# The effects of phonics instruction on L2 phonological decoding and vocabulary learning:

## An experimental trial on Chinese University EFL learners

Sha Li

Exeter College / Department of Education

Thesis submitted for the degree of DPhil, University of Oxford

Hilary Term, 2019

**Acknowledgments**

I would like to say a big thank you to the following people for their generous support during my DPhil study.

Firstly, I owe a huge debt of gratitude to my supervisors, Dr. Robert Woore and Dr. Catherine Walter. Their insights were invaluable and inspirational in so many ways, which greatly helped me develop my thinking. I have been extremely fortunate to be their student, and they are the kind of teachers I aspire to become. Thank you, Robert, for taking me in as your student, for sharing your expertise in the field, for offering me as many opportunities as possible to help pave the way for my future, and for always trusting me even when I doubted myself. Thank you, Catherine, for your wonderful ALSLA lecture on reading comprehension which inspired me to pursue this topic in the first place, and for your patience, generosity and warm encouragement throughout these years.

Secondly, my heartfelt thanks goes to all the students and teachers involved in this project, and their department for offering me this precious opportunity to work with them over a term. I would like to thank my students for putting up with a novice teacher like me, and for the many wonderful and joyful conversations. I am also very grateful for the valuable advice given by the teachers who observed my classes.

Thirdly, I am deeply indebted to my parents, for their unconditional love, support and encouragement throughout the years. Thank you for always having faith in me. I could never have completed this project without you.

Finally, I want to thank my husband, for always being happy to be the first one that I piloted my lectures and testing materials on, and for helping me in every way possible to finish this project. Thank you for your love, patience and understanding. I am very lucky to have found you as a participant in my master study, and as a partner from then on.

I will always cherish the years I spent in Oxford, where I met so many wonderful people and learned so much from them. This memory will always be held dear to my heart, and be an eternal source of encouragement throughout my future endeavors.

**Abstract**

Phonological decoding, defined as the ability of using the systematic knowledge of a language's grapheme-phoneme correspondences in order to generate pronunciation based on orthographic context, has been found to be facilitative of the acquisition of reading comprehension in both first language (L1) and second language (L2). Moreover, research into L1 learning has consistently demonstrated a strong relationship between decoding proficiency and vocabulary acquisition. Recently, the relationship between L2 decoding and word learning has begun to be explored. There is a strong theoretical support for a causal relationship between the variables: fast and accurate decoding of written forms provides reliable phonological representations which support the operation of phonological working memory; psycholinguistic evidence suggests that this, in turn, plays a central role in learning novel phonological forms. Further, knowledge of a language's grapheme-phoneme correspondences allows the orthographic and phonological representations of new words to be mutually reinforcing. Previous studies have indeed found positive correlations between both the speed and accuracy of decoding on the one hand and success in intentional word learning on the other amongst learners with alphabetic but not morphemic L1 backgrounds. This is consistent with the view that morphemic learners are more likely to process words visually as whole units. However, there has been no experimental evidence thus far, which could prove the causal link between decoding and vocabulary learning.

Against this backdrop, an intervention study was conducted in which a twelve-week programme of systematic phonics instruction, covering 101 English grapheme-phoneme correspondences, was implemented to three classes of first-year English majors in three universities in Wuhan, China. The comparison participants received a twelve-week English phonology instruction programme focusing on the pronunciation of 44 English phonemes and other pronunciation tips, but not explicit instruction on English grapheme-phoneme correspondences. To evaluate the effectiveness of the phonics instruction programme, two tests were administered both before and after the intervention: (a) an English decoding test; and (b) a vocabulary memorisation task, followed by immediate recall and recognition tests of the phonological and written forms of the new words. Participants' scores at the two time points and between the two groups were compared.

Participants who followed the phonics instruction programme demonstrated a clear and significant advantage over their counterparts in the comparison group in terms of the number of both (a) whole words and (b) individual graphemes that they pronounced correctly in the English decoding post-test. Moreover, comparison of participants' realisations of English graphemes before and after the instruction programme suggests the influence of cross-linguistic transfer. For instance, the English graphemes that also exist in Pinyin and share the same pronunciations witnessed the highest accuracy at t1, while those that exist in Pinyin but have different

pronunciations in the two languages appeared to be those associated with least improvement at t2. In the vocabulary memorisation task, the intervention participants achieved significantly higher scores at t2 in the oral recall, written recall and aural recognition test compared to the comparison groups. However, no significant differences between the two groups were observed in the written recognition test.

The results suggest that explicit phonics instruction can be effective in improving the English decoding proficiency of Chinese university EFL learners. Further, the findings are consistent with the hypothesis of a causal link between L2 decoding and vocabulary learning. The study may also inform the design of future phonics instruction programmes for this population of learners: for example, certain English graphemes appeared to need more explicit instruction than others.

**Contents**

## List of tables

## List of figures

## List of abbreviations

| | |
|---|---|
| BPVS | British Picture Vocabulary Scale |
| GPC | Grapheme-phoneme correspondences |
| L1 | First language |
| L2 | Second language |
| NCEEE | National College Entrance English Exam |
| SLA | Second language acquisition |

# Chapter 1. Introduction

## 1.1 The role of phonological decoding in L2 learning

Phonological decoding, defined as 'the process of converting the written symbols (or graphemes) of a language into the sounds (or phonemes) they represent, using knowledge of the language's symbol/sound correspondence' (Woore, 2009: 3), has long been argued to underlie various aspects of L1 learning, particularly reading and spelling, of languages written with alphabets or syllabaries. Over the past two decades, phonological decoding has also attracted enthusiastic attention from the research community of Second Language Acquisition (SLA) (Koda, 2007). This is not difficult to understand: though for literate learners, language learning involves constant processing in both symbol-to-sound and sound-to-symbol directions (Erler & Macaro, 2011), L2 learners generally (at least instructed/ classroom learners) have more exposure to L2 written input compared to aural input (Schmitt, 2010). Thus, reliable phonological representations of the written input, intuitively, should play an important role in much L2 learning.

In the past, phonological decoding has mainly been argued to play an important role in L2 reading comprehension (e.g. Koda, 2007; Walter, 2008). The past four decades have seen a growing emphasis on this lower-level processing in L2 reading, which is argued to interact with higher-level processing such as employing metacognition and

schema knowledge in order to achieve comprehension (Nassaji, 2014). As learners' attentional resources are limited (Perfetti & Lesgold, 1977), decoding should ideally be efficient and automatic, thus freeing up more resources for higher-level processing. Indeed, research has consistently demonstrated that decoding proficiency makes a unique and important contribution to L2 reading comprehension when other relevant variables are controlled for (Nassaji, 2014; Grant, Gottardo & Geva, 2011). Moreover, such a contribution appears to hold for L2 learners with various L1 writing systems, including both sound-based writing systems like Spanish and Portuguese (Grant, Gottardo & Geva, 2011), and meaning-based writing systems like Chinese (Geva & Wang, 2001).

In recent years, some studies have also begun to explore the possible contribution of decoding to L2 vocabulary learning (e.g. Hamada & Koda, 2008, 2010). One can reasonably see why there might be a causal relationship between decoding and vocabulary learning in L2. As learners' attentional resources are limited, fast and accurate decoding of written words provides reliable phonological representations, and frees up more resources to deal with other aspects of word learning. Moreover, phonics knowledge allows the orthographic information and the phonological information of new words to be mutually reinforcing. However, research in this area is still limited, and this calls for more empirical evidence to explore this potential causal relationship.

Some studies (e.g. Erler & Macaro, 2011; Erler, 2003, 2004) have also looked at the potential influence of decoding on L2 learners' self-efficacy and motivation regarding learning the language. For instance, Erler and Macaro (2011) demonstrated that the decoding proficiency of English secondary school learners of French accounted for approximately 10% of the variance in the decision to continue with French learning. Considering the many other possible variables contributing to this decision, this figure appears to be high, which again points to the importance of acquiring decoding proficiency in L2 learning.

Given the important contribution of L2 decoding to different aspects of L2 learning, it is therefore worth contemplating whether L2 decoding proficiency can be promoted with the help of systematic phonics instruction. Currently, studies in this area are still limited in number. However, a few existing studies on L2 phonics instruction (e.g. Woore, 2011, Woore, Graham, Porter, Courtney & Savory, 2018) have provided encouraging evidence for the effectiveness of such instruction on different aspects of L2 learning, including decoding proficiency, reading comprehension and vocabulary knowledge. However, none of these instruction programmes was conducted with Chinese EFL learners (or indeed with learners with a non-alphabetic L1). As a result, it still remains unknown whether a systematic English phonics instruction programme can be of help in promoting decoding proficiency and other aspects of English learning for Chinese EFL learners.

Even so, there seems to be a growing interest in English phonics learning in China in the past few years. In China, phonics, or *zi ran pin du fa* (自然拼读法), has been widely advertised by many educational institutions, with the claim that it can help promote various aspects of English learning. These educational institutions generally use English phonics textbooks designed for L1 English speakers, as there are no materials that are specifically designed for Chinese EFL learners. It is hoped that through conducting an intervention study exploring the effects of English phonics instruction for Chinese EFL learners, the question of whether this kind of instruction can be of help in promoting decoding proficiency and vocabulary learning for Chinese EFL learners can be investigated. Moreover, it is hoped that through detailed analysis of participants' pronunciations of unfamiliar English words before and after phonics instruction, Chinese EFL learners' realizations (i.e. pronunciations) of different English graphemes can be studied. This may serve as a useful reference for the design of future instruction programmes.

