Lexical richness of Chinese candidates in the graded oral English examinations

Article

Published Version

Lexical Richness of Chinese Candidates in the Graded Oral English Examinations

Jian Zhang **Michael H. Daller**[1]

Abstract: The main purpose of this study is to explore the lexical richness of Chinese candidates of different proficiency levels in a graded examination in spoken English (GESE), which is an exam developed by Trinity College, London and administered in Beijing, China by trained local examiners. We compared 5 lexical indices and the mean length of utterances (MLU) of the GEGE candidates of three proficiency levels. The quantitative results first indicate that lexical richness plays an important role in these oral interviews and there are significant correlations between the lexical indices, the MLU and the proficiency level of the candidates. Furthermore, candidates who pass the oral exams have significantly higher scores for lexical richness. There are significant differences between the lexical richness scores at the Initial level (GESE Grade 2) and at the Elementary level (Grade 5). But only some measures show significant differences between the Elementary level (Grade 5) and the Intermediate level (Grade 7), which casts some doubt on the validity of the classification system. One reason for this result might be the fact that a Grade 7 certificate leads to higher chances in the admissions process for prestigious secondary schools and there is a strong interest by candidates and by private preparatory schools to get a certificate at this level. Some candidates might have enrolled on this level without meeting the criteria fully. Overall, our results show that measures of lexical richness and MLU are good predictors for success in oral interviews, but that factors other than proficiency play a role when it comes to the placement of students in Grade 7. The unique contribution of the present study resides in the fact that we use a large sample drawn randomly from a huge corpus of oral interviews. On this basis, we can gain further insights in the role that vocabulary knowledge plays in oral interviews.

Keywords: oral interviews, language testing, vocabulary knowledge

[1] **Corresponding author: Michael Daller**, English Language and Applied Linguistics, University of Reading, Reading, UK, E-mail: M.Daller@Reading.ac.uk

Jian Zhang, English Department, Beijing Institute of Technology, Beijing, China

1. Introduction

Lexical knowledge has always been regarded as an important aspect in language assessment, and measures of vocabulary knowledge can be used to compare learners at different levels. According to Milton (2008: 334) "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". However, the relationship between lexical knowledge and lexical measures is not straight forward. The oral assessment of vocabulary knowledge is an under-researched area. In the last two decades or so, there has been much research on the lexical measures of written texts (Laufer & Nation 1995,1999; Malvern & Richards 2002; Meara & Bell 2001; Nation 2001), but only a few studies were carried out on oral data. Sandlund, Sundquvist and Nyroos (2016) carried out a meta study on oral exams, but their overview on studies published between 2004 and 2014 did not include any research on vocabulary knowledge and oral interviews. Malvern and Richards (2002) investigate lexical knowledge in oral interviews in French and come to the conclusion that it is difficult for teachers to rate "the range of vocabulary while listening to a tape recording" (p. 95). More research on the role of vocabulary knowledge in oral interviews is therefore needed. To our knowledge there are no studies that compare the role of vocabulary at different proficiency levels in oral examinations.

The present study aims to investigate the lexical richness in spoken English by Chinese candidates at different levels in the Graded Examinations in Spoken English for Speakers of Other Languages (GESE) co-sponsored by Trinity College, London and Beijing Educational Examinations Authority (BEEA). The lexical richness measures of candidates of 3 different proficiency levels (Initial, Elementary and Intermediate Stage in GESE) are compared. In addition, we investigate the role of vocabulary knowledge in the exam by comparing lexical richness scores of candidates who fail and those who pass the exam.

2. Literature review

2.`1 Vocabulary/lexical knowledge

While many researchers recognize the importance of vocabulary knowledge (e.g. Carter & McCarthy 1988; Krashen 1989; Li & Kirby 2015; Meara, 2002, 2005; Milton 2013; Nation 1990, 2001; Read 2000; Schmitt & McCarthy1997; Singleton1999; and Treffers-Daller & Milton 2013), they operationalize the nature of vocabulary knowledge in different ways. Two major approaches have been suggested: a detailed list with all key components of vocabulary knowledge or a holistic approach using global rating descriptors. For example, Richards (1976: 83) lists seven assumptions about knowing a word, which include knowing 1) the probability of encountering that word in oral or written discourse, which means intuitive knowledge about the frequency of words and the probability of words that are associated with it; 2) the limitation of the use of word in different functions and situations; 3) the syntactic behavior associated with the word; 4) the underlying form of a word and its possible derivations; 5) the network of association between that word and other words in the language; 6) the meaning and semantic value of a word and finally how many different meanings are associated with the word and 7) all dimensions continue to develop throughout a person's life span.

In Nation's framework (1990) there are eight subcategories, and each includes both receptive and productive knowledge: *form* (spoken form and written form), *position* (grammatical patterns and collocations), *functio*n (frequency and appropriateness) and *meaning* (concept and association). Later, Nation (2001:36-59) refines his vocabulary knowledge into three categories (*form, meaning* and *use)* and nine aspects: "spoken form; written form; concept and referents; word parts; connecting form and meaning; associations; grammatical functions; collocations; constraints on use".

