Measuring reading and vocabulary with the Test for English Majors Band 4: a concurrent validity study

Measuring reading and vocabulary with the Test for English Majors Band 4: a concurrent validity study

Jeanine Treffers-Daller (University of Reading) and Huang Jingyi (Tong Liao Vocational College)

Word count without title page, abstract and appendices (but including references): 9,149

Corresponding author

Jeanine Treffers-Daller (University of Reading)

Email: j.c.treffers-daller@reading.ac.uk

ORCID: https://orcid.org/0000-0002-6575-6736

Huang Jingyi ORCID:
https://orcid.org/0000-0003-1021-2579

Biostatement

Jeanine Treffers-Daller is Professor of Multilingualism at the University of Reading. She has published widely on the measurement of language ability (including language dominance) in bilinguals and second language learners, with a specific focus on vocabulary. She is on the Editorial Board of *Bilingualism, Language and Cognition*, as well as *the International Journal of Bilingualism*.

Jingyi Huang is Assistant teacher at Tong Liao Vocational College, China. Her main research interests include second language teaching, teacher assessments of vocabulary knowledge, and reading comprehension.

Abstract

The aim of the study was to investigate to what extent the vocabulary and reading components of the Test for English Majors Band 4 (TEM-4) do indeed measure the constructs they are supposed to measure. The concurrent validity of the reading and vocabulary components of the test was studied by correlating results on the TEM-4 with widely used tests of vocabulary and reading comprehension. A total of 60 English Major students, in their second year at a university in North China, were tested using the bilingual Mandarin-English version of the Vocabulary Size Test (Nation & Beglar, 2007), the Vocabulary Knowledge Scale (Brown, 2008), which assesses depth of vocabulary knowledge and the TEM-4. 30 out of these 60 students were also individually tested using the York Assessment of reading comprehension Secondary (Snowling et al., 2009). The results show that the reading component of the TEM-4 did not correlate with the YARC. Instead there were modest correlations between the TEM-4 reading component, on the one hand, and the VST and the VKS on the other hand. It is therefore likely that the reading component of the TEM-4 taps into vocabulary knowledge rather than reading. Regression analyses also confirmed these findings. We conclude that the reading component of the TEM-4 underrepresents the construct of reading, and that the theoretical underpinning and the ways in which reading is assessed in this test will need to be reconsidered.

Keywords: concurrent validity, TEM-4, vocabulary size, vocabulary depth, reading comprehension

1. Introduction

Testing students' English language proficiency is an enormous undertaking in China. Every year no less than 18 million students take the College English Test as part of their undergraduate studies (Yu & Jin, 2014). The current project focuses on another widely used test, namely the Test for English majors, Band IV (TEM-4), for which the number of test takers has soared to 270,000 between 1992 and 2015 (Xu & Liu, 2018). The TEM-4 is a criterion-referenced English language test for university undergraduates majoring in English Language and Literature in China and aimed at testing a wide range of aspects of students' English proficiency levels, as well as their knowledge of the content of the National College English Teaching Syllabus for English Majors. According to Jin and Fan (2011), the TEM-4 is considered to be a reliable and valid test. At the same time, they mention a study by Chen (2009), who notes there are issues with construct underrepresentation and construct-irrelevant variance, both of which are key aspects of the construct relevance of a test (Messick, 1995). Construct underrepresentation refers to situations where important dimensions of a construct are not included in a test. Construct-irrelevant variance, by contrast, is found in situations where a test is too broad and contains variance that is associated with other constructs. An example of how these issues affect a test of reading comprehension can be found in Ready, Chaudry, Schatz and Strazzullo's (2012) study of the Nelson-Denny reading comprehension test (Brown, Fishco & Hanna, 1993). The authors found that many test items could be answered correctly by testees who had not seen the reading passage (i.e. passageless administration). This means that the scores obtained did not really reflect reading comprehension of the passage: Instead, they were associated with general intelligence, levels of vocabulary and knowledge and broad

reading skills. In other words, there was a considerable amount of construct-irrelevant variance in the data set. As pointed out by Weir (2005, p.18), it is very important to ensure that the construct we are eliciting with a test is indeed the construct we aim to measure. Khalifa and Weir's (2009) model of reading presents a detailed overview of the different processes involved in reading. One of these is inferencing, that is the ability of readers to go beyond explicitly stated ideas and to build a mental representation of what a text is about (Khalifa & Weir, 2009, p. 50). Graesser, McNamara and Louwerse (2003) also mention the importance of measuring readers' inferencing skills. A reading test which only measures understanding of literal meanings as found in the text but does not assess respondents' ability to infer meanings that are not explicitly mentioned in the text is therefore likely to underrepresent the construct (see also the Methods section for further discussion).

From the literature available to researchers in Western Europe it is not clear to what extent the validity issues sketched above been investigated in any detail for the TEM-4. In the 1990s Zhou, Weir and Green's (1998) carried out a three-year validation study of the TEM-4 and the TEM-8 published by the TEM Test Centre in Shanghai[1]. Zhou et al. found that the TEM tests (both Band IV and Band VIII) were "reasonably valid and reliable tests" (p. 63). However, they also note that the concurrent validity of the version of the TEM-4 (from 1995) was not high, as correlations between scores on the different components of the TEM-4 and the corresponding scores on the Test of English for Educational Purposes (TEEP), developed at

[1] We are very grateful to Rita Green for having provided us with a copy of this report, and to Anthony Zhang and Changqing Zheng for sending us the final version.

Reading University (UK), were relatively low (r = 0.4037). The authors suggest that one of the reasons for the low correlations might be participants' lack of familiarity with the open-ended format of the TEEP questions. The reading component of the TEM-4 has changed considerably since the publication of Zhou et al's validation study in that the test no longer distinguishes between careful reading and speed reading. Therefore any conclusions from this report may not hold anymore for more recent versions of the TEM-4. Jin and Fan (2011) report that the average test reliability was good between 2008 and 2010 but also note that there are still very few published validation studies of the test: It is not clear whether, for example, a passageless administration of the TEM-4 has been attempted or whether any concurrent validity studies have been carried out. They therefore call for further validation studies of the test.

The current study sets out to evaluate the concurrent validity of the vocabulary and reading components of the TEM-4 by correlating respondents' scores on this test with those on widely used tests of vocabulary size and depth as well as a test of reading comprehension, the York Assessment of Reading Comprehension Secondary (from now on the YARC, Snowling et al., 2009). We are of course aware that tests are not suitable for all learners in all contexts (Schmitt, Nation & Kremmel, 2019). A potential issue with the YARC Secondary is that it was developed for students in the UK (both L1 and L2 users of English) who receive English input in their daily lives. We therefore also look in detail at the suitability of the test for the target group of L2 learners in China

In the current study we will, first of all, investigate to what extent the reading and vocabulary components of the TEM-4 correlate with widely used tests of vocabulary and

reading. Second, we will look at the relative contribution of size and depth of vocabulary knowledge to explaining reading comprehension as measured with the TEM-4 and the YARC.