### 1.2 The role of crosslinguistic influence of L1 on L2 learning

Another enduring topic in SLA is the crosslinguistic influence of L1 on L2 learning. This topic is especially interesting for Chinese EFL learners, given that the Chinese writing system is completely different from that of English, in that it is a meaning-based (often called 'logographic' or 'morphemic') system, whereas English is largely a sound-based ('phonographic') system. Yet, there is also a phonographic

system for Chinese: Pinyin. This uses the same Roman alphabet as English, and is used in most regions of mainland China to assist children in the early stages of learning to read Chinese characters. It is also used by adults as a system for typing Chinese characters on keyboards. Regrettably, few crosslinguistic studies have taken both these systems into consideration when examining the influence of Chinese EFL learners' L1 on various aspects of their English learning. Most of these studies seemed to focus on how the processing mechanisms associated with the morphemic Chinese writing system influence different aspects of English learning. For instance, Hamada and Koda (2008, 2010) found that Chinese EFL learners demonstrated poorer English decoding proficiency compared to proficiency-matched Korean EFL learners, with the possible explanation that Chinese EFL learners were less adept in engaging in intraword analysis due to the morphemic nature of their L1 writing system. Research on Chinese EFL learners' vocabulary learning (e.g. Ma, 2009, Gu & Johnson, 1996, Li, 2012) also seems to suggest that Chinese EFL learners tend to show heavy reliance on using visual-based strategies when learning new English vocabulary.

These studies have provided useful insights into the understanding of the crosslinguistic influence of the Chinese writing system on L2 English learning. However, it remains an under-researched question whether the knowledge of the phonographic Pinyin system also has an impact on various aspects of English learning, especially the decoding of English words. There seem to be valid reasons to

investigate this possible L1 transfer, given that Pinyin not only shares the same alphabet as English, but also shares many of the same graphemes. Studies of English learners' decoding of French words (e.g. Woore, 2009, 2011) have demonstrated the transfer of knowledge of L1 GPCs into the decoding of L2 words. For instance, Woore (2009) found that the overwhelming majority of the English secondary school children in his study decoded the French grapheme <ée> as /i/, pointing to the transfer of knowledge of the English grapheme <ee>. It is hoped that, through detailed analysis of participants' pronunciations of unfamiliar English words, the possible influence of Pinyin GPC knowledge on Chinese EFL learners' English decoding can be examined, which may provide a unique contribution to the already fruitful area of research into the influence of L1 on L2 learning.

## Chapter 2. Literature Review

This chapter firstly reviews research evidence on the importance of phonological decoding in different aspects of L2 learning, with a special focus on vocabulary acquisition. Then, section 2.2 reviews literature on the crosslinguistic influence of L1 on Chinese EFL learners' L2 decoding. In addition to the traditionally accepted influence of Chinese orthographic processing mechanisms, the potential influence of Pinyin is also discussed. Section 2.3 moves on to argue for the importance of phonics instruction for Chinese L2 learners, and discusses what a good phonics instruction programme might look like. Section 2.4 explores the ways in which Chinese EFL learners' vocabulary learning outcomes, their vocabulary learning strategies and their attitudes towards vocabulary learning may be influenced by the processing mechanisms associated with their L1 writing system(s) and their often inadequate English phonics knowledge. It also discusses how the acquisition of vocabulary knowledge, and the learning of English phonology and phonics – two aspects of learning to which Chinese EFL learners seem to attach very different levels of importance (Rao, 2002) – might actually be brought together. Section 2.5 summarises the arguments based on the review of the current literature and lays the ground for the research questions in this study. Finally, the research questions are proposed in section 2.6.

## 2.1 The importance of phonological decoding in learning an alphabetic L2

Knowing a word can be conceptualised on many different levels, but a minimum specification is to establish the link between its core or prototypical meaning and its form. In terms of the form of a lexical item, both its graphological form (spelling or ideogram) and its phonological form (pronunciation) are of importance to the literate learner, so that a word can be recognised and produced in both writing and speech. The link between the graphological form and the phonological form of a word is thus scrutinised by a robust line of research, especially in the research on L1 reading (Goff, Pratt & Ong, 2005). However, this link also underlies other aspects of L2 learning. This section reviews literature on the importance of phonological decoding in L2 learning from three different perspectives. (1) Since decoding has long been argued to be an indispensable component of L1 reading comprehension, this section starts by providing a brief overview of the importance of decoding in L2 reading comprehension. (2) The section then moves on to explore the link between decoding and L2 vocabulary learning, argued to be an under-researched area and also a key focus of the current study. (3) Finally, this section reviews literature examining the role of decoding in some other aspects related to L2 learning, such as motivation.

### 2.1.1 Phonological decoding and L2 reading comprehension

It is generally understood that reading comprehension is a complex cognitive process that involves multiple mental operations. For instance, Koda's review (2007: 4)

summarises three major components that are involved in reading comprehension, namely (a) *word recognition* (or 'decoding' Koda terms it), defined as 'extracting linguistic information directly from print'; (b) *text-information building*, defined as 'integrating the extracted information into phrases, sentences and paragraphs'; and (c) *reader-model construction*, defined as 'synthesizing the amalgamated text information with prior knowledge'. Of these three components, word recognition is argued to be the only component that is specific to reading comprehension, as in the so-called 'Simple View of Reading' (Gough and Tunmer, 1986).

Phonological decoding is generally believed to be a crucial component in word recognition (e.g. Bowers, Golden, Kennedy &Young, 1994; Droop & Verhoeven, 2003; Perfetti, Landi & Oakhill, 2005). The strongest theoretical support for the importance of decoding in word recognition comes from the 'Dual Route' model (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001), widely recognised as the most durable model of reading at the word-access level (Erler & Macaro, 2011). This model proposes that the meaning of a word can be accessed through two routes, namely: (a) a 'lexical route', through which the orthographic information directly evokes the word's meaning; and (b) a non-lexical, 'phonological' route, through which orthographic information needs to be phonologically decoded first. The decoding process triggers the aurally-stored memory of the word, which then evokes its meaning. In other words, poor decoding may harm the efficiency of the non-lexical route, as an incorrect phonological representation of a word makes it difficult to

connect its orthographic information to the already-stored phonological information and thus to its meaning, resulting in problems in word-level reading comprehension.

A series of studies by Walter (2004, 2007, 2008) have provided convincing evidence for the importance of reliable phonological representation of orthographic input in L2 reading. Walter (2004) found that, when reading in L2, lower-intermediate French learners of English had more difficulty than upper-intermediate learners in accessing their existing ability (as demonstrated in L1 reading comprehension) to construct mental representations of texts as a whole, despite understanding the texts well at the individual sentence level. The threshold for access appeared to be linked to L2 working memory. This was especially evident for the lower intermediate learners, for whom small advantages in working memory led to significant advantages in comprehension.

In order to identify the specific role of working memory in L2 reading comprehension, Walter's later study (2008) further investigated one part of working memory, namely the phonological loop, which serves to temporarily store and manipulate verbal information (Baddeley & Hitch, 1974). Two groups of French learners of English with different levels of proficiency in reading comprehension (high / low) were asked to complete a written recall test of visually-presented sequences of words which were either similar to each other (e.g. men / mine; met / might) or dissimilar (e.g. cheese / job; fine / yes), in both English and French. A group of native English speakers also

took the English version of the test as a point of comparison. It was found that the two groups of French learners (the high / low proficiency groups) performed similarly in their L1, recalling the dissimilar sequences somewhat better than the similar sequences. However, for the sequences in L2, the good comprehender group performed very much as they did in L1, and very much like the native English group, while the poor comprehender group showed a different pattern of results. They performed slightly worse than both the native group and the good comprehender group in recalling dissimilar sequences of words, and substantially worse than the other two groups in recalling similar sequences. In other words, when visually presented with words that differed by only one phoneme, regardless of smaller or larger differences in spelling, the poor comprehender group could not recall the words accurately. These results demonstrate that reliable phonological representation of written input may be of crucial importance to reading comprehension.

This led Walter (2008) to argue that a key to promoting L2 learners' reading comprehension is not to teach comprehension skills; rather, classroom time might be better spent on explicitly teaching learners to recognise and differentiate L2 phonemes, which may serve as the basis of reliable decoding. This is because if L2 learners 'have a reliable repertoire of L2 phonemes in long-term memory, they will decode written L2 text into well-differentiated words in the phonological loop, and will be able to use these words for comprehension' (p. 470).

Indeed, longitudinal studies of minority language children's L2 reading comprehension have also demonstrated that L2 decoding proficiency is a strong predictor of their L2 reading comprehension. For instance, Droop and Verhoeven (2003) followed a group of ethnic minority children (Moroccan and Turkish) living in Holland from third grade to fourth grade. Their L2 (Dutch) reading comprehension and other relevant variables, including L2 decoding proficiency, were measured at the start of the third grade, end of third grade and end of fourth grade. It was found that their L2 decoding proficiency was consistently correlated with their L2 reading comprehension, with a medium correlation coefficient of between .39 and .46 found at the three time points. Similar findings were also reported in Geva and Wang (2001), where L2 decoding proficiency was found to be a significant predictor of L2 reading comprehension for English-Hebrew bilingual children, throughout Grade 1 and Grade 2.

In a systematic review of the key component variables in passage-level L2 reading comprehension, Jeon and Yamashita (2014) investigated 10 correlates of L2 reading, namely L2 decoding, L2 vocabulary knowledge, L2 grammar knowledge, L1 reading comprehension, L2 phonological awareness, L2 orthographic knowledge, L2 morphological knowledge, L2 listening comprehension, working memory and metacognition. Sample sizes were weighted and measurement errors were corrected in the meta-analysis in order to ensure fair comparison among the correlates. It was found that L2 decoding was one of the three strongest correlates with L2 reading, with

an overall mean correlation of $r$ = .56 (second only to L2 grammar knowledge, $r$ = .85, and L2 vocabulary knowledge $r$ = .79). The results have again demonstrated the importance of L2 decoding in L2 reading comprehension.

### 2.1.2 Phonological decoding and L2 vocabulary learning

Similarly to reading comprehension, vocabulary acquisition is a complicated cognitive process that depends strongly on the function of working memory. The following section builds on the classic working memory model proposed by Baddeley and Hitch (1974), and updated by Baddeley (2000) to include the episodic buffer, theorized to integrate the other components of working memory, namely the phonological loop and the visuospatial sketchpad. The section reviews literature tapping into the relationship between decoding and vocabulary acquisition from two perspectives, namely the acquisition of phonological patterns and written forms.