From the above we conclude that defining lexical knowledge is complex and that measuring all these aspects is hardly possible in practice. In the same vain Meara (1996) argues that it is impracticable to measure all the attributes of word knowledge although it would be theoretically desirable. He proposes a model of lexical competence with only two dimensions: *size* (how large is the vocabulary) and *organization* (how well structured

is the vocabulary). These two dimensions "are characteristics of the system as a whole, rather than features of the individual words that make up the system" (Meara1996:3). Wesche and Paribakht (1996) also propose two aspects of lexical knowledge. In addition to breadth (size), which is similar to Meara's size dimension, they introduce the dimension of depth. They argue that the existing measures of vocabulary size (breadth) cannot show the quality of lexical knowledge (depth) that learners have. Henriksen (1999) proposes three dimensions of vocabulary knowledge: 1) a partial-precise knowledge dimension; 2) a depth of knowledge dimension, and 3) a receptive-productive dimension. The first two dimensions both involve the process of acquiring word meaning. The partial-precise knowledge refers to the development from "rough categorization or vagueness to more precision and mastery" of word meaning (311), whereas the second dimension is primarily associated with "understanding of sense relations"(314) or how a semantic network is built, which is similar to the organizational dimension described by Meara (1996).The three dimensions proposed reflect the vocabulary development process, and represent continua of the development of a learner's vocabulary.

Daller, Milton and Treffers-Daller (2007) argue that vocabulary knowledge is composed of breadth, depth and fluency. They add to the traditional dimensions of breadth and depth the dimension of fluency, which reflects the ease and speed of accessing and using vocabulary. "This hypothetical space allows learners with different types of vocabulary knowledge to be positioned differently in this space and systematically distinguished from each other" (Milton 2013:62).

2.2 Lexical richness of L2 learners in written and spoken discourse

"Lexical knowledge is now known to be an absolutely crucial factor across the whole spectrum of L2 activities" (Singleton 1999: 4-5), and there has been a rise in the search for reliable and valid measures of L2 learner's lexical knowledge over the last 20 years. A comprehensive discussion of all proposed measures is beyond the scope of the present

paper, but major developments are analyzed in the following. According to Read (2000: 200-203), lexical richness is the general term for vocabulary knowledge and there are four aspects of lexical richness in analyzing writing compositions: *lexical variation, lexical sophistication, lexical density* and *number of errors*. Good writing is assumed to have the lexical features of a wide range of words and expressions (*lexical variation),* the use of infrequent or difficult words appropriate to the topic and style of the writing (*lexical sophistication),* a high percentage of contents words rather than grammatical words (*lexical density*) and finally, few or a low level of lexical mistakes (*number of errors).*

One traditional measure of *lexical variation* or *diversity* is the type/token ratio (*TTR*), which is widely used in child and L2 acquisition research. However, this measure has been criticized by many researchers (Daller, van Hout &Treffers-Daller 2003; McCarthy & Jarvis 2007; Malvern & Richards 2002; Malvern et al. 2004, and Vermeer 2000;), because it is sensitive to text length. With increasing text length (number of tokens) the TTR is systematically decreasing and it is therefore not possible to compare texts with different lengths. An early measure that tries to overcome this problem is the *Index of Guiraud* (Guiraud 1954). This index tries to compensate for the falling *TTR* curve through a simple mathematical transformation: Types/√ Tokens. Vermeer（2000）discusses the reliability and validity of 10 measures of lexical richness (*tokens, types, lemmas, hapax legomena, TTR, corrected TTR, Guiraud, log TTR, Uber index* and *theoretical vocabulary*) and examines their behavior with spontaneous speech data. The results show that *Guiraud* gives the best indication of lexical richness, at least in the early stages of vocabulary acquisition (up to 3,000 words). Daller, van Hout and Treffers-Daller (2003) also argue that *Guiraud,* the mathematical transformation of TTR compensates for the systematically falling TTR with increasing text length, but that it is not always clear whether it over- or under compensates. Nevertheless, *Guiraud* has been used successfully in a recent study by Treffers-Daller, Parslow and Williams (2016), who show that *Guiraud* and simply the number of different words used by the candidates are the best predictors for language proficiency at different levels of the Common European Framework (Council of Europe, 2001).

In recent years, many researchers have investigated the strength and weakness of lexical indices and applied a variety of measures of lexical richness on their data (Daller, van Hout &Treffers-Daller; Durán et al. 2004; Jarvis 2002,2003; Jarvis & Daller, 2013; Lu 2012; Malvern & Richards, 2002, Malvern et al. 2004; McCarthy & Jarvis, 2007; Meara, 2005; Read 2000; Richards et al. 2009 and Vermeer 2000). One of the most popular measures is *D* proposed by Malvern and Richards (2002), which is claimed to overcome the text size effect of the *TTR* and other *TTR* transformations. Malvern and Richards (2002) studied 34 British students of two secondary school classes taking their oral exam in French for the General Certificate of Secondary Education (GCSE), which was conducted by their own teachers. The results of GCSE, ratings of the GCSE exam given by 24 experienced teachers and *D* values of both the teacher and student were obtained for the analyses. The study showed that *D* is a valid measure of lexical richness. Yu (2009) used *D* as a measure of lexical richness on both spoken and written data of the same subjects to investigate the relationship between lexical diversity and the holistic quality of both written and spoken discourse. He found that *D* was an effective measure of lexical richness and that it correlated significantly and positively with the overall writing and speaking performance of the candidates as well as their general language proficiency. He also found that *D* was a better indicator of speaking than for writing performance.

