 candidates in both Interactive phase and Conversation phase are investigated, and at the mean time the lexical variables of both Grade 7 and Grade 5 candidates in the same phase, the Conversation phase in this case are also compared. It might help us find out if there is a task type effect that may influence the lexical use of the candidate or not.

Another factor that might explain the unexpected outcome of the comparatively poor use of Grade 7 candidates is that the proficiency level of many Grade 7 candidates in 2008 did not meet the high standards for Grade 7. Although some of the lexical variables of Grade 7 candidates are a little higher than those of Grade 5 candidates, there is no significant difference between them. In future research, a better designed quantitative research and qualitative research might result in more convincing answers to the question. In addition, if some candidates and more Chinese local examiners of GESE are involved in the interview, we can also get a more complete picture of the situation.

Finally, *AG* is found to be unique in that it can distinguish the qualified candidates from the poor candidates of the same GESE grade both in the pooled data and in the separate data. In addition, it does not correlate highly with other lexical measures, which distinguished itself from other lexical richness measures. It might be an effective indicator of language proficiency than other lexical measures in the setting of GESE conducted by Chinese examiners. However, *AG* has not been fully studied in the present research because the validity of *AG* is not the main research question. In future research, the reliability and validity of *AG* need to be further explored with more data.

In conclusion, the present research has focused on the lexical richness and lexical accommodation in oral English examination of GESE. There is still a lot to explore and investigate in this field. More future research will definitely take our understanding of oral testing as well as vocabulary assessment and accommodation a step forward.

References:

Albrechtsen, D., Haastrup, K., and Henriksen, B. (2008) *Vocabulary and Writing in a First and Second Language: Process and Development.* Basingstoke: Palgrave Macmillan.

Alderson, J.C. (2011) *A Lifetime of Language Testing*. Shanghai: Shanghai Foreign Language Education Press.

Arlman-Rupp, A.J., Van Niekirk-de Hahn, D. and Van de Sandt-Koenderman, M. (1976) Brown's early stages: Some evidence from Dutch. *Journal of Child Language*. 3, pp.267-274.

Arnaud, P. (1992) Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate component tests. In Arnaud, P. and Bejoint, H. (eds.) *Vocabulary and Linguistics*. London: Macmillan, pp. 133-145.

Attride-Sterling, J. (2001) Thematic networks: an analytic tool for qualitative research. *Qualitative Research.* 1(3), pp.385-405.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bauer, L. and Nation, I.S.P. (1993) Word families. *International Journal of Lexicography.* 6(4), pp.253-279.

Boves, T.L.L. (1992) *Speech Accommodation in Co-operative and Competitive Conversations.* Nijmegen: KUN.

Brown, A. (2003) Examiner support strategies and test-taker vocabulary. *International Review of Applied Linguistics in Language Teaching.* 43, pp.239-258.

Brown, C. (1993) Factors affecting the acquisition of vocabulary: frequency and saliency of word. In Huckin, T., Haynes, M. and Coady, J. (eds.) *Second Language Reading and Vocabulary Learning.* Norwood, NJ: Ablex, pp.263-286.

Brown, R. (1973) *A First Language: The Early Stages.* Cambridge, MA: Harvard University Press.

Carter, R. (1998) *Vocabulary: Applied Linguistic Perspectives.* 2nd ed. London: Routledge.

Carter, R. and McCarthy, M. (1988) *Vocabulary and Language Teaching.* London: Longman.

Chomsky, N. (1986) *Knowledge of Language: Its Nature, Origin, and Use.* New York: Praeger.

Coady, J. and Huckin, T. (eds.) (2001) *Second Language Vocabulary Acquisition.* Shanghai: Shanghai Foreign Language Education Press.

Cohen, J. (1988) *Statistical Power Analysis for Behavioural Sciences.* 2nd ed. Hillsdale, NJ: Erlbaum.

Cook, V. (2000) *Linguistics and Second Language Acquisition*. Beijing: Foreign Language Teaching and Research Press.

Coupland, N. (1984) Accommodation at work: some phonological data and their implications. *International Journal of the Sociology of Language*. 46, pp.49-70.

Coupland, N., Coupland, J., and Giles, H. (1991) *Language, Society and the Elderly*. Oxford: Blackwell.

Daller, H., van Hout, R. and Treffers-Daller, J. (2003) Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistic.* 24, 197-222.

Daller, H. and Phelan, D. (2007) What is in a teacher's mind? The relation between teacher ratings of EFL essays and different aspects of lexical richness. In Daller, H., Milton, J. and Treffers-Daller, J. (eds.) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

Daller, H. and Xue, H. (2007) Lexical richness and the oral proficiency of Chinese EFL students – A comparison of different measures. In Daller, H., Milton, J. and Treffers-Daller, J. (eds.) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

Daller, H., Milton, J. and Treffers-Daller, J. (eds.) (2007) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

Duran, P., Malvern, D., Richards, B., and Chipere, N. (2004) Developmental trends in lexical diversity. *Applied Linguistics.* 25, pp.220- 242.

David, A. (2008) A developmental perspective on productive lexical knowledge in L2 oral interlanguage. In Treffers-Daller, J., Daller, H.M., Malvern, D., Richards, B. , Meara, P and Milton, J. (eds.) *Special Issue of French Language Studies.* 18, pp.269-276.

Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T., and O'Loughlin, K. (eds.) (2001) *Experimenting with Uncertainty: Essays in Honour of Alan Davies.* Cambridge: Cambridge University Press.

Ellis, R. (1985) *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (1994) *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (2002) Does form-focused instruction affect the acquisition of implicit knowledge? A review of the research. *Studies in Second Language Acquisition*. 24, pp.223–236.

Ellis, R. (2005) Principles of instructed language learning. *System.* 33, pp.209-224.

Gass, S. (1997) *Input, Interaction, and the Second Language Learner.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S.M. and Selinker, L. (2008) *Second Language Acquisition: An Introductory Course*. 3rd ed. Routledge: New York and London.

Gass, S.M. and Varonis, E.M. (1985) Variation in native speaker speech modification to non-native speakers. *Studies in Second Language Acquisition*. 7, pp.37-57.

Gass, S. M., & Varonis, E. M. (1994). Conversation in interactions and the development of L2 grammar. *Studies in Second Language Acquisition*. 16(3), 283-302.

Giles, H. (1973) Accent mobility: a model and some data. *Anthropological*

Linguistics. 15, pp.87-105.