#### 2.1.2.1 Decoding and the acquisition of phonological patterns

According to the widely acknowledged working memory model proposed by Baddeley and Hitch (1974), the phonological loop is the component responsible for the temporary storage and manipulation of verbal information. It consists of two subcomponents, the phonological short-term storage of verbal information and a subvocal rehearsal process. Incoming material is represented as sound-based

characteristics in the phonological short-term store. The resulting phonological representations are then subject to rapid decay over approximately two seconds. In the meantime, the subvocal rehearsal process serves to refresh the decaying representations and thus maintain them. In the context of vocabulary acquisition, newly-encountered words can gain access to the phonological loop via two routes. Orally-introduced words gain direct obligatory access to the phonological loop, while alternatively, non-speech inputs such as printed words are converted to phonological representations; and phonological representations from both kinds of sources are rehearsed in the phonological short-term store. Thus, phonological short-term memory is closely associated with vocabulary acquisition regardless of the learning context (i.e. whether words are encountered in spoken or written form). As this model posits, rapidly-fading traces in the phonological short-term store would 'give rise to errors in translation to an articulatory code' (Service, 1992: 44), therefore resulting in difficulty in the acquisition of novel phonological forms. Perhaps more importantly for the current discussion, inaccurately represented phonological forms, or representations of phonological forms that are insufficiently differentiated from those of similar words with different meanings, may cause confusion in the acquisition of novel phonological forms.

The importance of phonological short-term memory in vocabulary learning has been supported by a robust line of research. Studies of L1 vocabulary acquisition have consistently demonstrated a strong correlation between vocabulary achievement and

phonological short-term memory capacity, which has often been indexed by the ability to repeat nonwords (e.g. Gathercole and Baddeley, 1989; Gathercole, Service, Hitch, Adams & Martin, 1999; Gathercole, Willis, Emslie & Baddeley, 1992). The correlation coefficients observed in these studies generally fell within the range of .4 to .6 and were generally independent of nonverbal intelligence, which provided convincing evidence for the importance of phonological short-term memory in vocabulary acquisition. In experimental tasks which involved learning new vocabulary and immediate recall after the learning session, phonological short-term memory capacity was found to be a reliable predictor of the acquisition of the phonological forms of new words (e.g. Baddeley. Gathercole & Papagno, 1998; Gathercole, Hitch, Service & Martin, 1997). Moreover, the influence of phonological short-term memory capacity on the acquisition of novel phonological forms seems to remain strong as people age, as found by empirical studies of participants from different age groups. For instance, Gathercole and Baddeley (1990b) assigned children aged between 5 and 6 to learn the made-up names of toy monsters such as *Piemass* and *Sommel* and found that the group with higher phonological short-term memory capacity needed fewer learning trials to master the phonological form than the group with lower phonological short-term memory capacity (1.24 and 2.06 trials, respectively). Gupta (2003) also found similar results with adult learners, in that a correlation coefficient of .36 ($p < .05$) was observed between participants' phonological short-term memory capacity and their learning results of newly-encountered phonological forms such as *zonolambic* and *volukinster*.

From a different perspective, studies on children with specific language impairment (SLI) also suggest that those with particular difficulty in repeating nonwords and performing other phonological short-term memory tasks also perform poorly in acquiring the sound patterns of new words (e.g. Gathercole and Baddeley, 1990a). For instance, Weismer and Hesketh (1996) studied two groups of participants in a word learning task, one consisting of SLI children and the other of normally-developing children, and found that children with SLI recalled significantly fewer novel phonological forms than the other group, even though the two groups were matched in mental age.

Likewise, research on L2 vocabulary acquisition has demonstrated that individual differences in phonological short-term memory capacity constrain L2 vocabulary acquisition (Speciale, Ellis & Bywater, 2004). For instance, Cheung (1996) found that Hong Kong seventh graders' nonword repetition ability was significantly correlated with their learning results for three new English words, namely *egregious*, *succulent* and *jocular*. A correlation coefficient of -.45 ($p < .05$) was observed between their nonword repetition score and the number of trials needed to learn the pronunciations and Chinese translations of the words.

From a different perspective, a line of study involving articulatory suppression has showed that L2 vocabulary learning is less effective when the phonological loop is impaired. Articulatory suppression is a technique designed to interrupt the function of

working memory, usually by asking participants to repeat simple words or nonwords while performing a memorisation task. An oft-cited study by Papagno, Valentine and Baddeley (1991) has demonstrated that when articulatory suppression was used, Italian native speakers recalled significant fewer Russian words, as did English native speakers in recalling Finnish words. It should be noted that the participants in these experiments were not actually learners of these languages, and can thus be argued to have lower motivation in memorising words in these languages, which might have been reflected in their recall results. However, the finding that they performed significantly worse when the articulatory suppression technique was adopted still showed that the impairment of the phonological loop was associated with less effective L2 vocabulary learning. Similar findings have also been observed in the learning of Welsh words by English-speaking participants (Ellis & Sinclair, 1996).

The above studies have demonstrated the essential role of the phonological loop in both L1 and L2 vocabulary acquisition. As Baddeley et al. (1998) forcefully conclude, the phonological loop serves as an important learning device for the acquisition of novel phonological forms and lays the foundation for long-term learning, '[by storing] unfamiliar sound patterns while more permanent memory records are being constructed' (p. 158).

From a different perspective, inaccurate representations in the phonological loop could harm the efficiency of the acquisition of novel phonological forms, and possibly

result in errors in long-term memory. Based on this, a question that naturally arises is whether the efficiency of the phonological loop could be enhanced to promote the learning of novel phonological forms. According to Baddeley and Hitch's working memory model (1974), all the phonological information is held in the phonological short-term store in the same way, regardless of whether it conforms to familiar or alien sound patterns. In other words, the phonological loop is essentially 'knowledge-free' (Gathercole & Thorn, 1998:145). However, some research evidence has demonstrated that the phonological loop probably also interacts with lexical knowledge that has already been acquired. For instance, in Gathercole's longitudinal study (1995) of phonological memory and vocabulary acquisition, native English-speaking children were tested on their ability to repeat two groups of orally-presented nonwords varying in respect of 'wordlikeness' at age 4 and again at age 5. The wordlikeness of the stimuli was rated by 20 adult native English speakers on a 5-point scale. Examples of words that were scored low in wordlikeness were *perplisteronk*, *bannow*, *glistow*, while those considered high in wordlikeness included *hampent*, *sladding*, *confratually*. Both groups consisted of 14 words. The stimuli were matched for number of syllables as well as length (defined as the number of letters in a word). It was found that children consistently recalled significantly more words that were high in wordlikeness at both times. Moreover, a correlation coefficient of .33 ($p < .05$) was observed between wordlikeness and recall accuracy at the age of 4, and an even stronger correlation was found at the age of 5 ($r = .44$, $p < .05$).

Such findings point to an interesting phenomenon, in that the phonological loop and vocabulary knowledge operate in an interactive manner. In other words, pseudo words that conform to the phonotactic probabilities of the language tend to be easier to learn. Humle, Maughan and Brown (1991) argue that this 'lexicality effect' arises because phonotactic knowledge is used to reconstruct representations in the phonological loop. As a result, an increase in GPC knowledge would facilitate the acquisition of the phonological forms of novel words. This was reflected in the research by Gathercole (2005) cited above, as the mean number of correctly-recalled high-wordlikeness stimuli increased from 8.99 to 10.51 over a year, which could probably be explained by participants' incremental growth in vocabulary knowledge (and potentially better understanding of the phonotactics of the language). By contrast, the mean number of low-wordlikeness stimuli which were correctly recalled showed only a small increase, from 7.11 to 7.79.

Given the importance of the phonological loop in the acquisition of novel phonological forms, it is important that accurate phonological representations of written forms should be secured, to help lay good foundations for long-term learning. This points to the importance of phonological decoding, which is the process of converting written symbols to sounds using knowledge of the language's GPCs (i.e. phonics knowledge).

The relationship between decoding proficiency and L2 vocabulary learning results

was explored in a study conducted by Li (2012), using Chinese advanced EFL learners as participants (N = 60). The participants firstly had a decoding test consisting of 20 unfamiliar low-frequency English words, and then took part in an intentional word-learning session containing 21 English words (presented in written form) and their Chinese translations. After that the participants took three recall tests, namely an L2-to-L1 recognition test, where the participants were presented with English words and asked to write down their Chinese translations; an L1-to-L2 spelling test, where the participants saw the Chinese translations and were required to write down the equivalent English words, and an L2-to-L1 listening test, where the participants listened to a recording of English words and wrote down the Chinese translations. It was found that the participants' decoding proficiency significantly correlated with the results of all three recall tests, but the strongest correlation was found between their decoding proficiency and their L2-to-L1 listening test results ($r = .77$, $p < .05$). In other words, the participants who were better at phonological decoding were also better at recognising the phonological forms of the new words, which had been presented to them in written form. This is presumably because they had decoded the written forms of the new words more accurately in the learning phase, resulting in stored phonological forms which better matched what they subsequently heard in the testing phase. Overall, these results strongly suggest that decoding proficiency plays a crucial role in learning new L2 words, particularly their pronunciations.

Taken together, the research reviewed above has demonstrated that the phonological loop in working memory is closely associated with the learning of novel phonological forms in both L1 and L2. Moreover, the phonological loop functions in interaction with phonotactic knowledge that has already been acquired. Thus, reliable phonological representations of written input in the phonological loop, as well as adequate phonotactic knowledge, can facilitate the acquisition of novel phonological forms in vocabulary learning.

#### 2.1.2.2 Decoding and the acquisition of written forms

The above section has demonstrated that phonics knowledge or decoding is of great importance in the acquisition of the sound patterns of new words, as evidenced by research with both L1 and L2 learners. In contrast, few studies so far have explored the role of phonics knowledge or decoding proficiency in the learning of written forms of new words, which is another essential aspect of vocabulary learning. Yet from the theoretical point of view, decoding is also closely associated with the acquisition of written forms.