However, although *D* is widely used as a valid measure of lexical richness, some researchers have argued that there is a need for more research on this measure. 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 results indicate that the curve-fitting formulae of *D* provide accurate models of the *type-token* curves for short texts. However, McCarthy and Jarvis (2007) also point out that although *D* seems to be a reliable and valid indicator of lexical diversity in many earlier studies, its reliability was still in question because *D* is also significantly affected by text length when the size of the sample is above a certain range. Malvern et al. (2004) also acknowledged that *D* could be affected by text length, but they argue that these effects are not significant for the text lengths in most studies.

In addition to the research on lexical diversity as an indication of lexical knowledge, many researchers (Daller, van Hout & Treffers-Daller, 2003; Laufer& Nation, 1995; Wen, 1999;Wesche & Paribakht 1996 and Vermeer 2000) have argued that a more effective measure of lexical richness may involve the *lexical sophistication* or the *frequency* of words. Laufer and Nation (1995) first proposed a lexical richness measure, the Lexical Frequency Profile (*LFP*), which shows the proportion of words of different frequencies and the academic words used by learners. It 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 frequency levels. It can be used as diagnostic as well as a research tool. 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. The study shows that *advanced TTR* and *Guiraud Advanced* (*AG*), which include information about the frequency of the types, outperform other measures that do not include information about word frequency. *AG* 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.

Overall, the literature review shows that there are various measures of lexical richness, and some of them seem to be valid in certain contexts. However, "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" (Laufer, 2005: 587). In research on lexical richness in spoken data, many researchers use data that were collected on the availability basis, and the data sets are usually small. Large data sets as in the present study with learners at different levels are, to our knowledge, only rare, and random sampling is in most cases not used.

3. Research Questions and Hypotheses

3.1 Research Questions

3.1.1 How important is lexical richness for the grade classification of the candidates?

3.1.2 What is the validity of the measures used in the present study with regard to oral proficiency?

3.2 Hypotheses

3.2.1Candidates in higher grades will have higher scores for lexical richness and *MLU* than those on lower Grades.

3.2.2 Lexical richness scores and *MLU* can distinguish between candidates who passed the oral exams and those who failed (for a detailed description of the measures see the chapter of Methodology).

4. Methodology

4.1 GESE and GESE data

GESE is a set of international examinations sponsored by Trinity College, London, and was introduced in China in 1999. GESE has 12 Grades in 4 Stages, with three Grades in each Stage: Initial Stage (Grade 1 to 3), Elementary Stage (Grade 4 to 6), Intermediate Stage (Grade 7 to 9) and Advanced Stage (Grades 10 to 12). The examinations of the first three stages (Grade 1 to 9) are conducted by local Chinese examiners and are analyzed in the present study.

GESE is an oral interview between an examiner and a candidate. In the Initial Stage (Grade1 to 3), there is only one examination phase, *Conversation*. At this stage, the examiner asks simple questions and asks the candidate to do some actions according to the instructions. The examiner controls the conversation. In the Elementary Stage (Grades 4 to 6), there are two phases, the *Topic Phas*e and the *Conversation Phase.* In the Topic Phase, the candidate first gives a talk on a topic of the grade and then the examiner asks questions and answers the candidate's questions. 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 in both

phases. In the Intermediate Stage (Grades 7 to 9), in addition to the two phases of topic and conversation, a third *Interactive Phase* is added: it is also conversation, but the candidate has to keep the conversation going and maintain the interaction by asking questions based on an oral prompt given by the examiner.

In the present research, the GESE candidates of three different stages are expected to have different language proficiency levels according to the Syllabus of GESE. There are different conversation topics and requirements for candidates of the three grades. 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. Simple present tense is used. Topics in Grade 5 are festivals, means of transport and music etc., which are more difficult and need explanation or clarification, past tense and present perfect tense are used in the conversation. The candidate is expected to take more initiatives during the conversations by asking a couple of questions on each topic. While in Grade 7 more formal and abstract topics such as education, national customs and products and recycling are discussed, and the subjunctive mood is required to use in the conversation. The candidates from Grade 2, 5 and 7 are expected to be classified into different levels of the Common European Framework of Reference for Languages (CEFR).Grade 2 is in the Initial Stage, which relates to level A1 (Basic User) of the CEFR; Grade 5 is in the Elementary Stage which is between the level A2 to B1 (Basic User to Independent User) of the CEFR, and Grade 7 is in the Intermediate Stage which relates to level B2 (Independent User) of the CEFR. Table 1 gives an overview of the collected data and the different examination forms at different levels.

Table 1
The collected Data and the overview of GESE at different levels.

Proficiency level Grade (refer to***CEFR)**	**Initial Grade 2 *(A1)***	**Elemental Grade 5 *(A2-B1)***	**Intermediate Grade 7 *(B2)***
Data collected	***6 minutes***	***5 minutes***	***5 minutes***
	Conversation	***Conversation***	***Interactive tasks***
Examples of topics	Rooms; family and friends; days of the week, etc.	Festival, means of transport; music, etc.	Education; early memories; ecycling, etc.