The structure of the current paper is as follows. In Section 2 we first present the construct of reading and the ways in which this is measured in different tests. Section 3 focuses on vocabulary and its measurement. In section 4 we present the current study. The methods used in the study are given in Section 5, the results are presented and discussed in Section 6 and we finish with a summary and conclusion (Section 7).

2. Reading comprehension: the construct and its measurement

Under the Simple View of Reading (Gough & Tunmer, 1986), which is the most widely used model of reading ability, decoding is one of the two key dimensions of reading, the other one being linguistic comprehension. Decoding refers to readers' ability to recognise words, that is to make a link between the printed word and the appropriate entry in the mental lexicon. Reading comprehension, by contrast, is defined as the ability to understand written language. More specifically it refers to readers' ability to use lexical (semantic) information and to derive sentence and discourse level interpretations from it (Gough & Tunmer, 1986). Reading comprehension is thus different from linguistic comprehension, in that the former relates to written and the latter to aural language. While Gough and Tunmer recognise that there are many aspects to understanding a text, the two dimensions of decoding and linguistic comprehension are the essential ones without which no reading can take place.

In her discussion of the construct of reading, Snow (2002, p. 11) elaborates on the notion of reading comprehension, which she defines as "the process of simultaneously extracting and

constructing meaning through interaction and involvement with written language." To be able to construe meaning in this way, at sentence as well as textual levels, the reader needs to know about the domain and the topic, have the necessary linguistic and discourse knowledge, and rely on cognitive capacities (e.g., attention, memory, critical analytic ability, inferencing, visualization ability).

Reading fluency, that is the "ability to read rapidly with ease and accuracy, and to read with appropriate expression and phrasing" (Grabe 2009, p. 291), is another variable which has been found to correlate strongly with reading comprehension. Grabe (2010) suggests that readers who read fast and have very efficient word recognition skills are generally able to integrate information from different sources and construe text level interpretations, even under time pressure. Indeed, the available research on L2 reading fluency indicates that word reading fluency and passage reading fluency impact on reading comprehension. Conversely, a lack of reading fluency is also a reliable predictor of reading comprehension difficulties (Stanovich, 1991).

Among the linguistic variables that are relevant for reading, vocabulary has often been found to be a key predictor (Laufer & Ravenhorst-Kalovski, 2010). Since Stanovich's (1986) seminal publication on the Matthew effect in reading, it has been known that there is a reciprocal relationship between vocabulary knowledge and reading: readers who know more words are better readers, and better readers can learn new words from reading more easily. The relationship between reading and vocabulary has therefore been the focus of a wide range of studies. For the purposes of this chapter we will limit the presentation of the available literature

to studies which focus on adult L2 learners of English, as these are the target group for the current study.

According to Nation and Waring (1997) adult learners of English often have vocabularies smaller than 5,000 words, despite having learned English for several years. However, for reading a newspaper or a novel, however, around 8,000 – 9,000 words are needed (Nation, 2006). This means that both decoding and reading comprehension are likely to be more difficult for this group than for monolinguals. Clearly it is not just the size of person's vocabulary that matters but also how well words are known (vocabulary depth – see also Section 3). Qian (2005) found that depth of vocabulary knowledge contributes more to reading comprehension than readers' vocabulary size, although Binder, Cote, Lee, Bessette and Vu (2017) only vocabulary size explained unique variance in reading fluency. As the authors point out, it is quite challenging to measure vocabulary depth and the battery used in the study may not have been sufficient to tap this construct successfully.

Before looking in more detail at the construct of vocabulary and how this can be measured a few words must be said about the measurement of different components of reading. As Ready et al. (2012) point out, there are not many reading tests for adults, and even fewer that specifically target adult L2 learners. One of the tests for adult native speakers is the National Adult Reading Test (NART, Nelson, 1982). This test assesses word recognition and familiarity of words, and consists of 50 words of increasing difficulty, all of which have irregular grapheme-phoneme correspondences. However, this test is unlikely to be suitable for L2 learners who have small vocabularies because a Vocabprofile analysis of the items, provided by Tom Cobb's Lextutor (https://www.lextutor.ca/vp/), shows that 82% of the words in the

NART belong to frequency levels lower than 5k, and the test includes words from frequency layers up to 20k. L2 learners are therefore likely to know only very few of these items.

The YARC Secondary is a comprehensive test of reading, based on the Simple View of Reading (Gough & Tunmer, 1986). The test was developed for 11 to 16-year-old students in the UK (including non-native speakers of English). A sample of 89 students for whom English was an Additional Language was included in the standardisation sample. As might be expected, scores for non-native speakers were lower than those for native speakers (see https://www.gl-assessment.co.uk/support/yarc-support/). The reliability information as provided in the manual shows that Cronbach's alpha varied from .85 to .90 for most components, except for the summarization part, where reliability ranged from .65 to .74. The reading passages are accessible for readers with smaller vocabularies because they contain very few words beyond the 5k level (see Section 5 for further discussion). According to Stothard (2010), the reading comprehension questions include a range of inference questions that can be used to assess predictive, evaluative, knowledge-based and cohesive inference. As a detailed analysis of these different kinds of inferencing is beyond the scope of the current paper, the reader is referred to Bowyer-Crane and Snowling (2005) for an overview. In the YARC, reading comprehension is not only assessed with comprehension questions, but also with a summarization task (See Yu, 2008, for a discussion of summarization to assess reading comprehension).

Whether or not the YARC Secondary is also suitable for university students of English in China is an empirical question. The TEM-4 contains a reading component, which consists of 4-5 different reading texts and understanding of these texts is measured with multiple choice

questions. However, it is not clear which model of reading underpins the test, and it seems to only target reading comprehension, as measures of word recognition or fluency are not included. Although inferencing skills are mentioned in the 2015 test specification[2], most of the reading comprehension questions in the TEM-4 test paper used in the current study, and later TEM-4 papers from 2015 and 2016 which we have seen, appear to mainly assess literal information.

Before explaining the specific objectives of the current study we first briefly present two concepts which, as we have seen in section 1, are key to reading comprehension, namely vocabulary size and vocabulary depth.