As noted above, according to Baddeley and Hitch's working memory model (1974), working memory consists of two systems, namely the phonological loop and the visuospatial sketchpad, with the phonological loop being responsible for the short-term storage of phonological information. The visuospatial sketchpad, on the

other hand, stores visual and spatial information. The theoretical development of this component has been slower than that of the phonological loop (Maehler and Schuchardt, 2011). In the task of vocabulary acquisition, both the phonological loop and the visuospatial sketchpad may be actively involved. As the central executive regulates the distribution of attentional resources between the two components, efficient decoding not only enables the phonological loop to function more effectively, but also frees more resources for the visuospatial sketchpad to deal with the visual characteristics of new vocabulary, which serves as a powerful mnemonic to secure spelling in memory.

From a different perspective, as phonologically encoded information is more durable than any other form of representation (Kleiman, 1975; Just & Carpenter, 1980), converting visually presented word forms into their phonological forms, also known as decoding, enables the operation of phonological loop and thus holds the information of the new vocabulary longer in working memory, so that other aspects of learning can take place, such as the acquisition of written forms and the inference of meaning from written forms.

Though from a theoretical perspective decoding is closely associated with these aspects of vocabulary learning, this relationship is still relatively under researched. To the author's knowledge, there are only two studies which have directly investigated the influence of decoding on the acquisition of written forms and the acquisition/

inference of word meaning in L2 vocabulary, both conducted by Hamada and Koda (2008, 2010).

In Hamada and Koda (2008), the decoding proficiency of two groups of EFL learners with different native languages was firstly examined, followed by a word learning task which explored whether their decoding proficiency had any connection with their learning outcomes. 18 Chinese and 17 Korean EFL learners took part in the study. Their English proficiency was matched based on a TOEFL reading section, and no significant difference was found in terms of the two groups' performance in an L1 picture-naming task with stimuli consisting of common items, suggesting that their cognitive ability was also similar. The participants firstly took a decoding test including 28 pseudowords. They were given 5,000 ms to decode each stimulus, and only the utterances produced within this period were scored. The results showed that the Korean group performed more quickly and accurately than the Chinese group in the decoding test. In the later word learning session, participants were presented with 16 pseudowords paired with corresponding pictures on a computer screen. This was followed by three immediate recall tests, namely a spelling test, a picture recognition test and a word recognition test. It was found that the Korean group constantly and significantly outperformed the Chinese group across all three recall tests, suggesting that those who were more proficient in decoding also performed better in the acquisition of form-meaning links when learning new vocabulary. Another important finding was that a correlation coefficient of .31 ($p < .01$) was found between decoding

proficiency and spelling test accuracy in an overall correlation test of all the participants, which directly pointed to the strong connection between decoding and the learning of new L2 orthographic forms. However, an interesting observation is that the correlation between decoding and word learning was significant only for the Korean group, indicating that the Chinese group might have used other strategies in vocabulary learning.

In their later study, Hamada and Koda (2010) examined the role of decoding in another aspect of word learning, word-meaning inference. Two groups of EFL learners with contrasting L1 orthographic backgrounds, one being the alphabetic L1 group (15 native speakers of Korean and 1 native speaker of Turkish) and the other being the logographic L1 group (13 native speakers of Chinese and 4 native speakers of Japanese) participated in the study. Similarly to the previous study, English proficiency, which was tested by a TOEFL reading section, and working memory span, which was tested by a numerical digit memory span test, were examined first, and no significant differences were observed. Participants' decoding proficiency was examined using a naming task of 20 unfamiliar real English words and 20 pseudowords, all chosen from a well-established decoding test (Woodcock, 1987), and again both speed and accuracy of decoding were measured. Then the participants completed a word-meaning inference task, in which they read a passage containing 10 pseudo words and were asked to write down the inferred meanings. The results showed that the alphabetic group significantly outperformed the logographic group in

the decoding of both real and pseudo words, which again suggested an effect of L1 orthographic background, as indicated by their previous research. The results of the word-meaning inference task showed that the alphabetic L1 group also outperformed the logographic L1 group; however, the difference was not significant.

These two studies have provided useful first steps to explore the relationship between decoding and the acquisition of written forms in L2 vocabulary learning. However, some potential limitations should be noted. Firstly, though the English proficiency of the participants was tested and matched between groups, neither of the studies directly examined participants' vocabulary breadth, which could potentially be a confounding variable. As the pseudo words used in the decoding test in Hamada and Koda (2008) were adapted from real words by changing a letter, those who had prior knowledge of these words were more likely to correctly decode them, which might jeopardize the validity of the decoding test. Secondly, some other factors that could possibly confound the vocabulary learning results – such as motivation, strategic behaviour and educational background (including previous instructional approaches in both L1 and L2) – were not explored. Thirdly, both these studies involved relatively small numbers of participants, and thus the generalizability of the findings is open to question. Fourthly, a time limit was imposed on the decoding test in both studies. Though the proficiency level of the participants was relatively high, so that most of them probably did only need 5,000 ms to decode a word, there was still a chance that some participants might have needed more time. An interesting observation made in

Erler's study of Year 7 English learners of French (2003) was that some of the participants who took the longest in the decoding test turned out to produce the best pronunciations for English-resembling words. This suggests that a longer reaction time does not necessarily suggest poor decoding, but may rather indicate that the participants had not achieved automaticity in L2 decoding: that is, they may have needed more time to consciously figure out the pronunciations of these L2 words.

The studies reviewed above suggest that decoding plays an important role in the acquisition of both phonological and orthographic forms in L2 vocabulary learning. However, as current research evidence has only demonstrated a correlation between decoding and vocabulary acquisition, whether there is a causal relationship between them still awaits exploration. Hence, from a research perspective, intervention studies are called for, since if decoding instruction is found to promote vocabulary acquisition, it will 'afford ultimate proof of the causal linkage between decoding and word learning' (Koda, 2008: 25). From a pedagogical perspective, as considerable research evidence has suggested that EFL learners frequently encounter difficulty with both phonological and orthographic forms (Schmitt, 2010), this might also provide useful implications for English vocabulary instruction.

### 2.1.3 Phonological decoding and motivation

Decoding not only plays a crucial role in both reading comprehension and vocabulary learning in L2, but it is also argued to have an impact on other aspects of language learning, such as learners' motivation. For instance, Erler and Macaro (2011)'s large-scale study examined whether Key Stage 3 students' French decoding proficiency would influence their perceived self-efficacy in decoding-related tasks, their attitude toward learning French generally, and their motivation to continue French learning. It was found through a questionnaire completed before the decoding test that only a third of the participants thought they would get most of the items correct. Then, after the decoding test, two thirds of the participants reported that they did not feel they had done well in the test, again indicating their low confidence in regard to French decoding. Moreover, over half of the participants reported having negative feelings towards French learning in general. Interestingly, there was a significant positive correlation between the construct of 'positive feeling towards French learning' and the construct of 'decoding test self-efficacy', indicating that those who were confident in their French decoding proficiency were also likely to have a positive attitude towards learning the language. Not surprisingly, those who had confidence in French decoding and felt positive towards the language were also found to be more motivated to continue with their French learning, while those had low self-efficacy in French decoding were less likely to continue learning French. This is not difficult to understand, especially when taking into consideration that written input is one of the most common forms of L2 input in their learning process; thus, the lack of confidence in mentally processing L2 written input could naturally

result in frustration towards learning the language in general. In an earlier work, Erler (2004) argues that beginner learners of L2 French were effectively being put into the equivalent position of dyslexic beginner readers in L1, due to their poor GPC knowledge. This, in turn, could be associated with low motivation and interest in learning French.

Taken together, research reviewed in this section has demonstrated the important role of decoding in various aspects of L2 learning, not only including language skills like reading comprehension and vocabulary acquisition, but also including general attitude and motivation toward L2 learning. Though research evidence on Chinese EFL learners' English decoding proficiency was scarce, existing evidence seems to suggest that Chinese learners perform worse in English decoding compared to their proficiency-matched counterparts who are from an alphabetic L1 background (e.g. Hamada & Koda, 2008, 2010). Hence, the next section aims to offer a possible explanation for this by reviewing literature examining L1 influence on L2 decoding.

## 2.2. L1 influence on L2 decoding

Traditionally, research on learning to read in an L2 writing system has emphasised the crucial role of L1 influence, which can be traced back to the Contrastive Analysis Hypothesis proposed by Lado (1957). It is an intuitive assumption that the similarities

between an L1 and L2 writing system will facilitate learning to read in the L2, whereas differences will often cause problems. This section will examine how various aspects of the Chinese language, including both the Chinese character writing system and the phonographic Pinyin system, impact Chinese EFL learners' decoding of English words.

### 2.2.1 English decoding for Chinese learners: a tale of two writing systems

Languages vary with regard to their writing systems, also known as the ways in which written symbols connect to the spoken language (Daniels & Bright, 1996). The major divide between writing systems has been seen as whether their smallest semantically distinguishing units, also referred to as graphemes, connect with sounds or meanings (Cook & Bassetti, 2005), which also differentiates the English and Chinese writing systems.

English is primarily a sound-based (or 'phonographic') writing system. English words usually consist of a string of letters; and these strings can be further divided into several graphemes which map onto phonemes. There are some general correspondence rules linking graphemes and phonemes and vice versa, which serve as the basis of English decoding (Cook, 2004). For instance, as Berndt, Reggia and Mitchum (1987) conclude after analysing a corpus of 17,310 English words, the

grapheme <ay> is highly likely ($p$ = .97) to correspond with the phoneme /eɪ/ (e.g. *tray*, *play*, etc.), and the grapheme <qu> in most cases ($p$ = .88) is pronounced as /kw/ (e.g. *quick*, *quilt*, etc.). Thus, GPC knowledge is of great importance to English learners in order to efficiently decode written words.

In contrast, Chinese is a meaning-based writing system. In Chinese, graphemes, which are also frequently known as characters, directly represent syllabic morphemes, (DeFrancis, 1989; Collinge, 1992). Both semantic and phonological information is assigned holistically to a graphic symbol. Unlike English, where most vocabulary has its unique phonological form, many Chinese characters might share the same segmental phonology, but each has a different written form. For instance, when /li/ (third tone) means *plum* it is written as '李'; when it means *politeness* as '礼'; when it means *inside* as '里' and when it means *reason* as '理'.