Examples of some grammar required	Simple present	Past tense; present perfect	Conditional clause; passive voice
Examples of some Functions required	Simple answers	Description, facts and ideas	Opinions; advice; elicting futher information
Communicative abilities	Verbal and non-verbal responses	Answer questions and ask at least one question	Take control and keep the interaction going; take and give up turns appropriately

*The Common European Framework of Reference for Language: Learning, Teaching, Assessment (2001) (CEFR)

4.2 Participants

4.2.1 GESE Candidates

60 GESE examinations were collected randomly from 3 proficiency levels respectively and there are 180 candidates in total. The candidates are mainly primary school students and a very small number of school teachers. The average age of the three groups is 9.1,11.9 and 15.8 years. Most candidates have followed a three-month training course in commercial training schools in addition to the English class at public schools. The candidates usually start from Grade 1 or Grade 2 and then continue to pass higher grades of GESE. The candidates can skip grades instead of taking examinations one grade after another. The training schools administer mock exams that lead to suggestions for an appropriate grade. The candidates can, however, try a different grade. A certificate from Grade 7 gives the young candidates the possibility to enter prestigious secondary schools when they graduate from the Primary school, and entry into Grade 7 is therefore in high demand. An overview of the candidates' age, gender, pass rate and proficiency level is shown in Table 2.

Table 2.
Background information and proficiency levels of the candidates

Grades/Stages	Age (year)			Gender		Pass rate	*reference to CEFR**
	Min.	Max.	Mean	M	F		
Grade 2 (n=60) Initial Stage	6.0	14.5	9.1	34	26	83%	A1(Basic User)
Grade 5 (n=60) Elementary Stage	10.01	30.6	11.9	31	29	55%	between A2 and B1(Basic User to Independent User)
Grade 7 (n=60) Intermediate Stage	8.9	45.9	15.8	28	32	25%	B2 (Independent User)

*The Common European Framework of Reference for Language: Learning, Teaching, Assessment (2001) (CEFR)

4.2.2 The examiners

There are 23 examiners who conducted the 180 examinations collected for the present research. All examiners were experienced university lecturers of English working at universities in Beijing, and all had been GESE examiners for at least 6 years. Among them, there are 20 female examiners and 3 male examiners, which reflects the fact that there are many more female rather than male English teachers in China. All the 23 examiners conducted examinations at Grade 1 to 6. Among them there are 11 senior examiners, who had a certificate to conduct Grade 7-9 exams.

The examiners do not know the candidates. All the GESE examinations conducted by Chinese examiners were audio-recorded and supervised by panels both in China and at Trinity College, London in the UK. All Chinese examiners receive standardization training sponsored by Trinity London and BEEA twice annually.

4.3 The measures

Five measures of lexical richness are applied in the present study: *Types, Tokens, D, Guiraud (G)and Advanced Guiraud (AG*).*Token* refers to the total number of words used in a text and *Types* refers to the number of different words used. The number of *Tokens*, *Types* and *D* were obtained by using the software *CLAN* of the *CHILDES database* (MacWhinney 2000).The index *AG* was obtained by using the software tool *Guiraud Advanced* (Daller, 2010). In addition to the indices of lexical richness, the Mean Length of Utterances (*MLU*) is used as a measure of the candidates' general language proficiency, and it was also obtained from *CLAN.*

Brown (1973) proposed the *MLU* on the basis a morpheme count. It has been widely accepted and used as an index for the general language development of children, and it has also been used in SLA research. Many researchers (for example, Arlman-Rupp et al., 1976; Hichkey, 1991, Parker and Brorson, 2005) have argued that the *MLU* measured in words has advantages over the *MLU* in morphemes because words are easier to identify and to calculate. Richards and Malvern (2000) and Malvern and Richards (2002) used the *MLU* in words to analyze accommodation of teachers and students in oral interviews. the *MLU* in word count is used in the present study.

4.4 Data collection procedure

First, 60 oral examinations from each Grade (2, 5 and 7) were collected randomly from the GESE Examination Corpus (BEEA, 2008).The data for analysis in the present study were chosen from the examinations: the whole examination (conversation) of Grade 2 which lasts about 6 minutes, the *Conversation Phase* of Grade 5 and the *Interactive task* (conversation with focus on interaction) of Grade 7, which both last about 5 minutes. The information about the collected data is also presented in Table 1.

Next, the audio-recorded data were transcribed into the *CHAT* format of the *CHILDES* Language Data Exchange System (*CHILDES*, see MacWhinney 2000) by the first author with the help of Chen Hui and Wang Xiaoqing. The candidate and the examiner data were separated for further processing and calculation. In the present study, only the quantitative data from the candidates and the interviews with the examiners are discussed (for details about the quantitative data from the examiners see Zhang 2014). We deleted two outliers:

candidate 259 was not intelligible and candidate 260 produced less than 35 tokens and therefore *D* cannot be computed for technical reasons. Valid data from 158 candidates are analyzed in the present study. In the collected data, the range of the number of token is 66 (minimum) to 482 (maximum), with a mean score of 230 and a standard deviation of 82.3. According to previous research, this is within the range where our measures of lexical richness are valid.

Finally, three senior GESE examiners were interviewed concerning the performance of the candidates and some interactions of Grade 7 are also examined to get a further understanding of the quantitative analyses results. A detailed data processing procedure is discussed in Zhang (2014).

5. Results

5.1 Lexical richness of the candidates of three different grades

Table 3 shows that the mean scores for lexical richness increase with higher grades. This holds for *Tokens*, *Types*, *D* and *MLU*. *Table 3* also gives the results of a post-hoc comparison (Tukey) between the groups.