3. Vocabulary size and vocabulary depth: the constructs and their measurement

Most researchers in the field would agree that it is not only important for readers to know a large number of words, but how well they know these words (the depth of their knowledge) is relevant too. The most widely used model of vocabulary knowledge is that of Nation (2013) who proposes there are three basic components to vocabulary knowledge, namely form, meaning and use, each of which can be known to different degrees both receptively (passively) or productively (actively). There is a wide range of possible options for vocabulary tests for L2

[2] We are very grateful to Guoxing Yu for providing us with the information regarding the inferencing skills in the test specification for the TEM-4 (2015)

learners, depending on whether active or passive recognition or active or passive recall is measured (see Laufer & Goldstein, 2004), even though many widely used vocabulary tests have not sufficiently been validated (Schmitt, Nation & Kremmel, 2019).

Nation and Beglar's (2007) Vocabulary Size Test (VST) is a widely used free test of both first language and second language learners' written receptive vocabulary size, that is the vocabulary size needed for reading in English. As pointed out by Gyllstad et al (2015), there is a clear risk that the test overestimates the vocabulary learners know, as is often the case with multiple choice format. Further information about the validity of the test can be obtained from Beglar (2010).

Testing vocabulary depth is even more complex than testing vocabulary size. A widely used format is the Vocabulary Knowledge Scale (Paribakht &Wesche, 1996), which is a self-report form on which respondents indicate on a five-point scale how well they know a particular word. While there are obvious disadvantages to using self-report, the format has been widely used, also because users can include items they want to focus on. In addition, other tests of vocabulary depth, such as the Word Associates Test (Read, 1993), which taps into collocational knowledge as well as antonyms and synonyms, can be very complex and unsuitable for learners with relatively small vocabularies.

Measuring vocabulary knowledge is very important for studies of reading, because vocabulary is key determinant of reading comprehension (Laufer & Ravenhorst-Kalovski, 2010). In a comprehensive study of over 600 learners of English from eight different countries, Schmitt, Jiang and Grabe (2011) found that there is no specific vocabulary threshold for understanding text. Rather students with higher scores on different vocabulary measures could

demonstrate more in-depth understanding of the texts in the study. Establishing an exact threshold is also complicated because the results depend on the degree of comprehension that is required: if only 60% needs to be understood, then a coverage of 95% is probably sufficient. However, most teachers would probably want their students to understand more of the text. The authors therefore suggest that if 70% comprehension is required, a 98-99% coverage is needed. Importantly, they also found that even students who knew all the words did not always get full marks on the comprehension task. This is likely due to the fact that non-native speakers may not be familiar with the genre or the wider context or lack cultural information that is needed to comprehend a text.

In the Chinese context, according to figures from the College English Curriculum Requirements of the Ministry of Education from 2007, reported in Zhao, Wang, Coniam and Xie (2017), at the Basic Level, students should know 4,795 words and 700 phrases and expressions; at the Intermediate level it is 6,395 words, and 1,200 phrases and expressions; and at the Advanced level it is 7,675 words and 1,870 phrases and expressions However, actual vocabulary knowledge of students is often much more limited, even among students who are majoring in English. In a recent study among second year non-English major students in China in which Nation and Beglar's (2007) Vocabulary Size Test was used to measure vocabulary size, Wang and Treffers-Daller (2017) found that the students knew on average just under 3,000 words, which is far less than the 5,000 words they are required to know according to the syllabus. These figures may even be inflated as according to Gyllstad, Vilkaitė and Schmitt (2015), vocabulary sizes as measured with the VST overestimate students' knowledge by up to 26%. Other sources do indeed report lower actual vocabulary sizes for Chinese university

students. In a study among English majors and non-English majors, in which Schmitt, Schmitt and Clapham's (2001) vocabulary levels test was used to assess students' vocabulary knowledge, Zhang (2009) found students knew 2,156 words receptively and 859 productively. While a detailed overview of the vocabulary knowledge of Chinese students is beyond the scope of the current study, these studies suggest that Chinese university students do not always know enough words for independent reading of authentic materials such as texts from newspapers or novels.

The current study hopes to contribute to a further understanding of these issues, as will be explained in the next section.

4. The current study

The aim of the current project is, first of all, to evaluate the concurrent validity of the reading and vocabulary components of the TEM-4. Investigating concurrent validity entails investigating whether the data collected from one instrument correlate highly with the data collected from another instrument which purportedly measures the same construct (Cohen, Manion & Morrison, 2018, p. 258). We will assume this test to be a valid test of reading if there are positive and significant correlations between the results of the TEM-4 and different components of the YARC Secondary. The vocabulary component of the TEM-4 will be assessed in a similar way against two widely used vocabulary tests. It is important to note here that vocabulary as well as grammar are assessed together in one component of the TEM-4. Therefore, this component does not assess only one construct, but two. This is not necessarily

a problem, however, as many researchers assume with Halliday (1994, p. 14) that "grammar and vocabulary are merely different ends of the same continuum."

Next, as some studies have shown that vocabulary depth is more important than vocabulary size for reading comprehension, while other studies found the opposite, the second objective is to investigate to what extent these two dimensions of vocabulary knowledge can explain unique variance in reading comprehension in our study.

The following two research questions have guided our investigation:

RQ1: To what extent do the reading and vocabulary components of the TEM-4 tap into the constructs they are intended to measure?

RQ2: To what extent do vocabulary size and vocabulary depth explain unique variance in reading comprehension as measured with the TEM-4 and the YARC Secondary?

5. Methods

The participants in this study were 60 second-year English Major (Education) undergraduate students who were studying at a university in the North of China. The students' ages ranged from 18 to 22, and their first language was Mandarin. A brief questionnaire revealed that the students rarely spoke English after class and communicated only a few times a year with (near-) native speakers. All 60 students took the following three tests: the TEM-4, the VST, and the

VKS, and a subsample of 30 students were also administered the YARC (see further down in this section).

The TEM-4 contains six different parts: dictation, listening, cloze, vocabulary, reading, and writing, each of which is described in detail in Jin and Fan (2011). For the purposes of the current study it is important to know that the vocabulary and grammar component of the version of the TEM-4 we used (a past paper from 2011) consisted of 30 multiple choice test for which respondents need to recognise the meaning of a target word out of four options. Half a point was given for each correctly answered question, which means the maximum score was 15. The reading comprehension part of the TEM-4 contained four passages with 20 multiple-choice questions in total. Each of the four passages had a different theme (see Appendix A for an example). An analysis of the vocabulary in the texts revealed that 95% coverage of the texts was achieved at the 6k level. This means that the texts were probably relatively difficult for the students, although some rare words were translated with glosses in the text. Reliability of the individual components could not be computed as only total scores for each component were provided by the school. Students' results for the TEM-4 were obtained from the school administration.