Although Chinese was previously believed to be a logographic writing system, in which the relationship between symbol and sound is entirely arbitrary (Baron & Strawson, 1976), a growing consensus generated by recent research is that most Chinese characters can still be phonologically decoded to at least some extent (Lam et al., 2004). This can be attributed to a phonetic component of many Chinese characters, which is also known as the 'phonetic radical'. As its name implies, the phonetic radical serves to directly or indirectly reflect the phonological information of a character (Cook and Bassetti, 2005). For instance, '奇' (*wonder*) is the phonetic

radical of '骑' (*ride*) , '崎' (*bumpy*) and '绮' (*beautiful*) and they all have the same pronunciation (/tɕi/, second tone) . More than 80% of Chinese characters contain such a phonetic component (Feldman & Siok, 1997), among which 26% have the same pronunciation as their phonetic radicals (Fan et al., 1984). In other words, the phonological information of more than 21% of Chinese characters can be obtained directly from their orthographic forms. For other characters, their phonetic radicals still indicate part (though not all) of their phonological information. For instance, the phonetic radical of the character '吻' (/wən/, third tone, meaning *kiss*) is '勿' (/wu/, fourth tone, meaning *no*), which provides a correct first phoneme. However, it should be noted that though some phonological information is thus available in some Chinese characters, it is encoded at the syllabic level, rather than at the phonemic level, which is the case in most English words (Gottardo, Yan, Siegel & Wade-Woolley, 2001). Moreover, such phonological information can only be obtained if adequate orthographic knowledge is in place.

It has been argued that the contrasting features of the English and Chinese writing systems also lead to different processing mechanisms involved in the word-level reading of the two languages. A large body of research has demonstrated that decoding proficiency is a reliable predictor of English word-level reading (e.g. Hulme, Hatcher, Nation, Brown, Adams & Stuart, 2002; Byrne & Fielding-Barnsley, 1995; Muter. Hulme, Snowling & Taylor, 1998). In contrast, research evidence has suggested that the character is the basic processing component in the reading of

Chinese (Taft, Zhu & Peng, 1999). As Biederman and Tsao (1979: 131) put it, 'a reader of English cannot refrain from applying an abstract rule system to the word; a reader of Chinese may not be able to refrain from configurational processing of the ideography'. Such a difference in processing mechanisms is clearly observed in an oft-cited study by Huang and Hanley (1994). In this study, 8-year-old children from three primary schools in the UK, Taiwan and Hong Kong were tested on reading proficiency, visual form discrimination, non-verbal IQ, vocabulary knowledge and phonological awareness. All the tests were administered in participants' native language and the difficulty of the two sets of tests was matched. It was found that performance on the phonological awareness test per se did not predict reading proficiency of the participants in Taiwan and Hong Kong but was significantly related to reading test results of the British children, even after their performance on the other tests was controlled for. In contrast, test results on visual form discrimination were found to be significantly related to reading performance for the Taiwanese and Hong Kong children but not for their British counterparts. These results clearly demonstrate that visual processing is crucial in the word-level reading of Chinese, whereas phonological awareness is a key factor in the word-level reading of English. However, it should be noted that though some sets of the tests used in this study, namely the visual form discrimination set and non-verbal IQ set, were essentially different versions of the same test, other sets of tests, namely the tests of reading proficiency, vocabulary knowledge and phonological awareness, were designed differently for the English-speaking and the Chinese-speaking children. Therefore, it is hard to say with

confidence that these tests were actually 'matched' in terms of difficulty, which is a potential (though seemingly inevitable) limitation of this study.

As different processing mechanisms are entailed in the reading of English and Chinese at word level, the question of whether cross-linguistic transfer of processing mechanisms occurs in Chinese EFL learners' reading of English words is naturally raised. A robust line of studies in this area (see Koda, 2007) has provided a positive answer to this question. The most powerful evidence comes from a cognitive neuroscience study by Tan, Spinks, Feng, Siok, Perfetti, Xiong, Fox & Gao (2003). In this study, the brain activity of 12 Chinese-English bilinguals and 12 English monolinguals was monitored during phonological decoding in the two languages, using functional magnetic resonance imaging (fMRI). It was found that the neural system that the Chinese-dominant bilinguals relied on in the phonological processing of Chinese characters was different from that used by monolingual native speakers of English. Moreover, when the bilinguals processed English words, identical brain activity to that involved in decoding Chinese characters was observed. This study has lent strong indirect support to the transfer of L1 processing mechanisms in L2 word recognition. Direct support for such transfer is provided by many cross-linguistic studies, in which two main aspects of L1 effects are found.

The first aspect is the type of information (i.e. phonological or orthographic information) predominantly used in L2 word recognition. Koda (1988), for instance,

argues that an essential factor that differentiates processing mechanisms involved in word recognition in varying orthographies is the systematicity of the GPCs, or 'orthographic depth' as Katz and Frost (1992) propose in their Orthographic Depth Hypothesis. Katz and Frost argue that in an absolutely shallow orthography, phonological information is assembled in working memory mainly through grapheme-by-grapheme decoding and is highly reliable, while in an absolutely deep orthography, phonological information can only be obtained after the whole word is identified. Hence learners with a deep orthography L1 such as Chinese and Japanese tend to rely heavily on orthographic information in L1 word recognition, and can potentially transfer such processing mechanisms to L2 English word recognition. In contrast, those with a shallow orthography L1 such as Spanish do not seem to show such heavy reliance on orthographic information in either L1 or L2 word recognition (Taylor & Taylor, 1995).

This difference is observed in a study by Wang, Koda and Perfetti (2003), in which they compared two groups of college-level proficiency-matched EFL learners from distinct orthographic backgrounds, namely Chinese (morphemic) and Korean (non-Roman alphabetic) and examined which type of information – phonological or visual – was mainly used in the recognition of English words. The participants firstly performed a semantic category judgment task, in which they were presented with a category name (e.g. *a flower*) and then a word (e.g. *rows*) and decided whether the word belonged to the category. Two key variables were phonological similarity and

spelling similarity, which created four experimental conditions, namely similarly spelled homophones (e.g. *stair* - *stare*), similarly spelled controls (e.g. *stair* - *stars*), less similarly spelled homophones (e.g. *rose* - *rows*) and less similarly spelled controls (e.g. *rose* - *robs*). It was found that the Korean participants committed significantly more errors in judging both types of homophones than the corresponding spelling controls, while the Chinese participants demonstrated no such differences. In contrast, the Chinese group consistently made more mistakes in judging both types of similarly spelled words than the corresponding less similarly spelled ones. These results seem to demonstrate that the Chinese EFL participants relied more on visual information in the recognition of English words. This was further validated by a subsequent phoneme deletion task, in which the Chinese group performed significantly worse than the Korean group and made more mistakes that were phonologically incorrect but orthographically acceptable. This again suggests that Chinese EFL learners depend less on phonological decoding and more on orthographic information in the recognition of English words. However, a potential limitation of the study is that, though the two groups of participants were matched for English proficiency – based on self-report and on scores in standardized English proficiency tests (TOEFL and Michigan) – neither could be argued to accurately reflect participants' English proficiency at the time of the study, as the standardized English proficiency tests were probably taken a long time before the study.

In addition to the type of information predominantly used in word recognition, L1

effects are also reflected in EFL learners' sensitivity towards intraword structure, which is the internal orthographic structure of words (Koda, 1999). It has been found that L2 readers with a non-alphabetic L1 are less sensitive to intraword structure than those with an alphabetic L1, a difference possibly attributable to the difference in sound retrieval processes between alphabetic and nonalphabetic languages (*e.g*. Koda, 1999; Muljani, Koda & Moates, 1998; Akamatsu, 2003). Considerable research evidence has demonstrated that analysis and blending of the constituent letters is essential in learning to read an alphabetic orthography (e.g. Ehri, 1992). L1 readers of this type of orthography recognise words by analysing the intraword structure and blending the constituent components until the words become familiar and can be read as a whole (Akamatsu, 2003). In contrast, nonalphabetic languages such as Chinese generally do not require the amalgamation of intraword elements in order to retrieve sound. As a result, native speakers of Chinese tend to pay less attention to the intraword structure and use whole-word processing in the recognition of words (Taylor & Taylor, 1995).

Koda (1999), for instance, found that Chinese EFL learners performed much less well in rejecting illegal strings in written English than their proficiency-matched Korean counterparts, suggesting that Chinese EFL learners demonstrate less sensitivity toward English intraword structure than those with an alphabetic L1. Similar results can also be observed in Akamatsu (2003), in which Chinese EFL learners were more adversely affected by case alternation (e.g. ThErE, aBsOLUtE) in reading an English

passage than those with an alphabetic L1 (Persian). Moreover, Koda (1999) has also gone further, to examine the relationship between intraword sensitivity and decoding (pseudoword reading) proficiency, and found that intraword sensitivity accounted for more variance in the decoding proficiency of the Korean participants (54%) than the Chinese participants (32%), and the Korean group also outperformed the Chinese group on the decoding test. These results suggest that Chinese EFL learners lack sensitivity to English intraword structure, which may negatively affect their decoding of English words.

Again, however, some potential limitations of these studies need to be acknowledged. Firstly, though the English proficiency of the alphabetic L1 participants and nonalphabetic L1 participants were matched in these studies (judged by a reading comprehension measure such as a TOEFL reading section), other possible confounding variables were not taken into consideration, an important one being participants' English vocabulary knowledge, which is an important component of participants' English proficiency and intuitively can be closely connected with their English word recognition. Secondly, the nature of these cross-linguistic studies requires involving participants from various countries with different cultural backgrounds. It is difficult to get the whole picture of their educational background and how they were previously instructed to read English words, which could also play a role in their English word recognition. Nonetheless, this is arguably a 'necessary evil' in cross-linguistic studies, given that it is almost impossible to find two groups

of participants with different L1 writing systems but the same educational background, as they are from different countries.

In sum, the above research evidence has suggested that Chinese EFL learners are often influenced by the processing mechanisms of Chinese in the reading of English words. Their predominant use of orthographic information and their lack of sensitivity to intraword structure appear to pose great challenges for them in decoding English words.