Table 3
Lexical Richness and *MLU* mean scores between the different groups (all candidates) (ANOVA and post-hoc comparison Tukey)

Variables	Grade	N	Mean	SD	ANOVA			*Post-hoc comparison (Tukey)(p)*		
					F	*df*	P	2 vs 5	2 vs 7	5 vs 7
Tokens	*2*	58	166.9	42.4	37.6	2	<.001	<.001	<.001	ns
	5	60	247.1	87.6						
	7	60	273.9	70.3						
Types	*2*	58	72.3	14.8	35.6	2	<.001	<.001	<.001	ns
	5	60	100.4	25.8						
	7	60	104.2	24.6						
D	*2*	58	33.0	10.9	30.6	2	<.001	<.001	<.001	ns
	5	60	47.1	11.9						
	7	60	49.3	13.7						
G	*2*	58	4.4	.98	34.0	2	<.001	<.001	<.001	ns
	5	60	7.0	2.2						
	7	60	6.8	2.1						

AG	*2*	58	161.9	59.8	14.6	2	<.001	<.01	ns	<.001
	5	60	198.5	58.4						
	7	60	142.6	54.3						
MLU	*2*	58	4.5	1.0	111.4	2	<.001	<.001	<.000	<.001
	5	60	11.3	4.6						
	7	60	15.6	5.3						

The ANOVA shows that the differences between Grade 2 and Grade 5 are all highly significant as are most differences between Grade 2 and Grade 7. The differences between Grade 5 and Grade 7 are only significant for *MLU* and *AG*. For *AG* the differences for these two grades are counter-intuitive as Grade 5 gets higher scores. A possible explanation for this might be the low pass rate of Grade 7 students. In the present study, the pass rate of Grades 2, 5 and 7 are 83%, 55% and 25% respectively. Most students (75%) failed the Grade 7 examinations and the unqualified candidates might pull down the mean score of Grade 7. In the next step, only the data of the students who passed the examination are computed; the results are presented in Tables 4. and Table 5.

Table 4

Lexical Richness and MLU mean Scores between the different groups (candidates who passed)

(ANOVA and post-hoc comparison Tukey)

Measure	Grade	N	Mean	SD	*ANOVA*			*Post-hoc comparison (Tukey)*		
					F	*df*	*p*	2vs5	2 vs7	5vs7
Tokens	*2*	50	171.5	42.3	46.9	2	<.001	<.001	<.001	ns
	5	33	271.8	80.8						
	7	15	305.4	60.5						
Types	*2*	50	75.1	13.7	64.4	2	<.001	<.001	<.001	ns
	5	33	109.4	20.6						
	7	15	119.16	19.3						
D	*2*	50	33.7	12.1	39.7	2	<.001	<.001	<.001	ns
	5	33	49.8	9.6						
	7	15	58.1	10.4						
G	*2*	50	4.6	.92	51.1	2	<.001	<.001	<.001	ns
	5	33	7.7	2.1						
	7	15	7.5	2.0						
AG	*2*	50	170.4	55.7	6.3	2	<.01	<.01	ns	<.01
	5	33	213.3	54.9						
	7	15	170.8	66.8						
MLU	*2*	50	4.6	.93	75.1	2	<.001	<.001	<.001	ns
	5	33	12.1	4.8						
	7	15	15.7	6.2						

Table 4 shows that for the candidates who passed the examination, the mean scores of lexical richness increase with higher grades except *G* and *AG*. The Grade 5 *G* score (7.7) is higher than that of Grade 7 (7.5) and the Grade 5 *AG* score (213.3) is higher than that of Grade 7 (170.8). For *G*, the reason might be that there is no significant difference between *G* in Grade 5 and 7, the slight difference is very likely caused by chance. For *AG*, one reasons might be the biased classification of grades, and the second might be that the Grade 7 candidates has used less difficult or low-frequency words than the Grade 5 candidates .

The post-hoc comparison (Tukey) shows that there are significant differences between Grade 2 and 5, Grade 2 and 7 (except *AG)*, but still there is no significant difference between Grade 5 and 7 in most measures, which show a similar situation as presented in Table 3.

5.2 Differences in lexical richness scores between candidates who passed and who failed

In the following section, we compare the lexical richness and the *MLU* mean scores between candidates who passed and who failed for each Grade level separately. Table 5. shows the scores for Grade 2 candidates and the results of t-tests.

Table 5.
Differences between Grade 2 lexical mean scores of the Pass and Fail group

Measures	*Pass (n=49)*	*Fail (n=9)*	*t*	*Sig.(2-tailed)*	*p*
Type	75.1	55	5.68	.000	<.001
Token	171.5	138.8	2.61	.036	<.05
D	33.7	21.4	3.81	.007	<.05
G	4.6	3.4	11.07	.003	<.05
AG	170.4	114.9	2.48	.009	<.05
MLU	4.6	3.6	2.78	.006	<.05

For Grade 2, the Pass group has overall higher scores than the Fail group, and the independent samples t-tests show that all the *p value* < .05, which indicates that there are

statistically significant differences between the pass group and the fail group in all the measures studied. Table 6 shows the comparison for Grade 5 candidates.