The YARC consists of three parts: in Part 1 decoding is measured, in Part 2 reading comprehension and in part 3 reading fluency. For Part 1, the Single Word Reading test (SWRT), students need to read 70 single words aloud. One point is awarded for each word read correctly. An analysis with Vocabprofile showed that in the first half of the test all words except one (*yawned*) belong to the highest three frequency levels. The second half, however, contained words in frequency levels up to 13k. While this means that the test is easier than the NART

(Nelson, 1982), which contains words up to the 20k level, the second half of the SWRT is likely to be difficult for students.

Part two assesses Reading Comprehension. For this part there is a choice of two levels for the reading fluency passages (Level 1 and Level 2). For this study, Level 1 was chosen because a pre-test revealed this was the more appropriate one for the target group. As we expected the students to have relatively low levels of vocabulary, we chose two passages from level one: the School Boy (fiction) and Honey for You, Honey for me (non-fiction). A Vocabprofile analysis of these two stories revealed that 95% coverage was reached at K4. These texts were therefore likely to be a little easier for the students than the reading texts of the TEM-4, at least as far as the vocabulary is concerned. Analyses of the readability of the texts confirm this. We used the Flesch Reading Ease score, which is based on the number of words per sentence and the number of syllables per word, and found a score of 60.1 for the TEM-4 texts, while the YARC texts obtained 79.9. As texts for which a score of 30 is given are considered difficult and those which obtain 70 easy (Stajner, Evans, Orăsan & Mitkov, 2012), these results suggest that the YARC texts were easier than the TEM-4 ones. The results of the Flesch Kincaid readability scores, which are a simplified version of the Flesch Reading Ease score, and indicate US grade levels, point in the same direction: 8.5 for the TEM-4 and 5.5 for the YARC. According to the data provided by Stajner et al. (2012), this means that the TEM-4 readability scores are closer to those for news texts and the YARC ones closer to fictional texts. Both the lexical and the readability indices therefore show that the YARC texts were simpler than the TEM-4 texts. As one reviewer points out, readability indices present only one aspect of the difficulty levels of a text: the comprehension questions for a text may be easy for a difficult text and vice versa.

However, comparing the difficulty of the questions is unfortunately beyond the scope of the current study. For each text students had to respond to thirteen comprehension questions, each worth one point, and for the summarization part of each story eight and nine points could be obtained. In addition, they were required to summarize the texts.

Part 3 assessed reading fluency. The passage students read contained 137 words, and one point was awarded for each word read correctly (reading fluency). The time needed to read the passage (reading rate) was also recorded.

As the YARC Secondary had to be administered on a one-to-one basis, and it was not feasible to test all students with the available means, 30 of the 60 students were randomly selected and administered the YARC Secondary Test. Each tutor assessed ten students at different times over a period of two months. For the analysis, we did not make use of the ability scores because the students' ages were higher than those of the group for which the test was developed, and the students were classroom learners of English with relatively little contact with day-to-day English. The ability scores and their associated norms as found in the manual are therefore unlikely to be appropriate for the sample in the current study.

As the test was intended for UK-based students, we were interested in obtaining teachers' and students' opinions about the YARC too. Therefore interviews were held with a small sample of students and the classroom teachers involved in the study.

Both vocabulary tests that were used in this study are widely used with adult L2 learners. The bilingual Mandarin-English version of the VST was chosen to ensure students were able to understand the answer options. Considering the students' relatively low vocabulary levels, only the first eight levels of the fourteen levels in the VST were used in this

research. Thus, the maximum score was 80. Students' vocabulary sizes (word families) were computed by multiplying the scores with 100, as suggested in Nation's (2012) test specification for the VST. As the reliability of the VST was a little low (Cronbach's Alpha = .671), we decided to leave out the third level of the VST, which led to an improvement of the reliability (Cronbach's alpha = .712). The VST was administered to all participants during class time. Students also filled in a brief background questionnaire about their language learning history and personal background.

Brown's (2008) slightly simplified version of the VKS was used to assess vocabulary depth. No points were given when students ticked level 1 ("I don't know this word") or level 2 ("I have seen this word but I don't know what it means"), because recognition of the form of the word was not assessed at level 2, and therefore information provided by students could not be verified. The difference between "I think this word means X" (level 3) and "I know this word and it means X" (level 4) was considered to be indicative of students' confidence rather than their actual degree of knowledge, and therefore level 3 was not used. One point was given for receptive knowledge (ability to translate the word) and one for productive knowledge of a word (ability to use the word in a sentence), even if there was a spelling or grammar error in the answer. In total 20 words randomly selected from the 1K until the 8K levels of the VST were included in the VKS (see Appendix B). The maximum number of points that could be obtained was therefore 40. Participants were all given the VKS in class at the same time.

Before carrying out any further analyses, we investigated whether the scores on the different tests were normally distributed. No significant differences were found with the normal distribution for any of the test results. No floor or ceiling effects were found.

6. Results

In Section 6.1, we will first give an overview of the descriptive results for all tests. This will include an analysis of the suitability of the YARC Secondary for adult Chinese L1 learners of English. In 6.2 the correlations between the different components of the TEM-4, the YARC and the vocabulary tests will be discussed (RQ1) and, finally, we will look into the dimensions of vocabulary knowledge which can predict reading comprehension as measured with the TEM-4 and the YARC (RQ2).

6.1 Descriptive results

Students' total mean scores on the TEM-4 were 63.6 (SD 10.5), with a minimum of 37 and a maximum of 85. This means that, on average, students obtained a pass mark for the test, as scores between 60 and 69 are a "pass" (Jin & Fan, 2011). However, one third of the students in this group obtained a mark below 60. Students' English language levels are therefore likely to be relatively low. For vocabulary and grammar the mean score was 8.88 (SD 3.1), which means that students answered 59% of the questions correctly. For the reading component, the mean score was 13.81 (SD 3.31). As 69% of the answers were correct, for this part of the TEM-4 students therefore obtained slightly better scores.

The results for the VST show that students obtained a mean score of 41.5 (SD 7.3) out of

70 items (without level 3 which had to be deleted for reasons of reliability). The minimum score was 25 and the maximum 63. The total number of word families known by students was therefore on average around 4,000, although ten students had vocabularies smaller than 4,000 word families. Figure 1 reveals that students' performance decreased at the lower frequency levels, so that at the K6 and K7 levels they knew only half of the items, and at the K8 level they were just above chance level.

======== Figure 1 approximately here ========

For the VKS, the mean score was 20.1 (SD 4.4) and the minimum and maximum scores were 7 and 36. This means that the students knew on average on half of the items in the test, and 42% of the students knew less than half of the items.