Giles, H., Tayor, D.M. and Bourhis, R.Y. (1973) Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society*. 2, pp.177-192.

Giles, H. and Coupland, N. (1991) *Language: Contexts and Consequences*. Buckingham: Open University Press.

Giles, H. and Powesland, P. (1975) *Speech Style and Social Evaluation*. Cambridge: Cambridge University Press.

Giles, H. and Smith, P.M. (1979) Accommodation theory: optimal levels of convergence. In Giles, H. and St. Clair, R.N. (eds.) *Language and Social Psychology.* Oxford: Blackwell.

Gieve, S. and Clark, R. (2005) The Chinese approach to learning: cultural trait or situated response? The case of a self-directed learning programme. System. 33(2), pp.261-276.

Gregory, S. and Webster, S. (1996) A nonverbal signals in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology.*70, pp.1231-1240.

Hazenberg,S and Hulstijn, J.H. (1996) Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics.* 17, pp.145-163.

He, A.W. and Young, R. (1998) Language proficiency interviews: a discourse approach. In Young, R. and He, A.W. (eds.) *Talking and testing: discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins, pp.1–24.

Henriksen, B. (1999) Three dimensions of vocabulary development. *Studies in Second Language Acquisition.* 21, pp.303-317.

Hickey, T. (1991) Mean length of utterance and the acquisition of Irish. *Journal of Child Language*. 3, pp. 553-569.

Issidorides, D. and Hulstijn, Y. (1992) Comprehension of grammatically modified and nonmodified sentences by second language learners. *Applied Psycholinguistics*. 13, pp.147-171.

Jarvis, S. (2002) Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*. 19, pp.57-84.

Krashen, S. (1985) *The Input Hypothesis: Issues and Implications.* Longman, New York.

Krashen, S. (1989) We acquire vocabulary and spelling by reading: Additional evidence for input hypothesis. *The Modern Language Journal*. 73, pp.440-464.

Labov, W. (1972) *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Larsen-Freeman, D and Long. M.H. (2000) *An Introduction to Second Language Acquisition Research*. Beijing: Foreign Language Teaching and Research Press.

Laufer, B. (1995) Beyond 2000. A measure of productive lexicon in a second language. In Eubank, L., Selinker, L., and Sharwood-Smith, M. (eds.) *The*

Current State of Interlanguage. Studies in Honour of William E. Rutherford. Amsterdam: John Benjamins.

Laufer, B. (2001) Quantitative evaluation of vocabulary: How it can be done and what it is good for. In Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T., and O'Loughlin, K. (eds.) *Experimenting with Uncertainty: Essays in Honour of Alan Davies.* Cambridge: Cambridge University Press.

Laufer, B. (2003) Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *The Canadian Modern Language Review.* 59, pp.567-587.

Laufer, B. (2003) Lexical frequency profile: from Monte Carlo to the real world a response to Meara (2005). *Applied Linguistics*. 26(4), pp.582-588.

Laufer, B. and Z. Goldstein (2004). "Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning 54*(3): 399-436.

Laufer, B. and Nation, I.S.P. (1995) Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics.* 16, 302-322.

Laufer, B. and Nation, I.S.P. (1999) A vocabulary-size test of controlled productive ability. *Language Testing*.16, pp.36-55.

Lazaraton, A. (1996) Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing*.13, pp.151-172.

Lazaraton, A. (2002) *A Qualitative Approach to the Validation of Oral Language Tests.* Cambridge: Cambridge University Press.

Lorenzo-Dus, N. and Meara, P. (2005) Examiner support strategies and test-taker vocabulary. *IRAL.* 43, pp. 239-58

Linnarud, M. (1986). *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English.* Malmo: Liber Forlag.

Liu, R. and Dai, M. (2003) *Foreign Language Teaching Reform in Chinese Universities: a Study on the Status Quo and Development Strategies*. Beijing: Foreign Language Teaching and Research Press.

Long, M.H. (1983) Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*. 4, pp.126-141.

Long, M.H. (1991) Focus on form: A design feature in language teaching methodology. In de Bot, K.,Ginsberg, R and Kramsch, C. (eds.) *Foreign Language Research in Cross-Cultural Perspective.* Amsterdam: John Benjamins, pp.39-52.

Long, M.H. (1996) The role of the linguistic environment in second language acquisition. In W.C. Ritche and T.k. Bahtia (eds.) *Handbook of Second Language Acquisition*. New York: Academic Press, pp.413-468.

Long, M. H. and Richards, J.C. (2007) Series Editors' preface. In Daller, H., Milton, J. and Treffers-Daller, J. (eds.) Modelling and Assessing Vocabulary Knowledge. Cambridge: Cambridge University Press.

Loschky, L. (1994) Comprehensible input and second language acquisition: what is the relationship? *Studies in Second Language Acquisition*.16, pp.303-323.

Lu, X. (2011) The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*.11, pp.1-19.

MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk.* 3rd ed. Vol. 1: Transcription Format and Program. Mahwah, NJ: Erlbaum.

MacWhinney, B. and Snow, C.E. (1990) The Child Language Data Exchange System: an update. *Journal of Child Language*.17, pp. 457-472.

Malvern, D. and Richards, B. (1997) A new measure of lexical diversity. In Ryan, A. and Wray, A. (eds.) Evolving Models of Language. *Papers from the Annual Meeting of the BAAL.* Clevedon: Multilingual Matters, pp.58-71.

Malvern, D. and Richards, B. (2002) Investigating accommodation in language proficiency interviews using a new measure of lexical richness. *Language Testing*.19, pp.85-104.

Malvern, D., Richards, B., Chipere, N., and Duran, P. (2004) *Lexical Diversity and Language Development: Quantification and Assessment.* Houndmills: Palgrave Macmillan.

McCarthy, P.M. and Jarvis, S. (2007) vocd: a theoretical and empirical evaluation. *Language Testing*. 24(4), pp.459-488.

McCarthy, P.M. and Jarvis, S. (2010) MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to diversity assessment. *Behaviour Research Methods*. 42(2). pp.381-392.

Meara, P. (2002) The rediscovery of vocabulary. *Second Language Research.*18(4), pp.393-407

Meara, P. (2005) Lexical frequency profile: A Monte Carlo analysis. *Applied Linguistics*. 26, pp.32-47.

Meara, P. (1996) The dimensions of lexical competence. In Brown, G., Malmkjaer, K. and Williams, J. (eds.) *Performance and Competence in Second Language Acquisition*. Cambridge: Cambridge University Press.