Table 6

Differences between Grade 5 mean lexical scores of the Pass and Fail group

Measures	Pass (n=33)	Fail (n=27)	*t*	*Sig.(2-tailed)*	*p*
Type	109.4	89.4	3.14	.002	<.05
Token	271.8	216.8	2.51	.014	<.05
D	49.8	43.6	1.97	.047	<.05
G	7.7	5.7	.273	.008	<.05
AG	213.3	180.4	2.23	.029	<.05
MLU	12.1	10.3	1.53	.131	*ns*

Similar to the results from Grade 2, all scores expect MIU at Grade 5 can distinguish between the Pass and the Fail group in the expected direction. The independent samples t-tests show that except for MLU, the other *p value* < .05, which indicates that there are statistically significant differences between the Pass group and the Fail group in all the lexical measures studied, but no significant difference in *MLU*.

Table 7

Differences between Grade 7 lexical measures of the Pass and Fail group

Measures	Pass (n=15)	Fail (n=45)	*t*	*Sig.(2-tailed)*	*p*
Type	119.1	98.7	3.37	.002	<.05
Token	305.4	262.5	2.32	.027	<.05
D	58.1	46.1	3.62	.001	<.05
G	7.5	6.5	1.78	.092	*ns*
AG	170.8	132.3	2.13	.046	<.05
MLU	15.7	15.5	.13	.90	*ns*

Table 7 shows that in Grade 7 all the scores in the Fail group are lower than that of the Pass group. All the lexical scores except *G* and *MLU* show significant differences between the Pass and the Fail group at this level.

It is found from the results that all the lexical measures can distinguish the Pass group and the Fail group except *G* in Grade 7, which prove the validity of these lexical measures. However *MLU*, as a general indicator of language proficiency, can only distinguish the Pass and Fail groups at the initial stage of GESE but cannot distinguish the difference in elementary and intermediate stage. It may indicate that it is not as sensitive as the lexical richness measures in detect minute differences in language proficiency levels.

5.3 Results from the interviews with Grade 7 examiners

Three senior examiners coded as A, B and C were interviewed (see Zhang 2014 for details of research methods) and they were also prompted to talk about their opinions on the general performance and the vocabulary use of the candidates. One reoccurring theme was the poor performance of the Grade 7 candidates. A qualitative analysis of the interviews with the examiners led to the following explanation of this counter intuitive judgment that there was no significant difference between the vocabulary use of Grade 5 and Grade 7 GESE candidates.

First, most Grade 7 candidates in 2008 chose a grade that is higher than their real proficiency level because they wanted to gain the potential benefit of a Grade 7 GESE

certificate that may help them to enter a top middle school in Beijing, which is not possible with a Grade 5 certificate. A typical explanation given by examiner A is: "If a candidate had a Grade 7 certificate at that time (2008), he or she would be accepted by the best middle schools in Beijing…The parents were also very keen on it. The children in primary schools who wanted to enter a key middle school all take Grade 7".

Second, the interactive tasks of Grade 7 require communicative skills and abilities that many candidates do not have. Exam-oriented training and recitation of prepared monologues did not help them with their communicative abilities. Examiner B mentioned that "candidates from some training school had handouts, and they chose the same topics and recited a lot in interviews. They were not quite ready for the grade". Examiner A believes that there should be a difference between the vocabulary use of candidates in Grade 5 and candidates in Grade 7. There must be something wrong if there is no difference. All examiners expressed very similar views that most Grade 7 candidates were below the required proficiency level, and their vocabulary use was unsatisfactory.

5.4 Analysis of the interactions at Grade 7

Transcripts of the examinations were investigated and it was found that most interactions in the Interactive Tasks of Grade 7 are not smooth at all. As shown in the examples below, there was very often a long pause after the examiner's prompts, and many candidates just struggled to say something without fully understanding the examiner. They sometimes turned to the topics they had prepared instead of getting involved in the conversation. As a 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 Grade 7 than in Grade 5.

Transcript 1
T23: examiner
747: Grade 7 candidate

T23: right, thanks, thank you very much for the topic, and it's time to move on to

the next topic, for the next part, I will tell you something, then you need to ask me questions to find out more information.

747: ok

T23: and you need to keep the conversation going, after about four minutes, I will end the conversation.

747: ok.

T23: are you ready?

747: yes, I'm ready.

T23: right, many teenagers I know want to study at a university which is a long way from their home, I think there are two sides to this.

747: … (*long pause >5.0 seconds*)

T23: that's all.

747: but I think, umm, I beg your pardon?

T23: some young people want to study at a university far from their home, I think there are two sides to this.

747: (*long pause>5.0 seconds*) umm, I'm sorry.

…

Transcript 2

T21: examiner

737: Grade 7 candidate

T21: for the next part, I will tell you something, you have to ask me questions to find out more information, you need to keep the conversation going, after about four minutes, I'll end the conversation, are you ready?

737: yes

T21: I just moved to a new town and feeling lonely, I am wondering how I make some new friends?

737: you want to make some new friends?

T21: yes

737: what kind of friends you want to make?

T21: ordinary friends, we can talk with each other.

737: ok, where are you in now, I forgot.

T21: it's in the west part of Beijing.

737: in Beijing?

T21: yes.

737: have you ever been to other country?

T21: other country? yes.

737: where?

T21: I have been to some European countries and I've also been to Canada.

737: how do you think Canada, how do you think the food?