The results for the different components of the YARC are given in Table 1. The scores reveal that students obtained around 50% on most components, except for reading fluency, where they obtained almost full marks. The lowest scores were given for the reading comprehension part.

========= Table 1 approximately here ===========

As the Vocabprofile analyses and the readability indices indicate that the texts were relatively easy, certainly by comparison with the TEM-4 texts, the texts are unlikely to have been too

difficult for the students, except for the ones whose vocabulary contained less than 4,000 word families. Instead, the reading comprehension part may have been particularly difficult for students because they were not familiar with inferential questions. The reading comprehension component of the TEM-4 version we used contained mainly questions which assessed comprehension of information which had been provided in the text, and inferential skills were hardly assessed. In addition, students may not have been familiar with the question format, as the comprehension questions were open questions rather than multiple choice. The students' results are also low by comparison with the raw scores for secondary school pupils in the UK. In a large scale study among students in state schools in the UK, in which 8.2% of students were known to have English as an Additional Language, Stothard, Snowling and Hulme (2011) report that on the SWRT year 7 students (N =178) obtained mean scores of 47.88 (SD 9.20), with values ranging from 18 to 67. For the other components only the standardized scores are reported, so that a comparison with our sample is not possible.

That students were struggling was confirmed in interviews held with teachers and the students after the completion of the tests. Teachers reported that most students were able to accurately read the first 40 words of the SWRT, but struggled with the last 30 words. This is not surprising as we had already seen that the first 35 words on the list belonged to the highest three frequency levels, but in the second half words up to 13k were included. While this component was considered difficult, the teachers confirmed that the reading fluency test was relatively easy for the students, which was also clear from the high scores on this part of the test.

Students who were interviewed mainly pointed to problems with listening comprehension,

but the second most common problem the students encountered related to their grammar mistakes and limited vocabulary knowledge. While one student reported not knowing some keywords in the sentence, which made it difficult to understand the meaning of the entire sentence, the comments of another student pointed in the direction of her problems with integrating the information at sentence-level despite knowing the meaning of the words: “I know most of words, but I still not sure the meaning of the sentence”.

In light of the above, it is likely that the YARC Secondary was probably rather difficult for the students in the current sample for students with low vocabulary levels and because of their lack of familiarity with inferential questions. However, the vocabulary in the stories was not too difficult, as the Vocabprofile analyses have shown. In fact, the vocabulary in the stories was simpler than those in the TEM-4 and the readability indices confirmed this too (see Section 5).

6.2 Correlations between the TEM-4, the vocabulary tests and the YARC Secondary

The first aim of our study was to investigate to what extent the reading and vocabulary components of the TEM-4 tap into the constructs they are intended to measure. To enable us to answer this question we will first analyse the correlations between the different tests.

Table 2 gives an overview of all Pearson correlations between the variables. It reveals that the TEM-4 (total scores) correlates most strongly with the VKS (.521**), and the VST (.420**), but among the variables associated with the YARC only reading rate correlates significantly with the TEM-4 (-.371*). For the reading component of the TEM-4 the same picture emerges: significant correlations are again found only with the vocabulary tests (both around .352**),

although these are slightly less strong than the correlations with the overall TEM-4 scores. Interestingly, there is also a moderate correlation between the reading component of the TEM-4 and the grammar and vocabulary component of this test (.369*). As might be expected, the latter also correlates with both vocabulary tests. Again the correlations are slightly stronger with the VKS (.354**) than with the VST (.270*). Finally, there are some correlations between the different components of the YARC, which are less relevant for the aims of the current study.

==========Table 2 approximately here ========

In summary, these results show that the reading component of the TEM-4 is strongly related to students' vocabulary knowledge, and in particular to vocabulary depth. The absence of significant correlations between the YARC variables and the reading comprehension component of the TEM-4 is worrying, and makes the reader wonder if the TEM-4 reading comprehension component really taps into this construct. As one reviewer points out, it is also possible that the TEM-4 reading component measures different aspects of reading than the YARC. The fact that the *overall* scores on the TEM-4 (rather than the reading component on its own) correlated moderately but significantly with reading rate means that those who obtained higher scores on the TEM-4 are faster readers than those who obtained lower scores. In other words, the TEM-4 does indeed tap into one of the dimensions of reading that the YARC measures too, namely reading rate. However, the TEM-4 reading component, which is labelled "reading comprehension", targets the same construct as the YARC Secondary reading

comprehension task. The lack of correlations between these two tests therefore raises questions regarding the construct validity of the TEM-4 reading comprehension component.

As for the dimensions of vocabulary that are most relevant for reading, in the current study it appears to be the case that vocabulary depth is more important than vocabulary size as the TEM-4 correlates more strongly with the VKS than with the VST. In addition, only the VKS correlates significantly with different dimensions of the YARC (reading fluency and reading comprehension). It is possible of course, that the scores on the VST are not representative of students' actual vocabulary knowledge because of the guessing factor that might lead to an overestimation of students vocabulary size (Gyllstad, et al, 2015). Because the multiple choice format is not used in the VKS, it might present a more valid picture of their vocabulary knowledge. In the next section we will look into whether both measures make a unique contribution to reading comprehension.

6.2 The contribution of the VST and the VKS to reading comprehension as measured with the TEM-4 and the YARC Secondary

The second objective of the current study was to establish what extent the two vocabulary tests explain unique variance in reading comprehension as measured with the TEM-4 and the YARC. We first ran a hierarchical regression analysis with the TEM-4 reading component as the dependent variable, and the VKS and the VST as predictor variables. In the first model we entered the VKS in the first step and the VST in the second step. The β value for the VKS was .237 and for the VST it was .231. The overall model was significant ($F(2,57) = 5.65$, p

< .006) and the VIF and tolerance values were within acceptable ranges. Together these variables explained 16.6% of the variance in reading. The changes in R^2 were .126 for the VKS and .040 for the VST, which means the VST explained unique variance in reading over and above the contribution of the VKS. When the order of the entry of the predictors was reversed, the model was virtually identical: the β value for the VST was .231 and for the VKS it was .237. The changes in R^2 were .124 for the VST and .041 for the VKS.

As the TEM-4 contains a grammar and vocabulary component too, we also attempted a model which used this variable in addition to the VKS as a predictor for the reading component of the TEM-4. This model turned out to be significant ($F(2,57) = 6.848$, $p = 0.002$). Again multicollinearity values were within acceptable limits. The addition of this variable led to an increase in explained variance (19.4%), with β values of 0.256 for the VKS and 0.278 for the TEM-4 Grammar and vocabulary predictor. That this variable explained additional variance in the model may in part be explained by the fact that it covers not only vocabulary but also grammar. After entering the TEM-4 Grammar and vocabulary, the VST was no longer a significant predictor in this model.