Milton, J. (2008) French vocabulary breadth among learners in the British school and university system: comparing knowledge over time. In Treffers-Daller, J., Daller, H.M., Malvern, D., Richards, B., Meara, P and Milton, J. (eds.) *Special Issue of French Language Studies.* 18, pp.333-348.

Milton, J. (2009) *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters Limited.

Mitchell, R. and Myles, F. (2004) *Second Language Learning Theories.* 2nd ed. London: Hodder Arnold.

Meara, P. and Bell, H. (2001) P-Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*. 16, pp. 5-19.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral test. *Language Testing*. 28(4), pp.483–508.

Nation, I.S.P. (1990) *Teaching and learning vocabulary*. New York, NY: Heinle and Heinle.

Nation, I.S.P. (2001) *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, I.S.P. and Wang Ming-tzu, K. (1999) Graded readers and vocabulary. *Reading in a Foreign Language.* 12, pp.355-379.

Nation, I.S.P. and Waring, R. (1997) Vocabulary size, text coverage and word lists.

In Schmitt, N and McCarthy, M (eds.), *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press.

Nation, P. (2007) The four strands. *Innovation in Language Learning and Teaching.* 1(1), pp. 2-13.

O'Loughlin, K. (2002) The impact of gender in oral proficiency testing. *Language Testing*. 19(2), pp.169-192.

Parker, and Brorson (2005) A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language.* 2 (3), pp.365-376.

Parker, K. and Chaudron, C. (1987) The effects of linguistic simplification and elaborative modification on L2 comprehension. *University of Hawaii Working Papers in ESL.* 6 (2), pp.107-133.

Pica, T. (1994) Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning.* 44*,* pp.493–527.

Pica, T., Holliday, L., Lewis, N., Berducci, D and Newman, J. (1991) Language learning through interaction: what role does gender play? *Studies in Second Language Acquisition.* 13, pp.343-376.

Pica, T. (1987) Second-language acquisition, social interaction, and the classroom. *Applied Linguistics*. 8, pp.3-21.

Polio, C. G.. (1997) Measures of linguistic accuracy in second language writing research. *Language Testing.* 47, pp.101-143.

Read, J. (1997) Vocabulary and testing. In Schmitt, N and McCarthy, M (eds.) *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press. pp. 303-320.

Read, J. (2000) *Assessing Vocabulary.* Cambridge: Cambridge University Press.

Read, J. and Chapelle, C.A. (2001) A framework for second language vocabulary assessment. *Language Testing.*18, pp.1-32.

Richards, B.J. and Chambers, F. (1996) Reliability and validity in the GCSE oral examination. *Language Learning Journal*. 14, pp. 28-34.

Richards, B.J. and Malvern, D.D. (2000) Accommodation in oral interviews between foreign language learners and teachers who are not native speakers. *Studia Linguistic.* 54, pp. 260-271.

Richards, B., Daller, M.H., Malvern, D.D., Meara, P., Milton, J. & Treffers-Daller, J. (eds.) (2009) *Vocabulary Studies in First and Second Language Acquisition: the Interface between Theory and Application*. Cambridge: Cambridge University Press.

Richards, J.C. (1976) The role of vocabulary testing. *TESOL Quarterly.* 10, pp.77-89.

Ross, S. (1992) Accommodative questions in oral proficiency interviews. *Language Testing.* 9, pp. 173-186.

Ross, S. and Berwick, R. (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*.14, pp.159-176.

Selinger, H. and Shohamy, E. (1997) *Second Language Research Methods.*

Shanghai :Shanghai Foreign Teaching Press.

Schmidt, R. (1990) The role of consciousness in second language learning. *Applied Linguistics.* 11, pp. 129-158.

Schmidt, R. (1994) Deconstructing Consciousness in Search of Useful Definitions for Applied Linguistics. *AILA Review.* 11, pp. 11-26.

Schmidt, R. 2001: Attention. In Robinson, P. (ed.) *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Schmitt, N. (2000) *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schmitt, N and McCarthy, M. (1997) *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press.

Schoonen, R. (2001) Book Review: Assessing Vocabulary. *Language Testing.*18, pp.18-125.

Singleton, D. (1999) *Exploring the Second Mental Lexicon.* Cambridge: Cambridge University Press.

Swain, M. (1985) Communicative competence: some roles of comprehensible input and comprehensive output in its development. In Gass, S. and Madden, C. (eds.) *Input in Second Language Acquisition.* Rowley, MA: Newbury House, pp.235-253.

Swain, M. (1995) Three functions of output in second language learning. In Cook, G. & Seidlhofer, B. (eds.) *Principle and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson.* Oxford : Oxford University Press.

Talamas, A., Kroll, J.F., and Dufour, R. (1999) From form to meaning: Stages in the acquisition of second language vocabulary. *Bilingualism: Language and Cognition.* 2, pp.45-58.

Thakerar, J., Giles, H. and Cheshire, J. (1982) Psychological and Linguistic parameters of speech accommodation theory. In Fraser, C and Scherer, K. (eds.) *Psychological Dimensions of Language Behaviour.* Cambridge: Cambridge University Press.

Trinity College London. (2002) *Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages, from 2002*.

Trinity College London. (2004) *Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages, 2004-2007*.

Trinity College London. (2010) *Trinity's International Syllabus for Graded Examinations in Spoken English for Speakers of Other Languages, from 2010*.

Underhill, N. (1987) *Testing Spoken Language: a Handbook of Oral Testing Technique.* Cambridge: Cambridge University Press.

Van Patten, B. (1996) *Input Processing and Grammar Instruction.* New Jersey: Ablex.

Van Patten, B. (2002) Processing instruction: An update. *Language Learning.* 52(4), pp. 755–803.

Vermeer, A. (2000) Coming to grips with lexical richness in spontaneous speech data. *Language Testing.* 17, pp.65-83.

Wen, Q. (1999) *Oral English Testing and Teaching*. Shanghai: Shanghai Foreign

Languages Education Press.

Wesche, M. & Paribakht, T.S. (1996) Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review.* 53, pp.13-40.

Willing, K. (1987) *Learning Style in Adult Migrant Education.* Adelaide: Adult Migrant Education.

Xue, G and Nation, I.S.P. (1984) A university word list. *Language Learning and Communication.* 3, pp. 215-219.

Yano, Y., Long, M.H. and Ross, S. (1994) The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning.* 44(2), pp.189-219.