…

Transcript 3

T21: examiner
735: Grade 7 candidate

T21: for the next part , I'll tell you something , then you have to ask me questions to find out more information , you need to keep the conversation going , after about four minutes , I'll end the conversation.
T21: are you ready?
735: ok
T21: my brother was told by his doctor that he needs to lose weight, he is finding it very difficult.
735: losing weight, is very difficult, there are many reasons, first, umm (*long pause >5.0 seconds*),
there are many ways to lose weight, first, take sports, … second, eat healthy food **…**
(The candidate starts a talk on how to lose weight without any turn-taking. Instead of engaging in negotiation with the examiner, the candidate turns an interactive task into a monologue.)

In the first transcript, the candidate didn't catch what the examiner was saying. Even after the examiner rephrased her prompt, the candidate still couldn't understand her and there was a breakdown of communication. In the given five minutes, the candidate didn't fulfill the tasks.

In Transcript 2, although the candidate asked questions to keep the conversation going, the questions except the first two were rather irrelevant. The candidate didn't talk about making friends but suddenly turned to some irrelevant questions that they may have prepared beforehand. In Transcript 3, instead of getting more information from the examiner by asking questions, the candidate turned the dialogue into a prepared monologue on how to lose weight. The problems of irrelevant questions in Transcript 2 and inappropriate monologues in Transcript 3 are very common among Grade 7 candidates in the phase of interactive tasks, which shows that the candidates were not ready for the grade. They had rather weak communicative abilities and turned to the strategy of prepared monologue when they are unable to engage in a meaningful conversation. This explains why 75% Grade 7 candidates failed in the examinations. Just as examiner A commented "I think the candidate and their parents were just trying their luck. Many Grade 7 candidates took a wrong grade". Most candidates did not meet the requirements of Grade 7.

6. Discussion

This research is quantitative in nature although we interviewed some examiners and examined some interactions and communicative failure in Grade 7. The main results are based on quantitative measures of lexical richness, which can indicate to some extent the learners'use of vocabulary but cannot determine the quality of the speech. In addition, as the lexical measures cannot be calculated during the process of oral interview, they are investigated mainly for research purposes.

Hypothesis 3.2.1., which states that the measures of lexical richness can distinguish between the different grade levels, is supported by our findings in principle. The fact that all lexical richness scores are significantly higher for Grade 5 than for Grade 2 and for Grade 7 than for Grade 2 supports the validity of these measures. Lexical richness measures and *MLU* can show differences between Grade 2 and 5 and Grade 2 and 7, which is an indication of the validity of these measures in the given context. There is a general trend that the scores of all lexical richness measures increase with higher grades. In other words, the higher the grade, the higher are the score of the lexical richness measures. However, there is no significant difference between Grade 5 and Grade 7 in most lexical measures. We therefore assume that Grade 5 and Grade 7 candidates are basically at the same vocabulary knowledge level measured by these lexical measures.

The result might be caused by several factors. Firstly, the grade classification is biased because a Grade 7 certificate could lead to a place in an outstanding middle school. As a result, students who actually had not meet the requirements of this grade were nonetheless placed there to "win" a place at a prestigious school. They assumed that the training school could help them to achieve their goals in a short period of time through intensive exam-oriented training. This is supported by data from interviews with GESE examiners and analysis of the interactions in Grade 7. The examiners agreed that both learners and training schools are exam-orientated rather than proficiency-orientated. They seemed to put more emphasis on the efficiency of passing a grade and get a certificate than increasing language proficiency as such. Instead of having negotiations and interactions with the examiner,

many Grade 7 candidates turned to recitation of the memorized information in case of communicative difficulty

Secondly, there is a three-grade difference between Grade 2 and Grade 5, but there is only a two-grade difference between Grade 5 and Grade 7 (most candidates took Grade 7 in the Intermediate Stage as explained in the chapter of Methodology), which might probably cause the fact that some lexical richness measures are not sensitive enough to distinguish between Grade 5 and 7.

The result might also be caused by task type. Many researchers (e.g.Yuan and Ellis, 2003; Tavakoli and Foster 2011) found that preparation time and the task type may affect the L2 learner's oral performance, including lexical diversity. The data chosen from Grade 5 is conversation based on given topics whereas the data of Grade 7 is the interactive task, a conversation task in which the candidate is required to take initiatives to keep the conversation going. It is challenging but it is a required phrase for GESE candidates of Grade 7 and above. Hopefully in future research the task effects will be investigated for candidates of different proficiency levels as well as the candidates at the same level so as to provide more insights into the issue.

Hypothesis 3.2.2 states that the measures of lexical richness and *MLU* can distinguish between candidates who fail or pass the oral exams. This is supported to a large extend by our data. At Grade 2 all lexical richness scores and *MLU* are significantly higher for students who pass than students who fail. However, *MLU* does not distinguish between the pass and fail group at Grade 5 and Grade 7, and *G* does not distinguish between the fail and pass group at Grade 7. In addition to the biased classification as we discussed earlier, the reason for this might be that *MLU*, as well as *G*, a measure of diversity based on the traditional *TTR*, is not sensitive enough to detect small differences in language proficiency at the Intermediate Stage.