Thus, on the basis of the models presented above, we can conclude that vocabulary size and vocabulary depth both explain unique variance in reading as measured with the TEM-4. As the total explained variance is relatively low, we wondered whether other variables, such as reading rate or reading fluency, as measured with the YARC, would explain additional variance but that was not the case.

We subsequently ran regression models with the YARC fluency score as the dependent variable. Recall that we did not compute the ability scores as the group differed in age from

the group for which the test was created. The VKS turned out to be the only significant predictor of fluency (F (1,28) = 4.89, p = 0.35). It explained 14.9% of the variance in fluency (β =.386). The VST was not a significant predictor of fluency, neither on its own, nor in combination with the VKS. Adding reading rate (or the grammar and vocabulary component of the TEM4) to the model did not bring about a significant change in R^2.

When the reading comprehension score of the YARC was used as the dependent variable, the result was very similar. Again the VKS was the only significant predictor of reading comprehension (F(1,28(= 4.70, p = 0.039). The VKS explained 14.4 percent of the variance in reading comprehension (β = .379). Adding reading fluency to the model brought the total explained variance to 24.1% (F (2,27) = 4.29, p =0.024), because a further 9 percent of variance was explained by reading fluency. The addition of any other variables could not improve the model.

The results of the regression analyses confirm not only that vocabulary is an important predictor of reading ability as measured by both tests, but also suggest that the TEM-4 and the YARC measure different aspects of reading ability. For reading as measured with the TEM-4, vocabulary size and vocabulary depth play an approximately equally important role, but for the YARC it is only vocabulary depth that matters and not vocabulary size. The findings from the regression models based on the YARC therefore confirm the findings of Qian (2005) who emphasized the importance of vocabulary depth for reading comprehension. In addition, reading fluency was a significant predictor of reading comprehension as measured with the YARC. The model based on the YARC data therefore supports the view of Grabe (2010) that reading fluency is strongly linked to reading comprehension.

A limitation of the current study was that we used a 2011 version of the TEM-4 as this was the only one available to us. A limitation of the analyses was that other variables which are known to affect reading comprehension, such as phonological or morphological awareness, working memory and non-verbal cognitive abilities were not measured. If these had been included, more variance could have been explained.

7. Summary and conclusion

The current paper set out to investigate the validity of the TEM-4 reading and vocabulary components in a study among 60 English major students between the ages of 18-22 from a university in Northern China. To investigate the concurrent validity of the test, students also took two vocabulary tests, the Vocabulary Size Test (Nation & Beglar, 2007) and a modified version of the Vocabulary Knowledge Scale (Brown, 2008). In addition, 30 students were administered the YARC reading comprehension Secondary (Snowling et al., 2009). We found that the YARC Secondary was rather difficult for the students in China, but this was not so much related to their smaller vocabularies (as the vocabulary and the readability of the texts showed they were easier than those in the TEM-4). Instead it is likely that students were not familiar with inferential questions and the test format of some of the other components. Nevertheless, the results of the YARC Secondary should not be completely discarded: first of all, for one component (reading fluency), the results were very promising, and there were no floor or ceiling effects in any of the components. Second, there were moderate correlations between reading comprehension and reading fluency, as predicted by Grabe (2010). Finally, there were moderate correlations between reading comprehension and the VKS, which

confirms findings of Qian (2005). The existence of these correlations lend support to the assumption that the YARC Secondary did provide useful information about students' abilities, despite the fact that they found some parts rather difficult.

There were also moderate to strong correlations between the reading component of the TEM-4 on the one hand, and vocabulary size as measured with the VST and vocabulary depth as measured with the VKS on the other hand. It was rather unexpected that the TEM-4 reading component did not correlate with any of the components of the YARC Secondary. This pattern of correlations leads us to the conclusion that the TEM-4 reading component mainly taps into vocabulary knowledge, and there is little evidence that this component measures dimensions of reading as distinguished in the YARC Secondary.

In the final part of our study we looked at whether or not vocabulary size or vocabulary depth explained unique variance in reading as measured with both tests. We found that the VST, the VKS and the TEM-4 grammar and vocabulary component all explained unique variance in the TEM-4 reading scores. In total, almost 20% of the variance could be explained.

For the YARC Secondary a different model emerged. The VST was not found to be a significant predictor of the YARC Secondary reading fluency component, nor of the reading comprehension component. The VKS, by contrast, was a significant predictor of both fluency and comprehension, and explained on its own between 14 and 15% of the variance. The addition of reading fluency to the model meant that an additional 10% of the variance in comprehension could be explained.

Overall we conclude that if the reading component of most recent versions of the TEM is similar to the reading component of the TEM-4 we used (which dated from 2011), it is in need of an overhaul. The construct validity of the test should be improved as it is not clear on which model of reading it is built, which is of crucial import in the process of test development (Weir, 2005). The fact that students' inferencing ability is hardly assessed in the test means that it taps virtually only into students' ability to provide answers to literal meanings that are found in the texts. While it is therefore likely that the test underrepresents the construct of reading, a possible way forward in this would be to obtain further clarification about the meaning of inferencing and its role in the TEM-4. The low scores obtained by students on the test are worrying: these could be due to the texts containing many low frequency items and to the complexity of the texts as measured with the readability indices, as our analyses have revealed. However, as the questions are all multiple choice, it is likely that scores are inflated because of guessing. The use of alternative formats, such as a gap filling task which is not based on multiple choice, can help to reduce construct-irrelevant variance (in particular students' ability to guess). In this context it is important to note that the VKS, which does not make use of multiple choice questions, turned out to be a stronger predictor of students' reading ability than the VST, the scores of which may also have been in part the result of students' guessing (see Gyllstad et al., 2015). While for the current group the YARC worked reasonably well, this does not mean that it can be used with any group of adult L2 learners. Prior to using the test with adult L2 learners, students' vocabulary levels should be carefully checked against the vocabulary that is used in the texts. In addition, staff and students would need to receive training in formulating and answering inferential

questions. We hope that the findings of the current study have provided useful information for test developers in China and that it can inform further discussions on the ways in which reading is measured in the TEM-4, and other tests that are widely used in China.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27(1), 101–118.

Binder, K.S., Cote, N.G., Lee, C., Bessette, E. & Vu, H. (2017). Beyond breadth: the contributions of vocabulary depth to reading comprehension among skilled readers. *Journal of Research in Reading*, 40(3), 333-343. DOI:10.1111/1467-9817.12069.