### 2.2.2 An underplayed factor in English decoding: Pinyin, friend or foe?

Research on the effects of L1 Chinese on learning L2 English has predominantly emphasized the 'negative transfer' (Odlin, 1989), or 'interference' (e.g. Koda & Reddy, 2008) from Chinese EFL learners' native language, probably due to the distinct differences between the two writing systems and the resulting processing mechanisms associated with the two orthographies, as mentioned in the previous section. However, little research has examined the possible influence of Pinyin on the reading and decoding of English words. This section aims to discuss why the Pinyin system might be an underplayed factor in the research of L1 influence on L2 learning for Chinese EFL learners.

Pinyin is an alphabetic system using the Roman alphabet, which is the same alphabet as used in English, to denote the pronunciations of Chinese characters. Though sharing the same alphabet, Pinyin and English are very different in terms of the depth of orthography. As discussed in the previous section, English is generally considered to be a 'deep' writing system (with low phonological transparency), in which the correspondences between graphemes and phonemes are often not congruent. In contrast, Pinyin is a highly transparent system, where the GPC rules are always clear and straightforward.

Until recently, the Pinyin system was viewed as a tool to facilitate the reading of Chinese characters for native-speaking children, as Pinyin is used to 'transcribe the printed words into phonological codes, which may correspond to concepts they already know orally' (Lü, 2017: 310). Native-speaking Chinese children generally start to learn Pinyin in the first year of primary school (e.g. Cheung & Ng, 2003), at the age of 6 or 7 years, though many children might already have had some access to Pinyin knowledge at a younger age, thanks to the teaching of their parents and kindergarten teachers. Though Pinyin has been included in the primary school syllabus since 1958 and most adults in Mainland China learned the system when they were young children (Li & Rao, 2000), it was not routinely used by adults – until recently. This situation has changed, however, as a result of the prevalence of computers and mobile phones, as Pinyin is the most popular typing method to enter Chinese characters via keyboard (McBride-Chang, 2012). In addition, Pinyin

transcriptions are also commonly used together with Chinese characters in public life, such as on road signs and billboards. As a result, most Chinese native speakers residing in Mainland China use Pinyin on a daily basis nowadays.

Considering that Pinyin is actively used by a very large population regardless of their age, it seems an omission not to contemplate the possible influence of Pinyin when it comes to research into L1 effects on L2 English learning, especially because Pinyin and English share the same Roman alphabet. Regrettably, many previous studies seemed to gloss over this issue and simply treated Chinese speakers as having a 'logographic L1', which is an oversimplification. Currently, studies in this area are very limited in number, which is understandable given that computers and mobile phones have been prevalent in China only for approximately the last decade. Before that, as noted above, Pinyin was mainly used only by young children, most of whom had not started to learn English. This is evident in the current research, as most of the research on Pinyin focused on two types of learners: (a) Young native-speaking Chinese children who use Pinyin as a tool to learn Chinese characters, and (b) Non-native speakers of Chinese who are learning Chinese as a foreign language. The following provides a brief summary of the general findings in these two domains of research, which can potentially shed some light on the influence of Pinyin on English learning.

Research on the influence of Pinyin on L1 literacy has consistently found that

learning Pinyin promotes phonological awareness in Chinese for native-speaking children. For instance, a study conducted by Xu and Ren (2004) examined the relationship between Pinyin knowledge and one aspect of the wider construct of phonological awareness, namely phonemic awareness in native-speaking Chinese children in Years 1, 3 and 5 of primary school. Participants' Pinyin knowledge was measured using a dictation task, where they heard some Chinese words and were asked to write down the Pinyin transcription, and a translation task, where they saw some Pinyin syllables and were asked to write down corresponding Chinese characters. Participants' phonemic awareness was measured using an oddity task, where they were asked to pick from a list of syllables the one containing a different onset or rime, and a phoneme deletion task, where they heard a Pinyin syllable and were asked to remove a phoneme and pronounce the new sound. The results revealed a clear correlation between Pinyin knowledge and Chinese phonemic awareness for participants of all three age groups, and Pinyin knowledge accounted for 46% of the variance in their phonemic awareness.

From a different perspective, research with non-native learners of Chinese has also pointed to the crucial role of Pinyin knowledge in acquiring Chinese phonological awareness. Lü (2017) examined the role of Pinyin knowledge in Chinese-English biliteracy learning using a group of Year 2 English native-speaking children in a Chinese immersion programme in the US. Two sets of data were collected at two time points, namely the beginning and the end of Year 2. It was found that Pinyin

knowledge strongly correlated with Chinese phonemic awareness and tone awareness (measured by an oddity task similar to the one motioned above) at both time points.

Packard (1990) studied the contribution of Pinyin to learning Chinese as a foreign language, working with two groups of native English-speaking students in an American University. One group started learning Chinese characters at the beginning of the programme without receiving any instruction on Pinyin, while the other group spent three weeks learning Pinyin before moving on to learning Chinese characters. The results demonstrated that the Pinyin-learning group outperformed the character-only group in terms of Chinese syllabic awareness (judged by a task of distinguishing unfamiliar Chinese syllables), and the Pinyin-learning group also became more fluent in terms of spoken Chinese over a semester (13 weeks).

It can be seen that Pinyin knowledge can play an important role in acquiring different aspects of phonological awareness in Chinese, such as phonemic awareness, tone awareness and syllabic awareness, regardless of whether the learners are native-speaking Chinese children or English native speakers who are learning Chinese as a foreign language. This naturally invites a question, which is, whether phonological awareness in L1 can be transferrable to learning another language. This question has been scrutinized by a robust line of studies involving alphabetic L1 speakers learning an alphabetic L2 (e.g. Melby-Lervåg & Lervåg, 2011; Sparks, Patton, Granschow & Humbach, 2009; Schiff & Calif, 2007), which generally point

to the existence of cross-linguistic transfer of L1 phonological awareness to L2 learning. Another, oft-cited study is Durgunoğlu, Nagy & Hancin-Bhatt (1993), which examined the role of L1 phonological awareness in L2 word recognition, using a group of L1 Spanish-speaking beginning-level learners of L2 English. It was found that their phonological awareness in Spanish, judged by a set of tasks including segmentation, blending and matching at phoneme/ syllable/ onset-rime level, correlated strongly with their performance on reading English words and pseudowords, providing evidence that phonological awareness in L1 can be transferred to learning another language. Interestingly, their Spanish phonological awareness accounted for more variance in decoding pseudowords that are pronounced differently in the two languages than the ones that are pronounced similarly, suggesting that stronger phonological awareness in L1 helped curb the automatized triggering of L1 in reading L2 words. Based on these findings, the authors argue that phonological awareness – defined as the ability to identify, segment and manipulate the phonological subcomponents in a word – is not language-specific, at least in the case of two alphabetic languages.

Considering that phonological awareness can be transferred from one alphabetic language to another, it is natural to wonder whether phonological awareness in Chinese, a language with a non-alphabetic writing system, can be transferred to learning English, an alphabetic language, especially when learners are also equipped with Pinyin knowledge. Research evidence seems to provide convincing evidence to

support this hypothesis. Leong, Cheng & Tan (2005) compared the phonological awareness of two groups of Chinese children in Years 4 and 5 of primary school, one group from Beijing who had previously learned about the Pinyin system, and the other group from Hong Kong, who (as is routinely the case for Hong Kong children) had not received any instruction in Pinyin. It was found that the Beijing group consistently outperformed the Hong Kong group across all measurements of English phonological awareness, including a rhyme detection task, a rhyme discrimination task, and a phoneme deletion task. Moreover, the Beijing group also performed better in decoding English pseudowords than the Hong Kong group. Such findings are not surprising, as learning Pinyin entails combining different onsets and rimes to produce various syllables, and the ability to manipulate subcomponents of Pinyin syllables can also be used in the reading and decoding of English words.

Taking a look at the research evidence reviewed in this and the previous section, one cannot help noticing a gap in the current literature. On the one hand, a robust line of cross-linguistic studies – most of which were conducted in the 1990s and early 2000s (i.e. before the prevalence of computers and mobile phones and the resulting daily use of Pinyin), and mostly involving adult Chinese learners of English – demonstrates that Chinese EFL learners are less sensitive to intra-word structures of English words and are more reliant on visual-orthographic processing compared to those with an alphabetic L1. On the other hand, research looking at primary school Chinese children, who have just received systematic Pinyin instruction, point to the facilitative role of

Pinyin in developing both Chinese and English phonological awareness, as well as in the decoding of English words. One cannot help but wonder, in the current context where native-speaking Chinese adults use Pinyin on a daily basis, to what extent and in what ways (both facilitative and disruptive) Pinyin knowledge affects the decoding of English words.

An initial, intuitive answer to this question would likely be that Pinyin knowledge would have a considerable facilitative effect on L2 English decoding, as Pinyin seems to promote various aspects of phonological awareness such as phonemic awareness and syllabic awareness, which can be beneficial in decoding English words. In other words, Pinyin is a 'friend' rather than a 'foe' for L2 English decoding. However, the real picture might be more complicated than this. Several factors need to be accounted for in evaluating the role of Pinyin in decoding English words.

Firstly, the differences between the structure of English words and Pinyin syllables might have an impact on the level of phonological awareness required in decoding. Pinyin syllables are always monosyllabic and have a simple structure, normally consisting of an onset and a rime. In contrast, many English words have more than one syllable or have complex consonant clusters in their onsets and codas (e.g. *string*, *brunch*). This makes the 'burden' of decoding English words heavier than decoding Pinyin syllables, as decoding polysyllabic English words firstly requires correctly segmenting a word into several syllables, and then further segmenting syllables into

graphemes. This first step is not needed in reading Pinyin.