Young, R. and He, A.W. (eds.) (1998) *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam: John Benjamins.

Young, R. and Milanovic, M. (1992) Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*. 14, pp.403-404.

Yu, G. (2009) Lexical diversity in writing and speaking task performances. *Applied Linguistics*. 31(2), pp.236-259.

Zhu Hua. (2010) Language socialization and interculturality: address terms in intergenerational talk in Chinese diasporic families. *Language and Interculture Communication*. 10(3). pp.189-205.

Zhu Hua. (2011) *The Language and Intercultural Communication Reader*. London: Routledge.

Zhang, J. (2012) "Ask you a question?": Analysis of Chinese candidates' questions in GESE interactions. The Annual Bloomsbury Round Table on Communication, Cognition and Culture, London.

Appendix 1 The questionnaire of the pilot study

How Do You Assess Vocabulary in GESE?

1. The higher the examinee's over-all language proficiency, the higher the score of vocabulary.

(strongly agree) 5 4 3 2 1 (strongly disagree)

2. I mark vocabulary according to specific rules derived from assessment categories.

(strongly agree) 5 4 3 2 1 (strongly disagree)

3. Experience and professional instinct are more important than assessment categories.

(strongly agree) 5 4 3 2 1 (strongly disagree)

4. The more a student talks in the examination, the larger the vocabulary he can use.

(strongly agree) 5 4 3 2 1 (strongly disagree)

5. I tend to give a high mark of vocabulary if the examinee uses synonyms or rephrasing to avoid repetition.

(strongly agree) 5 4 3 2 1 (strongly disagree)

6. I tend to give a high mark of vocabulary if the examinee uses many difficult or rare words.

(strongly agree) 5 4 3 2 1 (strongly disagree)

7. I tend to give a high mark of vocabulary if the examinee uses very complicated sentence structures.

(strongly agree) 5 4 3 2 1 (strongly disagree)

8. The more grammatical errors the examinee makes, the lower the mark of vocabulary.

(strongly agree) 5 4 3 2 1 (strongly disagree)

9. I tent to give a high mark for vocabulary if the examinee has a good pronunciation and intonation and can express himself/herself clearly.

(strongly agree) 5 4 3 2 1 (strongly disagree)

10. Please list other indicators of lexical richness:

Appendix 2 The CHAT format of the transcription

```
@Begin
@Languages:	eng
@Participants:	T11 Zhanghongbin Teacher, 249 Student
@ID:	eng|zhang|T11|39;|male|grade 2||Teacher||
@ID:	eng|zhang|249|||grade 2||Student||
@Transcriber:	chen
@Coder:	Jian Zhang
*T11:	hello .
*249:	hello .
*T11:	what's your name ?
*249:	my name is &Angela .
*T11:	what's your chinese name ?
*249:	my chinese name is &cuichengxi .
*T11:	&cuichengxi , where do you come from ?
*T11:	  where do you come from ?
*T11:	are you from &Beijing ?
*249:	yes .
*T11:	yes , how many people are there in your family ?
*249:	there are three people in my family .
*T11:	do you have any pets ?
*249:	yes , I do .
*T11:	what's your pet ?
*249:	I have , I have a rabbit .
*T11:	what color is the rabbit ?
*249:	it's white .
*T11:	how old is it ?
*249:	it's two .
*T11:	what does your rabbit like to eat ?
*249:	rabbit likes eat xxx and carrot .
*T11:	do you like carrots ?
*249:	yes , I like .
*T11:	how many rooms are there in your house ?
*249:	there are four rooms in house .
*T11:	how many bedrooms ?
*249:	one bedroom .
*T11:	is there a kitchen ?
*249:	yes .
*T11:	yes , what do you have for your breakfast ?
*T11:	do you have milk ?
*249:	chicken .
```

*T11:	you have chicken .
*T11:	now let's look at the picture , how many people are there ?
*249:	there are two people in your picture .
*T11:	where are they ?
*T11:	are they at home ?
*249:	they are sisters and my friend .
*T11:	this girl , is she wearing a black coat ?
*249:	no , it isn't , it's red coat .
*T11:	what are you wearing today ?
*249:	my wearing is red teeshirt and red shirt .
*T11:	now , let's look at this picture , what's this ?
*249:	computer .
*T11:	where is the boy ?
*249:	the computer .
*T11:	what's that ?
*249:	coat .
*T11:	and this ?
*249:	bag .
*T11:	where is the boy ?
*249:	it's under the bed .
*T11:	is he on the bed ?
*249:	no , he isn't .
*249:	he is behind bed .
*T11:	what's this ?
*249:	it's table .
*T11:	and this ?
*249:	chair .
*T11:	how many chairs are there ?
*249:	there are four chairs .
*T11:	where are they ?
*249:	inside the table .
*T11:	what's that ?
*T11:	what's this ?
*249:	it's book .
*T11:	now put the pen on the book , put it on the book .
*T11:	put this one under the book .
*T11:	what's this number ?
*249:	twenty five .
*T11:	his one ?
*249:	eighteen .
*T11:	this one ?
*249:	twenty seven .
*T11:	this one ?
*249:	fifty .

```
*T11:	what day is it today ?
*249:	it's thursday today .
*T11:	and tomorrow ?
*249:	it's friday .
*T11:	what's the month now ?
*T11:	is it july ?
*249:	it's yes , it is .
*T11:	what's next month ?
*249:	it's , it's august .
*T11:	thank you , that's all for your test , bye .
@End
```

Appendix 3: The output of the *mor post* command on the transcription

```
@Begin
@Languages:	eng
@Participants:T11 Zhanghongbin Teacher, 249 Student
@ID:	eng|zhang|T11|39;|male|grade 2||Teacher||
@ID:	eng|zhang|249|||grade 2||Student||
@Transcriber: chen
@Coder:  Jian Zhang
*T11:	hello .
%mor:	co|hello .
*249:	hello .
%mor:	co|hello .
*T11:	what's your name ?
%mor:	pro:wh|what~v:cop|be&3S pro:poss:det|your n|name ?
*249:	my name is &Angela .
%mor:	pro:poss:det|my n|name v:cop|be&3S .
*T11:	what's your chinese name ?
%mor:	pro:wh|what~v:cop|be&3S pro:poss:det|your adj|chinese n|name ?
*249:	my chinese name is &cuichengxi .
%mor:	pro:poss:det|my adj|chinese n|name v:cop|be&3S .
*T11:	&cuichengxi , where do you come from ?
%mor:	adv:wh|where mod|do pro|you v|come prep|from ?
*T11:	  where do you come from ?
%mor:	adv:wh|where mod|do pro|you v|come prep|from ?
*T11:	are you from &Beijing ?
%mor:	aux|be&PRES pro|you prep|from ?
*249:	yes .
%mor:	co|yes .
*T11:	yes , how many people are there in your family ?
%mor:	co|yes adv:wh|how qn|many n|person&PL v:cop|be&PRES adv:loc|there prep|in
	pro:poss:det|your n|family ?
*249:	there are three people in my family .
```