Token, Type, D and AG are the measure that can distinguish between qualified and poor performers at the same grade, which shows the validity of the lexical measures. The result also supports the argument proposed by many researchers (Daller, van Hout and Treffers-Daller 2003, Laufer and Nation 1995, Vermeer 2000, Wen 1999 and Wesche & Paribakht 1996) that the more effective measure of lexical richness may involve both

lexical sophistication and lexical diversity. Further research on *D* and *AG* may promote our understanding of the global indicator of lexical richness and help refine the existing tools of vocabulary research.

7. Conclusion

Vocabulary knowledge and lexical richness are an important aspect of language proficiency and play an important role in oral interviews. Generally, the candidates of higher grade tend to have higher lexical richness scores, and those candidates who pass have higher scores in lexical richness but also produce longer spoken texts than students who fail. However task type and the placement of candidates in Grade 7 may also play a role as *G* and *MLU* do not distinguish between passed and failed students at Grade 7. Among the lexical richness measures*, Token, Type, D* and *AG* can distinguish between good and poor performers of candidates at all grades. This is an indication that examiners are also sensitive to the lexical diversity and lexical sophistication of the candidates. The present study may have implication for vocabulary assessment and examiner training.

The present study has some limitations. Firstly, the quantitative measures do not indicate the quality of the speech or the exact vocabulary use of the candidate in GESE, and they are mainly for research purposes. In addition, only 6 variables that mainly involve lexical diversity and sophistication are chosen for the study. *Lexical Density*, one aspect of the lexical richness according to Read (2000), a measure that calculate the ratio of the content words and function words are not applied, and some new measures such as *MTLD* and *H-DD* (McCarthy and Jarvis 2010) are not applied in the study either. Future research with more measures that show different aspects of vocabulary use is welcome.

As the predictive validity of the measures varies between different proficiency levels, the results might not be generalizable and might not apply for other grades of GESE and other oral English examinations. More research is needed combining quantitative and qualitative methods, including more detailed discourse analysis and interviews with candidates as well as examiners. This might allow a more fine-grained analysis of the

factors that involve oral examinations. Nonetheless, the present study shows clearly the role of lexical richness as important factor in oral interviews.

Acknowledgements

This article was supported by Beijing Educational Examinations Authorities (BEEA) for providing research data and the State Fund the first author received from China Scholarship Council (No. 201706035053) for conducting research as a visiting scholar in the United Kingdom.

References:

Arlman-Rupp, A., Van Niekerk-de Haan, D., and Van de Sandt-Koenderman, M. 1976. Brown's early stages: some evidence from Dutch. *Journal of Child Language,* 3, 267-274.

BEEA (Beijing Educational Examinations Authorities), 2008. GESE Examinations Corpus. Beijing.

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

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

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Daller, M. (2010) Guirauds index of lexical richness. In: British Association of Applied Linguistics (conference proceedings), September 2010. Available from: http://eprints.uwe.ac.uk/11902

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

Daller, H., Milton, J., and Treffers-Daller, J. 2007. Editors' introduction: conventions, terminology and an overview of the book. In Daller, Milton, &Treffers-Daller (eds.), *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

Durán, P., Malvern, D., Richards, B., and Chipere, N. 2004. Developmental trends in lexical diversity. *AppliedLinguistics*, 25, 220- 242.

Guiraud, P. 1954. *Les caract´eristiques du vocabulaire*. Paris: Presses Universitaires de France.

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

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

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

Jarvis, S. and Daller, M. (eds.), 2013. *Vocabulary Knowledge: Human Ratings and Automated Measures.* Amsterdam: John Benjamins.

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

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world: A response to Meara (2005). *Applied linguistics*, *26*(4), 582-588.

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

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

Li, M and Kirby, J.R. 2015.The effects of vocabulary breadth and depth on English reading. *Applied Linguistics,* 36 (5), 611-634.

Lu, X. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190– 208.

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

Nation, P. 2001. *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

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

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

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

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

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*(2), 381-392.

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.

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

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

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, 333-348. Cambridge: Cambridge University Press.

Milton, J. 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In BardelC., Lindqvist, C. and Laufer, B(eds) *Eurosla Monographs Series 2, L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis***,** 57-78. Eurosla Monograph Series.

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.

Parker, M.D. and K. 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), 365-376.

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

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

Richards, B., Daller, M.H., Malvern, D., Meara, P., Milton, J. and 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, 77-89.

Sandlund, E., Sundqvist, P. and Nyroos, L. 2016. Testing L2 talk: A view of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass,*10 (1), 14-29.

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

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

Tavakoli, P and Foster P. 2011.Task Design and Second Language Performance: The Effect of Narrative Type on Learner Output. *Language Learning,* 61(S1), 37-72.

Treffers-Daller, J and Milton, J. 2013. Vocabulary size revisited; the link between vocabulary size and academic achievement. *Applied Linguistics Review,* 4 (1), 151-172.

Treffers-Daller, J., Parslow, P., and Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics, 39*(3), 302-327.

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

Wen, Q. 1999. *Oral English Testing and Teaching*. Shanghai: Shanghai Foreign Languages Education Press.

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

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

Yuan, F. and Ellis, R. 2003. The effect of pre-task planning and on-line planning on fluency, complexity, accuracy in L2 monologic oral production. *Applied Linguistics 23*(1), 1–27.

Zhang, J. 2014. Lexical richness and accommodation in oral English examinations with Chinese examiners. PhD. thesis. Bristol: University of the West of England, Bristol.