This naturally leads to a conjecture, that Chinese native speakers have better syllable-level phonological awareness, while English native speakers have better phoneme-level phonological awareness. This conjecture has been tested in a study conducted by Mcbride-Chang, Bialystok, Chong & Li (2004). Three groups of kindergarten and first-grade children from different backgrounds, including a group of native Chinese-speaking children in Mainland China who had learned Pinyin (N = 105), a group of native Cantonese-speaking children in Hong Kong who had not learned Pinyin (N = 70), and a group of native English-speaking children in Canada (N = 134), were tested on different levels of phonological awareness in both English and Chinese. It was found that the Mainland China group had the best syllable-level phonological awareness, as measured by a syllable deletion test (deleting a syllable from a three-syllable phrase, e.g. red ball pen → red pen), in not only Chinese but also English. The Canada group achieved the best phoneme-level phonological awareness in English, as measured by a phoneme deletion test (deleting the initial phoneme of a one-syllable word, e.g. cage → age), and similar syllable-level phonological awareness in Chinese as the Mainland China group. The Hong Kong group, who had not received any Pinyin instruction, achieved far lower phoneme-level phonological awareness than the other two groups in both Chinese and English. The results demonstrate that learning an alphabetic writing system (e.g. Pinyin, English) helped promote the development of both syllable-level and phoneme-level phonological

awareness.

Secondly, the levels of consistency of the GPCs ('orthographic depth') in Pinyin and in English are very different. As mentioned before, Pinyin is a highly consistent system where GPCs are always congruent, whereas English is generally seen as a 'deep' (phonologically opaque and inconsistent) phonographic writing system. A robust line of research has compared word recognition and decoding in languages of different orthographic depths, and has demonstrated that word recognition and decoding, especially when measured by pseudoword reading tasks, are easier in shallow orthographies, as grapheme-level decoding is effective in terms of word recognition and decoding. On the other hand, in deep orthographies, decoding an individual word not only takes more time, but is also 'mediated by the lexical representation of the word' (Frost & Bentin, 1987: 105). In other words, larger chunks, or 'grain sizes' of orthographic information (Ziegler & Goswami, 2005), rather than individual graphemes, need to be processed together in order to obtain correct phonological representations. For instance, the grapheme <a> is pronounced differently in 'has' and 'was', even though the two words have only one different grapheme; in these cases, grapheme-by-grapheme decoding is not enough to generate correct pronunciations of both words. As a result, one may argue that word-level decoding in shallow orthographies is somewhat easier than in deep orthographies.

This claim has been supported by research comparing the pseudo-word decoding of

young native speakers with different L1s. For instance, Wimmer and Goswami (1994) found that Austrian children performed better in a German pseudo word reading task compared to a group of age-matched English children in an English pseudo word reading task. The researchers attributed the differences to the fact that English is a deeper orthography than German. In the context of cross-linguistic transfer studies, it is generally believed that it may be easier to move from a deep orthography to a shallow orthography than the other way around (Koda, 2007). Considering that the orthographic depth of Pinyin is shallower than that of English, whether Chinese EFL learners would subconsciously transfer the grapheme-by-grapheme decoding mechanisms, which is sufficient for decoding Pinyin syllables, into decoding English words remains to be explored.

Thirdly, considering that Pinyin and English share the same Roman alphabet, it is worth considering the potential transfer of Pinyin GPC knowledge when decoding L2 English words. Though it is difficult to find research that has examined this potential impact of Pinyin on L2 English decoding in the current literature, research into language learners whose L1 and L2 share the same alphabet seems to confirm the existence of transfer of L1 GPC knowledge in L2 decoding. For instance, in his examination of English learners' decoding of French words, Woore (2014) found that the overwhelming majority of the participants decoded the French grapheme <ée> as /i/, pointing to the transfer of knowledge of the English grapheme <ee>.

However, the L1 GPCs may not always be disruptive in L2 decoding, especially when certain graphemes have the same or closely similar pronunciations in the two languages. Woore (2014) listed the French word 'voile' (/vwal/, meaning 'sail') as an example, and argues that English native speakers would probably produce correct pronunciations of the graphemes <v> and <l> in this word, thanks to their L1 GPC knowledge. It is worth exploring whether Pinyin GPC knowledge has any impact on Chinese EFL learners' decoding of English words. In line with Woore (2014), it is predicted that the graphemes that have the same or closely similar pronunciations in English and Chinese would play a facilitative role in Chinese EFL learners' decoding of English words, while those that have different pronunciations in the two languages would be disruptive.

Fourthly, how Pinyin is instructed might also have some impact on how Chinese EFL learners process English words in decoding. Considering that Pinyin is an alphabetic system, it is natural to assume that some degree of 'spelling' is required when learning Pinyin syllables. However, this might not necessarily be the case, as the limited number of Pinyin syllables (approximately 400 if not taking tones into consideration) makes the holistic teaching approach feasible (Li & Thompson, 1981). Given the vast differences in the instructional approaches taken in different areas in China, it is difficult to estimate the percentages of children who learned Pinyin in a holistic approach compared to those who learned Pinyin in a phonics-based approach (i.e. learning onsets and rimes / graphemes separately, and the blending of these).

However, research that examined two groups of Pinyin learners who had experienced these different instructional approaches did find some differences in their reading and writing of Chinese characters. McBride-Chang, Lin, Liu, Aram, Levin, Cho, Shu & Zhang (2012) evaluated the levels of maternal mediation on preschoolers' Pinyin learning, with mothers who taught Pinyin in a holistic approach considered as offering low-level mediation, and mothers who taught Pinyin in a phonics-based approach as offering high-level mediation. It was found that maternal mediation accounted for 6% of the variance in word writing and 7% of the variance in word reading. It is worth entertaining the hypothesis that these two types of Pinyin learners also have some differences in their processing of English words. Those who learned Pinyin in a holistic approach might be less inclined to analyse the intraword structures of English words, as for them both Pinyin and Chinese characters can be processed as visual wholes.

Taken together, though few studies have examined the potential influence of Pinyin knowledge in Chinese EFL learners' decoding of English words, there are plenty of reasons to believe that such influence exists and should not be downplayed. Both Pinyin and English share the same Roman script, and phonological awareness of Pinyin can plausibly be transferred to support the decoding of L2 English words. However, from the Contrastive Analysis point of view (Lado, 1957), Pinyin and English have many different properties, such as orthographic depth and differences in individual GPCs, which could potentially be disruptive in the decoding of English

words. In addition, how Pinyin was taught to Chinese students might have some impact on the way they process written input in English. Considering the current context where Chinese adults use both Chinese characters and Pinyin on a daily basis, it is important that studies of Chinese EFL learners reflect this changing context, and pay attention to the potential influence of Pinyin in English decoding.

## 2.3 The need for explicit instruction in L2 phonics

The previous section has demonstrated that Chinese EFL leaners, at least those who reside in Mainland China, are potentially influenced by both the Chinese writing system and the Pinyin writing system when decoding English words. However, the dual systems can potentially present challenges for Chinese EFL learners' English decoding, as research has demonstrated that Chinese EFL learners generally perform less well in English decoding tests compared to their proficiency-matched counterparts whose L1 is an alphabetic language (e.g. Hamada & Koda, 2008). Similarly, Japanese EFL learners, who have both the alphabetic Rōmaji system (similar to Pinyin) and the non-alphabetic kanji system (similar to Chinese characters) at their disposal, are also often found to have more trouble in decoding English words than those with an alphabetic L1 (e.g. Akamatsu, 2003).

Even for those learners whose L1 and L2 are both alphabetic languages, acquiring L2

decoding proficiency is not always an easy process, as has been shown in many studies reporting L2 decoding or word-recognition problems (e.g. Verhoeven, 1990, 2001; Geva, Wade-Woolley & Shany, 1997). It is worth contemplating why L2 decoding proficiency may not be acquired easily through the course of traditional L2 instruction programmes, which generally takes several hours per week (2 hours for Key Stage 3 students in England and 4 hours for middle school and high school students in China), with no specific focus on GPC knowledge (Woore, 2009).

It is generally believed that the mechanisms involved in the acquisition of L2 decoding proficiency are different from those involved in the acquisition of L1 decoding proficiency. L1 decoding proficiency is generally developed when L1 oracy skills have already reached a certain level. When they start learning to read, they can map the words they already know orally to their written forms, and gradually learn about the correspondences between graphemes and phonemes (Woore, 2011), and thus Share (1995) argues that phonological decoding acts as a 'self-teaching mechanism'. As their literacy skills advance, their knowledge of GPCs is constantly shaped and tuned by their increasing vocabulary. In contrast, L2 learners have limited oral vocabulary to begin with, yet have often already developed fluent L1 lower-level literacy processing mechanisms that cannot always be used successfully in processing L2 written words. On top of that, L2 learners generally do not have as much vocabulary as L1 learners, and thus have fewer chances to tune and update their understanding of the L2 GPC system. As a result, explicit instruction in L2 GPC

knowledge may be of particular importance to L2 learners. This will be discussed from the following two perspectives, namely the acquisition of (a) phonotactic competence and (b) GPC knowledge.

### 2.3.1 The acquisition of phonotactic competence

Phonotactic competence, defined as the ability to determine the permissible combination of phonemes in a language (Speciale, Ellis & Bywater, 2004), can be argued to be an important component of decoding. For instance, in order to correctly decode the word 'best', one needs to possess (at least implicitly) the phonotactic knowledge that consonant clusters are permissible in the English language. (Note that these are not permissible in some other languages, such as Chinese).

The phonotactics of a language are believed to be acquired implicitly by L1 learners. As Ellis (2002: 148) argues, L1 learners 'do not try to learn phonotactics… phonotactic competence simply emerges from using the language'. Studies of L1 learning have demonstrated that native speakers are sensitive to the phonotactic features of the language even at a very young age. For instance, Saffran and Thiessen (2003) discovered that infants aged nine months old could detect phonological regularities in their native speech. In this study, participating infants were firstly exposed to one of two lists of bisyllabic nonwords, one consisting of all CVCV words

(*e.g. boga*, *diku*) and the other consisting of all CVCCVC words (*e.g. bikrub*, *gadkug*) in the induction phase. Then, in the following segmentation phase, the infants listened to a continuous speech consisting of four unfamiliar nonwords, two conforming to the CVCV pattern (*baku*, *dola*) and the other two conforming to the CVCCVC pattern (*tupgod*, *girbup*). The four words appeared in a random order in the speech, with no acoustic stop or prosodic cues signalling word boundaries (*e.g. tupgodbakugirbupdolabaku*…). The testing was conducted immediately after this, where they listened to 12 lists of words, half of which consisting of all CVCV words and the other half of which consisting of all CVCCVC words. It was found that the infants listened significantly longer to the lists of words conforming to the phonological patterns they heard in the induction phase, indicating that they can acquire phonological structure knowledge after a very brief period of exposure. Then, Saffran and Thiessen went further to explore whether the infants could acquire a specific pattern of consonant voicing in bisyllabic words. The experiment procedure was similar to the previous one; the infants listened to either a list of nonwords in which syllable-initial consonants were voiceless and syllable-final consonants were voiced (*e.g. todkad*, or –V+V condition), or a list of nonwords in which syllable-initial consonants were voiced and the syllable-final consonants were voiceless (*e.g. dakdot*, or +V-V condition) in the induction phase. The testing results again revealed significant differences between their listening time on the lists of words conforming to their familiar phonological patterns versus those not conforming to their familiar phonological patterns, indicating that infants can ‘detect and learn

relatively fine-grained phonotactic generalisations' (p. 489).