Bowyer, Crane, C. & Snowling, M. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology* 75, 189 189-201.

Brown, D. (2008). Using a modified version of the vocabulary knowledge scale to aid vocabulary development. *The Language Teacher* 32(12). No page numbers.

Brown, J.A., Fishco, V.V. & Hanna, G. (1993). *Nelson-Denny reading test: Manual for scoring and interpretation, Forms G & H*. Rolling Meadows, IL: Riverside Publishing.

Chen, X. (2009). *The validation study of the objective items in TEM4*. Shanghai: Fudan University Press.

Cohen, L., Manion, L. & Morrison, K. (2018). *Research methods in education*. Oxford: Routledge.

Gough, P. B., &Tunmer, W. E. (1986).Decoding, reading, and reading disability. *Remedial and Special Education, 7(1),* 6-10.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Grabe (2010). Fluency in reading: thirty-five years later. *Reading in a Foreign Language* 22 (1)|, 71–83.

Graesser, A., McNamara, D., & Louwerse, M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). New York: Guilford Press.

Gyllstad, H., Vilkaitė, L. & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats. Issues with guessing and sampling rates. *International Journal of Applied Linguistics* 166:2 (2015), 278–306. doi 10.1075/itl.166.2.04gyl

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). London: Arnold.

Jin, Y. & Fan, J. (2011). Test for English Majors (TEM) in China. *Language Testing* 28(4) 589–596.

Khalifa, H. & Weir, C.J. (2009). *Examining reading. Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436. DOI: 10.1111/j.0023-8333.2004.00260.x

Laufer & Ravenhorst-Kalovski, G.C. (2010). Reading in a Foreign Language April 2010, Volume 22(1), 15–30 Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons'responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. DOI: 10.1037/0003-066X.50.9.741

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, 63(1), 59–82. DOI: 10.3138/cmlr.63.1.59

Nation, P. (2013).*Learning Vocabulary in Another Language*, 2nd Edition. Cambridge: Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9–13.

Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In Schmitt, N. and M McCarthy (Eds.). *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nelson, H.E. (1982). *National Adult Reading Test (NART): Test Manual*. Windsor: NFER-NELSON.

Paribakht, T.S. and Wesche, M. (1996). Enhancing vocabulary acquisition through reading: A hierarchy of text-related exercise types. *The Canadian Modern Language Review* 52(2): 155-178.

Qian D. D. (2005) Demystifying Lexical Inferencing: The role of aspects of vocabulary knowledge. *TESL Canada Journal* 22(2), 34-54.

Read, J. (1993) The development of a new measure of L2 vocabulary knowledge. *Language*

Testing, *10*(3), 355-371.

Ready, R.E., Chaudhry, M.F., Schatz, K.C., Strazullo, S. (2012). “Passageless" administration of the Nelson-Denny Reading Comprehension Test: associations with IQ and reading skills. *Journal of Learning Disabilities*. 46(4):377-84. doi: 10.1177/0022219412468160

Schmitt, N., Jiang, X. & Grabe, W.P. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Learning Journal* 95(1), 26-43.

Schmitt, N., Nation, P. & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching* (first view). https://doi.org/10.1017/S0261444819000326.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55–88.

Snow, C. (2002). *Reading for Understanding. Toward an R&D Program in Reading Comprehension*. RAND Corporation.

Snowling, M.J., Stothard, S.E., Clarke. P., Bowyer-Crane C., Harrington A., Truelove, E., Hulme, C. (2009). *York Assessment of Reading for Comprehension*. London: GL Assessment.

Stajner, S., Evans, R., Orăsan, C. & Mitkov, R. (2012). What Can Readability Measures Really Tell Us About Text Complexity? In L. Rello and H. Saggion (Eds.) *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility* (NLP4ITA). European Language Resources Association (ELRA), Istanbul, Turkey.

Stanovich, K.E. (1991). Word recognition: Changing perspectives. In R. Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of Reading Research* (Vol. 2, pp. 418–452). New York: Longman.

Stothard, S.E. (2010). Identifying reading difficulties in the secondary school years. *Literacy Today, 64*, 34-35.

Wang, Y. & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System* 65, 139-150.doi: 10.1016/j.system.2016.12.0

Weir, C. (2005). Language testing and validation: An evidence-based approach. Houndmills: Palgrave MacMillan.

Xu, Q. & Liu, J. (2018). *A study on the washback effects of the Test for English Majors (TEM).*

Springer.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing, 25* (4) 521–551.

Yu, G. & Jin, Y. (2014). English language assessment in China: policies,practices and impacts, *Assessment in Education: Principles, Policy & Practice*, 21:3, 245-250, DOI: 10.1080/0969594X.2014.937936

Zhang, B. (2009). FL Vocabulary Learning of Undergraduate English Majors in Western China: Perspective, Strategy Use and Vocabulary Size. *English Language Teaching* 2(3), 178-185.

Zhao, W., Wang, B., Coniam, D. & Xie, B. (2017). Calibrating the CEFR against the China Standards of English for College English vocabulary education in China. *Language Testing in Asia*, 7(5). DOI 10.1186/s40468-017-0036-1.

Zhou, S., Weir, C. J. and Green, R. (1998). *The Test for English Majors Validation Project*. Shanghai: Foreign Language Education Press.

Table 1. Results (raw mean scores) from the YARC Secondary

	Minimum	Maximum	Raw mean (%)	SD
SWRT	25	50	37.37 (53.39)	7.33
Fluency Accuracy score	128	137	133.10 (97.15)	2.43
Fluency Time	52	83	65.87	7.38
Reading Comprehension	4.00	18.00	10.73 (41.27)	3.25
Summarization	4.00	16.00	8.67 (51)	2.47

SWRT = mean raw scores on the Single Word Reading Test (maximum 70)

Fluency Accuracy Score = mean raw scores on the Reading Fluency Items (maximum 137)

Fluency time: number of seconds needed to read the words

Reading Comprehension = mean raw scores on two passages (maximum 26)

Summarization = mean raw scores of two passage summaries (maximum 17)