%mor:	adv:loc|there v:cop|be&PRES det:num|three n|person&PL prep|in pro:poss:det|my
	n|family .
*T11:	do you have any pets ?
%mor:	mod|do pro|you v|have qn|any n|pet-PL ?
*249:	yes , I do .
%mor:	co|yes pro|I v|do .
*T11:	what's your pet ?
%mor:	pro:wh|what~v:cop|be&3S pro:poss:det|your n|pet ?
*249:	I have , I have a rabbit .
%mor:	pro|I v|have pro|I v|have det|a n|rabbit .
*T11:	what color is the rabbit ?
%mor:	pro:wh|what n|color v:cop|be&3S det|the n|rabbit ?
*249:	it's white .
%mor:	pro|it~v:cop|be&3S adj|white .
*T11:	how old is it ?
%mor:	adv:wh|how adj|old v:cop|be&3S pro|it ?
*249:	it's two .
%mor:	pro|it~v:cop|be&3S det:num|two .
*T11:	what does your rabbit like to eat ?
%mor:	pro:wh|what mod|do&3S pro:poss:det|your n|rabbit v|like inf|to v|eat ?
*249:	rabbit likes eat xxx and carrot .
%mor:	n|rabbit v|like-3S v|eat unk|xxx conj|and n|carrot .
*T11:	do you like carrots ?
%mor:	mod|do pro|you v|like n|carrot-PL ?
*249:	yes , I like .
%mor:	co|yes pro|I v|like .
*T11:	how many rooms are there in your house ?
%mor:	adv:wh|how qn|many n|room-PL v:cop|be&PRES adv:loc|there prep|in pro:poss:det|your
	n|house ?
*249:	there are four rooms in house .
%mor:	adv:loc|there v:cop|be&PRES det:num|four n|room-PL prep|in n|house .
*T11:	how many bedrooms ?
%mor:	adv:wh|how qn|many n|+n|bed+n|room-PL ?
*249:	one bedroom .
%mor:	det:num|one n|+n|bed+n|room .
*T11:	is there a kitchen ?
%mor:	v:cop|be&3S adv:loc|there det|a n|kitchen ?
*249:	yes .
%mor:	co|yes .
*T11:	yes , what do you have for your breakfast ?
%mor:	co|yes pro:wh|what mod|do pro|you v|have prep|for pro:poss:det|your n|breakfast
	?
*T11:	do you have milk ?

%mor:	mod|do pro|you v|have n|milk ?
*249:	chicken .
%mor:	n|chicken .
*T11:	you have chicken .
%mor:	pro|you v|have n|chicken .
*T11:	now let's look at the picture , how many people are there ?
%mor:	adv|now v|let~pro|us v|look prep|at det|the n|picture adv:wh|how qn|many n|person&PL v:cop|be&PRES adv:loc|there ?
*249:	there are two people in your picture .
%mor:	adv:loc|there v:cop|be&PRES det:num|two n|person&PL prep|in pro:poss:det|your n|picture .
*T11:	where are they ?
%mor:	adv:wh|where aux|be&PRES pro|they ?
*T11:	are they at home ?
%mor:	aux|be&PRES pro|they prep|at n|home ?
*249:	they are sisters and my friend .
%mor:	pro|they v:cop|be&PRES n|sister-PL conj|and pro:poss:det|my n|friend .
*T11:	this girl , is she wearing a black coat ?
%mor:	det|this n|girl v:cop|be&3S pro|she part|wear-PROG det|a adj|black n|coat ?
*249:	no , it isn't , it's red coat .
%mor:	co|no pro|it v:cop|be&3S~neg|not pro|it~v:cop|be&3S adj|red n|coat .
*T11:	what are you wearing today ?
%mor:	pro:wh|what aux|be&PRES pro|you part|wear-PROG adv:tem|today ?
*249:	my wearing is red teeshirt and red shirt .
%mor:	pro:poss:det|my part|wear-PROG v:cop|be&3S adj|red n|teeshirt coord|and n|red n|shirt .
*T11:	now , let's look at this picture , what's this ?
%mor:	adv|now v|let~pro|us v|look prep|at det|this n|picture pro:wh|what~v:cop|be&3S pro:dem|this ?
*249:	computer .
%mor:	n|computer .
*T11:	where is the boy ?
%mor:	adv:wh|where v:cop|be&3S det|the n|boy ?
*249:	the computer .
%mor:	det|the n|computer .
*T11:	what's that ?
%mor:	pro:wh|what~v:cop|be&3S pro:dem|that ?
*249:	coat .
%mor:	n|coat .
*T11:	and this ?
%mor:	conj|and pro:dem|this ?
*249:	bag .
%mor:	n|bag .