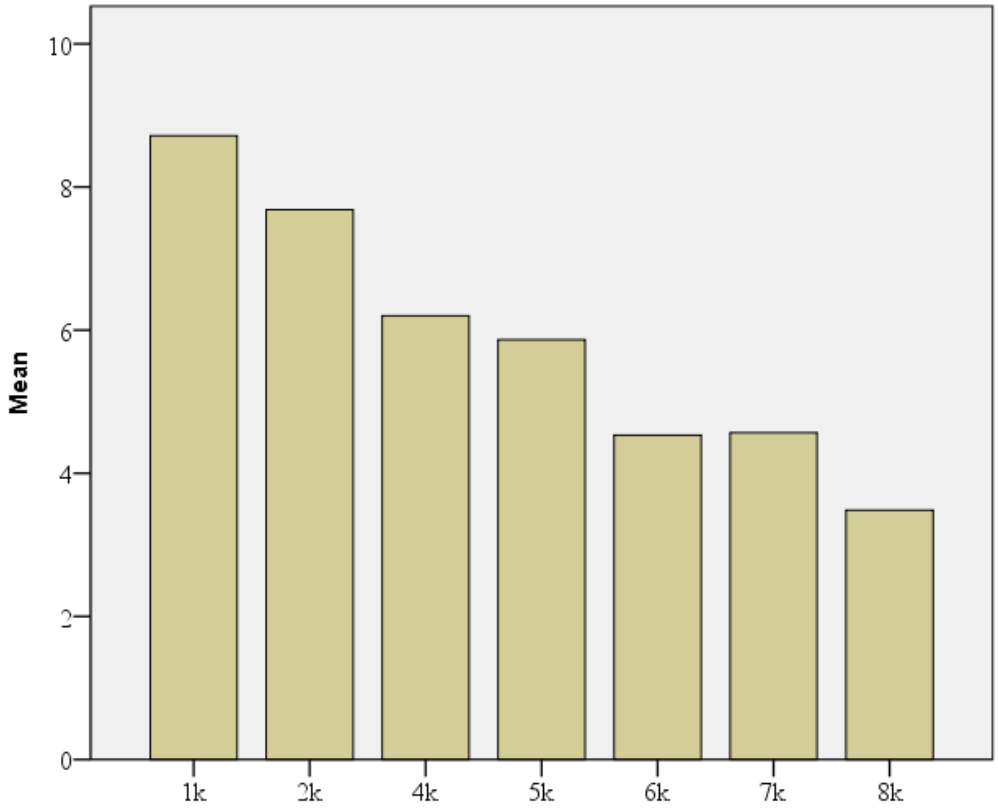
Table 2. Correlations between TEM4, VST, VKS and YARC Secondary

	TEM4 Reading	TEM Grammar and Vocab	VST	VKS	YARC SWRT	YARC Fluency	YARC Rate	YARC Comprehension	YARC Summarisation
TEM4 Total	.690**	.738**	.420**	.521**	0.295	0.265	-.371*	0.071	0.211
TEM4 reading		.369**	.352**	.355**	0.086	0.173	-0.136	0.107	0.012
TEM4 Grammar Vocab			.270*	.354**	0.287	0.141	-0.298	0.040	0.049
VST				.510**	0.340	0.287	-0.219	0.285	0.047
VKS					0.255	.386*	-0.200	.379*	0.157
YARC SWRT						.696**	-0.236	0.311	0.301
YARC Fluency							-.375*	.406*	.369*
YARC rate								0.013	-0.273

YARC comprehension		0.152
YARC summarisation		

* = correlation significant at p < .05; ** correlation significant at p < .01

Figure 1. VST results



Appendix A

TEM-4 TEXT B

I know when the snow melts and the first robins (知更鸟) come to call, when the laughter of children returns to the parks and playgrounds, something wonderful is about to happen. Spring cleaning.

I'll admit *spring cleaning is a difficult notion for modern families to grasp*. Today's busy families hardly have time to load the dishwasher, much less clean the doormat. Asking the family to spend the weekend collecting winter dog piles from the melting snow in the backyard is like announcing there will be no more Wi-Fi. It interrupts the natural order.

"Honey, what say we spend the weekend beating the rugs, sorting through the boxes in the basement and painting our bedroom a nice lemony yellow?" I say.

"Can we at least wait until the NBA matches are over?" my husband answers.

But I tell my family, *spring cleaning can't wait*. The temperature has risen just enough to melt snow but not enough for Little League practice to start. Some flowers are peeking out of the thawing ground, but there is no lawn to seed, nor garden to tend. Newly wakened from our winter's hibernation (冬眠), yet still needing extra blankets at night, we open our windows to the first fresh air floating on the breeze and all of the natural world demanding "Awake and be clean!"

Biologists offer a theory about this primal impulse to clean out every drawer and closet in the house at spring's first light, which has to do with melatonin, the sleepytime hormone (激素) our bodies produce when it's dark. When spring's light comes, the melatonin diminishes, and suddenly we are awakened to the dusty, virus-filled house we've been hibernating in for four months.

I tell my family about the science and psychology of a good healthy cleaning at spring's arrival. I speak to them about life's greatest rewards waiting in the removal of soap scum from the bathtub, which hasn't been properly cleaned since the first snowfall.

"I'll do it," says the eldest child, a 21-year-old college student who lives at home.

"You will? Wow!" I exclaim.

Maybe after all these years, he's finally grasped the concept. Maybe he's expressing his rightful position as eldest child and role model. Or maybe he's going to Florida for a break in a couple of weeks and he's being nice to me who is the financial-aid officer.

No matter. Seeing my adult son willingly cleaning that dirty bathtub gives me hope for the future of his 12-year-old brother who, instead of working, is found to be sleeping in the seat of the window he is supposed to be cleaning.

"Awake and be clean!" I say.

86. According to the passage, "*...spring cleaning is difficult notion for modern families to grasp*" means that spring cleaning

A. is no longer an easy practice to understand.

B. is no longer part of modern family life.

C. requires more family members to be involved.

D. calls for more complicated skills and knowledge.

87. Which of the following is LEAST likely to be included in family spring cleaning?

A. Beating the rugs.
B. Cleaning the window.
C. Restoring Wi-Fi services.
D. Cleaning the backyard.
88. Why does the author say "*spring cleaning can't wait*"?
A. Because there will be more activities when it gets warmer.
B. Because the air is fresher and the breeze is lighter.
C. Because the whole family is full of energy at spring time.
D. Because the snow is melting and the ground is thawing.
89.Which of the following interpretations of the biologists' theory about melatonin is INCORRECT?
A. The production of melatonin in our bodies varies at different times. B. Melatonin is more likely to cause sleepiness in our bodies.
C. The reduction of melatonin will cause wakefulness in our bodies.
D. The amount of melatonin remains constant in our bodies.
90. Which of the following can best sum up the author's overall reaction to her adult son's positive response to spring cleaning?
A. Surprised and skeptical.
B. Elated and hesitant.
C. Relieved and optimistic.
D. Optimistic and hesitant

Appendix B: words randomly selected from the VST
accessory
allege
compost
compound
deficit
devious
drawer
drive
hallmark
haunt
jug
latter
maintain
olive
soldier
standard
strangle
threshold
upset
yoghurt