*T11:	where is the boy ?
%mor:	adv:wh|where v:cop|be&3S det|the n|boy ?
*249:	it's under the bed .
%mor:	pro|it~v:cop|be&3S prep|under det|the n|bed .
*T11:	is he on the bed ?
%mor:	v:cop|be&3S pro|he prep|on det|the n|bed ?
*249:	no , he isn't .
%mor:	co|no pro|he v:cop|be&3S~neg|not .
*249:	he is behind bed .
%mor:	pro|he v:cop|be&3S prep|behind n|bed .
*T11:	what's this ?
%mor:	pro:wh|what~v:cop|be&3S pro:dem|this ?
*249:	it's table .
%mor:	pro|it~v:cop|be&3S n|table .
*T11:	and this ?
%mor:	conj|and pro:dem|this ?
*249:	chair .
%mor:	n|chair .
*T11:	how many chairs are there ?
%mor:	adv:wh|how qn|many n|chair-PL v:cop|be&PRES adv:loc|there ?
*249:	there are four chairs .
%mor:	adv:loc|there v:cop|be&PRES det:num|four n|chair-PL .
*T11:	where are they ?
%mor:	adv:wh|where aux|be&PRES pro|they ?
*249:	inside the table .
%mor:	adj|inside det|the n|table .
*T11:	what's that ?
%mor:	pro:wh|what~v:cop|be&3S pro:dem|that ?
*T11:	what's this ?
%mor:	pro:wh|what~v:cop|be&3S pro:dem|this ?
*249:	it's book .
%mor:	pro|it~v:cop|be&3S n|book .
*T11:	now put the pen on the book , put it on the book .
%mor:	adv|now v|put&ZERO det|the n|pen prep|on det|the n|book v|put&ZERO pro|it prep|on det|the n|book .
*T11:	put this one under the book .
%mor:	v|put&ZERO det|this pro:indef|one prep|under det|the n|book .
*T11:	what's this number ?
%mor:	pro:wh|what~v:cop|be&3S det|this n|number ?
*249:	twenty five .
%mor:	det:num|twenty det:num|five .
*T11:	his one ?
%mor:	pro:poss:det|his pro:indef|one ?
*249:	eighteen .

```
%mor:	det:num|eighteen .
*T11:	this one ?
%mor:	det|this pro:indef|one ?
*249:	twenty seven .
%mor:	det:num|twenty det:num|seven .
*T11:	this one ?
%mor:	det|this pro:indef|one ?
*249:	fifty .
%mor:	det:num|fifty .
*T11:	what day is it today ?
%mor:	pro:wh|what n|day v:cop|be&3S pro|it adv:tem|today ?
*249:	it's thursday today .
%mor:	pro|it~v:cop|be&3S n|thursday adv:tem|today .
*T11:	and tomorrow ?
%mor:	conj|and adv:tem|tomorrow ?
*249:	it's friday .
%mor:	pro|it~v:cop|be&3S n|friday .
*T11:	what's the month now ?
%mor:	pro:wh|what~v:cop|be&3S det|the n|month adv|now ?
*T11:	is it july ?
%mor:	v:cop|be&3S pro|it n|july ?
*249:	it's yes , it is .
%mor:	pro|it~v:cop|be&3S co|yes pro|it v:cop|be&3S .
*T11:	what's next month ?
%mor:	pro:wh|what~v:cop|be&3S adj|next n|month ?
*249:	it's , it's august .
%mor:	pro|it~aux|be&3S pro|it~v:cop|be&3S adj|august .
*T11:	thank you , that's all for your test , bye .
%mor:	v|thank pro|you rel|that~aux|be&3S qn|all prep|for pro:poss:det|your n|test
    co|bye .
@End
```

Appendix 4: The CLAN results of MLU of both examiner and candidate

```
mlu +f
Tue Oct 18 05:07:06 2011
mlu (10-Oct-2011) is conducting analyses on:
  ONLY dependent tiers matching: %MOR;
****************************************
From file <d:\NEW CLAN DATA 新 CLAN 语料 \MOR POST\chen 2\zhanghongbin 080703-95.tin.mor.pst.cex>
MLU for Speaker: *249:
  MLU (xxx, yyy and www are EXCLUDED from the utterance and morpheme counts):
        Number of: utterances = 38, morphemes = 133
        Ratio of morphemes over utterances = 3.500
        Standard deviation = 2.221

MLU for Speaker: *t11:
    MLU (xxx, yyy and www are EXCLUDED from the utterance and morpheme counts):
        Number of: utterances = 48, morphemes = 234
        Ratio of morphemes over utterances = 4.875
        Standard deviation = 2.579
```

Appendix 5 The CLAN result of VOCD of the candidate

```
vocd +t*249 +t%mor -t* +s*-%% +s*&%% +s*~%% +f
Tue Oct 18 07:54:46 2011
vocd (10-Oct-2011) is conducting analyses on:
    ONLY dependent tiers matching: %MOR;
****************************************
From file <d:\NEW CLAN DATA 新CLAN语料\MOR POST\CHEN 2\zhanghongbin 080703-95.tin.mor.pst.cex>
co|hello
pro:poss:det|my n|name v:cop|be
pro:poss:det|my adj|chinese n|name v:cop|be
co|yes
adv:loc|there v:cop|be det:num|three n|person prep|in pro:poss:det|my n|family
co|yes pro|i v|do
pro|i v|have pro|i v|have det|a n|rabbit
pro|it adj|white
pro|it det:num|two
n|rabbit v|like v|eat conj|and n|carrot
co|yes pro|i v|like
adv:loc|there v:cop|be det:num|four n|room prep|in n|house
det:num|one n|+n|bed+n|room
co|yes
n|chicken
adv:loc|there v:cop|be det:num|two n|person prep|in pro:poss:det|your n|picture
pro|they v:cop|be n|sister conj|and pro:poss:det|my n|friend
co|no pro|it v:cop|be pro|it adj|red n|coat
pro:poss:det|my part|wear v:cop|be adj|red n|teeshirt coord|and n|red n|shirt
n|computer
det|the n|computer
n|coat
n|bag
```

```
pro|it prep|under det|the n|bed
co|no pro|he v:cop|be
pro|he v:cop|be prep|behind n|bed
pro|it n|table
n|chair
adv:loc|there v:cop|be det:num|four n|chair
adj|inside det|the n|table
pro|it n|book
det:num|twenty det:num|five
det:num|eighteen
det:num|twenty det:num|seven
det:num|fifty
pro|it n|thursday adv:tem|today
pro|it n|friday
pro|it co|yes pro|it v:cop|be
pro|it pro|it adj|august

tokens  samples    ttr      st.dev      D
   35      100    0.7286    0.054     34.224
   36      100    0.7339    0.062     36.431
   37      100    0.7297    0.055     36.450
   38      100    0.7232    0.057     35.891
   39      100    0.7313    0.055     38.807
   40      100    0.7187    0.053     36.736
   41      100    0.7154    0.055     36.857
   42      100    0.7026    0.060     34.861
   43      100    0.7077    0.053     36.833
   44      100    0.7093    0.048     38.079
   45      100    0.6989    0.046     36.498
   46      100    0.7002    0.052     37.617
   47      100    0.6887    0.055     35.811
   48      100    0.6873    0.041     36.254
   49      100    0.6816    0.054     35.755
   50      100    0.6816    0.051     36.478

D: average = 36.474; std dev. = 1.081
D_optimum     <36.46; min least sq val = 0.000>

tokens  samples    ttr      st.dev      D
   35      100    0.7389    0.062     36.583
   36      100    0.7397    0.059     37.842
   37      100    0.7324    0.056     37.091
   38      100    0.7100    0.063     33.027
   39      100    0.7254    0.059     37.363
   40        100     0.7200     0.060       37.029
   41        100     0.7088     0.056       35.364
   42        100     0.7098     0.054       36.449
   43        100     0.7042     0.047       36.041
   44        100     0.6973     0.052       35.333
   45        100     0.6960     0.049       35.853
   46        100     0.6920     0.043       35.750
   47        100     0.6881     0.051       35.671
   48        100     0.6844     0.049       35.615
   49        100     0.6845     0.051       36.382
   50        100     0.6764     0.047       35.346

D: average = 36.046; std dev. = 1.079
D_optimum        <36.01; min least sq val = 0.000>

tokens   samples      ttr       st.dev        D
   35        100     0.7360     0.063       35.908
   36        100     0.7211     0.054       33.562
   37        100     0.7330     0.065       37.221
   38        100     0.7126     0.063       33.577
   39        100     0.7228     0.058       36.757
   40        100     0.7205     0.063       37.146
   41        100     0.7151     0.062       36.801
   42        100     0.7050     0.049       35.381
   43        100     0.7002     0.056       35.167
   44        100     0.7030     0.055       36.598
   45        100     0.6978     0.057       36.249
   46        100     0.6828     0.054       33.810
   47        100     0.6930     0.039       36.757
   48        100     0.6802     0.053       34.724
   49        100     0.6786     0.047       35.097
   50        100     0.6836     0.045       36.924

D: average = 35.730; std dev. = 1.246
D_optimum        <35.71; min least sq val = 0.001>

VOCD RESULTS SUMMARY
====================
     Types,Tokens,TTR:   <62,120,0.516667>
   D_optimum    values:   <36.46, 36.01, 35.71>
   D_optimum average:   36.06
```

Appendix 6   The CLAN result of VOCD of the examiner

```
vocd +t*t11 +t%mor -t* +s*-%% +s*&%% +s*~%% +f
Tue Oct 18 07:56:58 2011
vocd (10-Oct-2011) is conducting analyses on:
   ONLY dependent tiers matching: %MOR;
****************************************
From file <d:\NEW CLAN DATA 新CLAN语料\MOR POST\CHEN 2\zhanghongbin
080703-95.tin.mor.pst.cex>
co|hello
pro:wh|what pro:poss:det|your n|name
pro:wh|what pro:poss:det|your adj|chinese n|name
adv:wh|where mod|do pro|you v|come prep|from
adv:wh|where mod|do pro|you v|come prep|from
aux|be pro|you prep|from
co|yes adv:wh|how qn|many n|person v:cop|be adv:loc|there prep|in pro:poss:det|your
n|family
mod|do pro|you v|have qn|any n|pet
pro:wh|what pro:poss:det|your n|pet
pro:wh|what n|color v:cop|be det|the n|rabbit
adv:wh|how adj|old v:cop|be pro|it
pro:wh|what mod|do pro:poss:det|your n|rabbit v|like inf|to v|eat
mod|do pro|you v|like n|carrot
adv:wh|how qn|many n|room v:cop|be adv:loc|there prep|in pro:poss:det|your n|house
adv:wh|how qn|many n|+n|bed+n|room
v:cop|be adv:loc|there det|a n|kitchen
co|yes pro:wh|what mod|do pro|you v|have prep|for pro:poss:det|your n|breakfast
mod|do pro|you v|have n|milk
pro|you v|have n|chicken
adv|now v|let v|look prep|at det|the n|picture adv:wh|how qn|many n|person v:cop|be
adv:loc|there
adv:wh|where aux|be pro|they
aux|be pro|they prep|at n|home
det|this n|girl v:cop|be pro|she part|wear det|a adj|black n|coat
pro:wh|what aux|be pro|you part|wear adv:tem|today
adv|now v|let v|look prep|at det|this n|picture pro:wh|what pro:dem|this
adv:wh|where v:cop|be det|the n|boy
pro:wh|what pro:dem|that
conj|and pro:dem|this
adv:wh|where v:cop|be det|the n|boy
v:cop|be pro|he prep|on det|the n|bed
pro:wh|what pro:dem|this
conj|and pro:dem|this
adv:wh|how qn|many n|chair v:cop|be adv:loc|there
adv:wh|where aux|be pro|they
```

```
pro:wh|what pro:dem|that
pro:wh|what pro:dem|this
adv|now v|put det|the n|pen prep|on det|the n|book v|put pro|it prep|on det|the n|book
v|put det|this pro:indef|one prep|under det|the n|book
pro:wh|what det|this n|number
pro:poss:det|his pro:indef|one
det|this pro:indef|one
det|this pro:indef|one
pro:wh|what n|day v:cop|be pro|it adv:tem|today
conj|and adv:tem|tomorrow
pro:wh|what det|the n|month adv|now
v:cop|be pro|it n|july
pro:wh|what adj|next n|month
v|thank pro|you rel|that qn|all prep|for pro:poss:det|your n|test co|bye

tokens  samples   ttr     st.dev     D
   35      100   0.7346   0.066    35.576
   36      100   0.7286   0.070    35.210
   37      100   0.7430   0.067    39.732
   38      100   0.7387   0.055    39.674
   39      100   0.7269   0.060    37.733
   40      100   0.7200   0.058    37.029
   41      100   0.7068   0.058    34.935
   42      100   0.7057   0.051    35.539
   43      100   0.6935   0.045    33.734
   44      100   0.7018   0.057    36.340
   45      100   0.6891   0.050    34.368
   46      100   0.6909   0.053    35.512
   47      100   0.6964   0.050    37.535
   48      100   0.6885   0.052    36.532
   49      100   0.6722   0.056    33.781
   50      100   0.6786   0.054    35.820

D: average = 36.191; std dev. = 1.746
D_optimum     <36.11; min least sq val = 0.001>

tokens  samples   ttr     st.dev     D
   35      100   0.7506   0.057    39.525
   36      100   0.7397   0.059    37.842
   37      100   0.7278   0.055    36.009
   38      100   0.7263   0.059    36.623
   39      100   0.7126   0.062    34.446
   40      100   0.7143   0.058    35.706
   41      100   0.7029   0.051    34.097
   42        100      0.7031      0.058        34.965
   43        100      0.7053      0.058        36.303
   44        100      0.7036      0.056        36.753
   45        100      0.7002      0.053        36.801
   46        100      0.6978      0.056        37.065
   47        100      0.6832      0.053        34.622
   48        100      0.6902      0.050        36.906
   49        100      0.6733      0.049        33.989
   50        100      0.6776      0.047        35.603

D: average = 36.079; std dev. = 1.429
D_optimum        <36.02; min least sq val = 0.001>

tokens   samples      ttr       st.dev        D
   35        100      0.7434      0.063        37.697
   36        100      0.7361      0.066        36.961
   37        100      0.7300      0.069        36.514
   38        100      0.7239      0.059        36.072
   39        100      0.7308      0.061        38.679
   40        100      0.7073      0.065        34.173
   41        100      0.7205      0.061        38.072
   42        100      0.7176      0.059        38.298
   43        100      0.7053      0.056        36.303
   44        100      0.6993      0.063        35.782
   45        100      0.7016      0.056        37.106
   46        100      0.6920      0.056        35.750
   47        100      0.6870      0.060        35.440
   48        100      0.6821      0.058        35.121
   49        100      0.6841      0.053        36.292
   50        100      0.6702      0.059        34.049

D: average = 36.394; std dev. = 1.325
D_optimum        <36.34; min least sq val = 0.001>

VOCD RESULTS SUMMARY
===================
    Types,Tokens,TTR:    <79,212,0.372642>
   D_optimum    values:    <36.11, 36.02, 36.34>
   D_optimum average:    36.16
```

Appendix 7: SPSS output of the comparisons of candidate Type

```
ONEWAY typesS BY level
STATISTICS DESCRIPTIVES BROWNFORSYTHE WELCH
```

```
PLOT MEANS
MISSING ANALYSIS
POSTHOC=TUKEY ALPHA(0.05).
```

Descriptives

typesS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
2	58	72.2931	14.81821	1.94573	68.3969	76.1894	40.00	106.00
5	60	100.3833	25.75444	3.32488	93.7303	107.0364	29.00	170.00
7	60	104.1667	24.60892	3.17700	97.8095	110.5238	67.00	192.00
Total	178	92.5056	26.35680	1.97553	88.6070	96.4042	29.00	192.00

ANOVA

typesS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	35577.960	2	17788.980	35.627	.000
Within Groups	87380.534	175	499.317		
Total	122958.494	177			

Robust Tests of Equality of Means

typesS

	Statistic[a]	df1	df2	Sig.
Welch	49.919	2	109.827	.000
Brown-Forsythe	35.966	2	153.057	.000

a. Asymptotically F distributed.

Multiple Comparisons

typesS

Tukey HSD

(I) level	(J) level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2	5	-28.09023*	4.11471	.000	-37.8167	-18.3638
	7	-31.87356*	4.11471	.000	-41.6000	-22.1471

5	2	28.09023*	4.11471	.000	18.3638	37.8167
	7	-3.78333	4.07969	.624	-13.4270	5.8604
7	2	31.87356*	4.11471	.000	22.1471	41.6000
	5	3.78333	4.07969	.624	-5.8604	13.4270

*. The mean difference is significant at the 0.05 level.

typesS

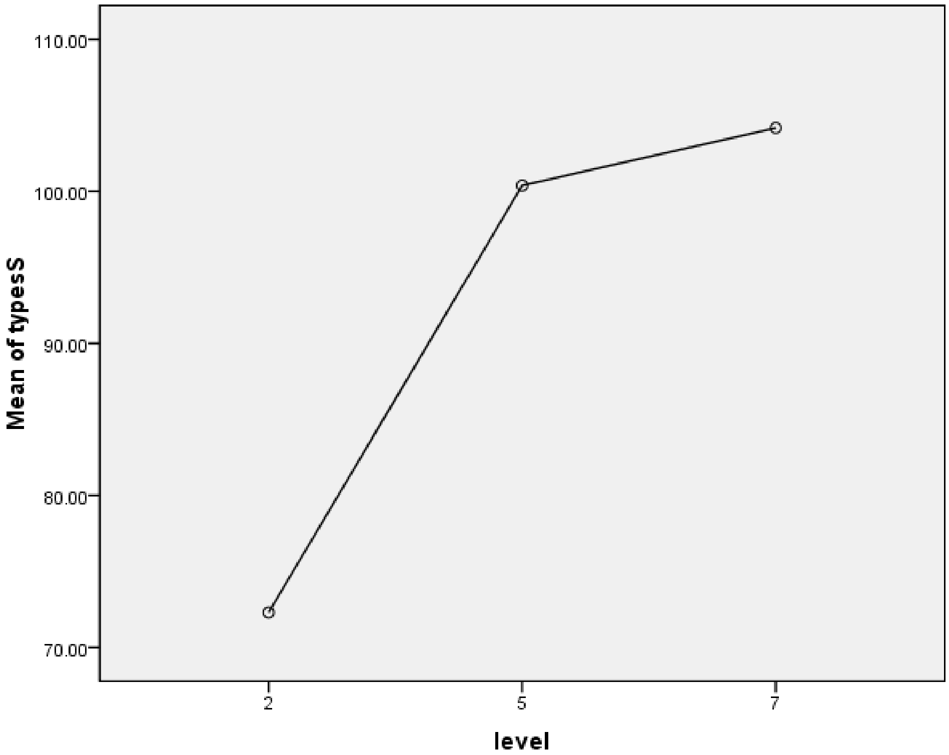
Tukey HSD[a,,b]

		Subset for alpha = 0.05	
level	N	1	2
2	58	72.2931	
5	60		100.3833
7	60		104.1667
Sig.		1.000	.627

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 59.318.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.



Appendix 8. The interview outline for the qualitative research

1. This is a Grade 2 (5, or 7) examination you conducted in 2008. Would you please listen to it again and mark the candidate. Please mark the candidate according to the criteria of Readiness, Pronunciation, Usage (and also Focus if it is Grade 5 or 7), and finally a final score is also needed.
2. Could you talk me through your decisions, how and why you gave this mark?
3. What do you think of the candidate's vocabulary use (if the examiner didn't talk about vocabulary) ?
4. Here are scores you marked in 2008. What do you think?

Any immediate reactions?

5. I found from the quantitative analysis that there is significant diffrnece of the lexical variables between Grade 2 and Grade 5 candidates and Grade 2 and Grade 7 candidates. However, there is no significant differnece between the lexical variables between Grade 5 and Grade 7 candidates. What do you think of it?