Given that learners as young as nine months old can acquire phonotactic features of their native language, it is hardly surprising that L1 learners generally achieve a very high level of phonotactic competence after years of using the language. This is confirmed by a line of psycholinguistic studies on learners' awareness of well-formedness of nonwords. For instance, Frisch, Large, Zawaydeh & Pisoni (2001) asked adult native English speakers to rate a list of nonwords on a 7-point familiarity scale, with higher scores suggesting more resemblance to real English words. The participants' ratings were compared to the expected probabilities of the nonwords, which were computed by 'taking the logarithm of the product of probabilities of the onset and rime constituents of the nonwords' (p. 162). The results show that the well-formedness ratings given by the participants were strongly related to the expected probabilities ($r$ = .87), providing convincing evidence that L1 learners naturally acquire good understanding of the phonotactic features of their native language. Frisch and colleagues attribute this finding to the hypothesis that L1 learners judge the well-formedness of nonwords based on the patterns of distribution of phonological constituents in their mental lexicon. In other words, L1 learners have good phonotactic competence because they have enough vocabulary. To test this hypothesis, another word familiarity test was conducted to assess participants' vocabulary, and the results were compared to their ratings on the nonwords well-formedness judgement task. It was found that participants with larger vocabulary

were more likely to judge less probable nonwords as well formed, possibly because less probable nonwords have low-frequency phonotactic patterns, which are only knowable to those who with many low-frequency words in their vocabulary. This clearly established a link between L1 learners' vocabulary and their phonotactic competence.

In contrast, L2 learners are not always equipped with good L2 phonotactic competence, as their vocabulary is generally smaller than that of L1 learners, and their exposure to L2 is limited, considering they only study L2 for a few hours each week. The lack of L2 phonotactic knowledge will almost certainly lead to problems in L2 decoding. Again taking the word 'best' as an example, if an EFL learner does not possess the phonotactic knowledge that consonant clusters are permissible in the English language, they would probably find it difficult to correctly decode this word. The importance of phonotactic knowledge is often neglected in the design of L2 instruction programmes, as learners are expected to pick it up in the course of L2 learning. However, even advanced level L2 learners who do not seem to have any problem in understanding aural input can still face difficulties in oral production. Altenberg (2005) investigated a commonly observed error in Spanish learners' English pronunciation, which was the epenthesis of a vowel phoneme before consonant cluster onsets starting with the phoneme /s/, such as pronouncing *school* as /ɛskul/. In order to examine whether this error was rooted in perception or production, three tasks were designed. The first one was a metalinguistic judgement task, where a

group of L1 English speakers were asked to judge whether a list of nonwords with different onsets could pass as English words, and a group of Spanish EFL learners with varying English proficiency levels were asked to judge the same list of nonwords twice, to see if they could pass as English or Spanish words. The nonwords were divided into three categories, namely (a) both phonotactically possible in English and Spanish (ES); (b) phonotactically possible only in English and not Spanish (E*S); and (c) not phonotactically possible in either English or Spanish (*E*S). The second task was a perception task, where the two groups of participants listened to some nonwords and wrote down their onsets. The third task was a production task, where participants saw pictures of words with different consonant cluster onsets starting with the phoneme /s/, and were asked to name them in English. It was found that in the metalinguistic task, the Spanish group performed very much like the English group in judging whether the nonwords can pass as English words. Moreover, the Spanish group rated the same E*S words as acceptable in English but unacceptable in Spanish, suggesting that they had good understanding of English-specific phonotactic features. In the perception task, the English group and the Spanish group also performed similarly in judging the onsets of both ES and E*S nonwords, suggesting no transfer from L1 at the perception stage. In the production task, however, the Spanish group made many errors, most of which affected E*S nonwords, with word-initial epenthesis of vowel phonemes. The findings demonstrated that the Spanish learners of English could correctly understand and perceive English-specific phonotactic features from aural input, yet even advanced level leaners still faced

problems in correctly pronouncing them. In other words, they failed to correctly decode words with English-specific phonotactic features, because the output of their decoding was filtered through their L1 phonotactics.

Such findings recall Broselow and Park (1995)'s Split Parameter Hypothesis, which claims that in the course of acquiring L2 phonotactic features, there is a stage that L2 settings govern perception only, but not production. The findings led Altenberg (2005) to argue for the necessity for any second language acquisition model to 'account for production and perception as well as metalinguistic knowledge in order to be complete' (p. 76). Given the importance of L2 phonotactic knowledge in L2 decoding and the difficulty in acquiring such knowledge, it is therefore of great importance for L2 decoding instruction programmes to cover this knowledge, especially the production of L2-specific phonotactic features.

### 2.3.2 The acquisition of GPC knowledge

Another important aspect of decoding is knowledge of GPCs. L2 learners have often been found to encounter difficulties in acquiring this kind of knowledge. A robust line of studies conducted in England has consistently identified that students at Key Stage 3 demonstrate poor understanding of French GPCs. For instance, Erler (2003) investigated a group of 359 Year 7 English students' decoding proficiency in L2

French by asking them to complete a written rhyme judgment task. It was found that most participants performed poorly in the test, which led to the conclusion that these students had very limited knowledge about GPCs rules in French. In a later and larger-scale study conducted by Erler and Macaro (2011), 1735 Year 7 to Year 9 English students' decoding proficiency in L2 French was examined using both a rhyme judgment task and a word segmentation task. It was again found that the participants did not perform well in these two tests. On average, they only made correct judgments for approximately 12 pairs out of the total 25 pairs of French words in the rhyme judgment task. Considering that the participants only needed to answer yes or no in the rhyme judgment task, the less than 50% accuracy achieved by the participants was, as the authors put it, no better than 'a matter of chance or luck' (p. 511). Similarly, they correctly segmented only approximately 10 out the total 16 French words on average in the word segmentation task. When comparing the performance of the three year groups, it was found that even though the scores of both tests were higher for older students than younger ones and the statistical analysis also suggested that the differences between Year 7 and Year 8 and between Year 8 and Year 9 were significant, the differences between each group of participants were very small: in the rhyme judgment task, the Year 7 group scored on average 12.13 out of 25, the Year 8 group 12.28 and the Year 9 group 12.64; in the word segmentation task, the Year 7 group scored on average 9.15 out of 16, the Year 8 group 9.61 and the Year 9 group 10.38. The statistical significance was therefore a function of the very large sample size, rather than the actual size of the effect, which was very small.

Given that the participants from different year groups were from the same schools and were instructed by the same teachers, the small differences between the scores of each year group suggested that little progress was made with increasing years of French learning experience.

The findings of the aforementioned two studies, which were derived from large samples of participants, point to the complex nature of acquiring GPC knowledge in an L2, even when learners' L1 and L2 share the same alphabet. However, it should be noted that the test instruments used in these two studies, namely pen-and-paper test forms, may not be the most accurate measure of L2 decoding proficiency. It can be argued that the rhyme judgment task, apart from its binary nature, only tapped into participants' knowledge of the vowel system in French. The word segmentation task, on the other hand, was designed to evaluate participants' syllable-level awareness. In other words, these tests did not directly reflect participants' GPC knowledge. Another thing that warrants caution is that both these studies were cross-sectional rather than longitudinal in nature. As a result, the conclusion that participants made little progress in French decoding over time must be treated with some caution.

Taking these limitations into consideration, Woore's longitudinal study (2009) followed a group of 94 students for a year in order to examine their progress in L2 French decoding. The participants were tested at the end of Year 7 and again at the end of Year 8 using the same Reading Aloud test, rather than a pen-and-paper test, in

order to provide more reliable insight into their French decoding proficiency. They were asked to pronounce 55 unusual French words covering a wide range of 249 grapheme tokens (defined as 'the individual instantiations' of a grapheme). It was found that the participants achieved very similar scores at the two time points, with mean scores (out of 249) of 120.5 at time 1 and 119.4 at time 2, which were not statistically different. The results echoed the findings in Erler (2003) and Erler and Macaro (2011), pointing to the poor performance in L2 French decoding and little, if any progress made in this area among Key Stage 3 students under the prevailing curriculum at the time, where 'there was no systematic, explicit instruction in L2 decoding' (Woore, 2009: 14).

Given the research evidence reviewed in this and the previous section together, it can be seen that the two important aspects contributing to L2 decoding proficiency, namely phonotactic competence and GPC knowledge, cannot always be implicitly acquired from L2 instruction programmes. A probable reason for this, according to Schmidt (1990)'s Noticing Hypothesis, is that special attention and conscious identification are not dedicated to L2 phonics knowledge. Based on the Noticing Hypothesis, the subjective experience of noticing is the prerequisite for any L2 input to become intake. Schmidt (1994) later advanced the Noticing Hypothesis, arguing that in order to acquire a specific aspect of the target language, learners cannot only notice the input of the language in a global sense; instead, conscious attention needs to be directed to the specific aspect before intake can happen. In other words, in order

for L2 GPC knowledge to be acquired, L2 learners must pay conscious attention to those GPCs. This, as discussed before, is not the case for many L2 learners. It is a possibility that explicit instruction could help direct L2 learners' attention to the phonic knowledge, so that proficient L2 decoding can be achieved.

### 2.3.3 Effective L2 phonics instruction: what should it look like?

In order to achieve best results from an L2 phonics instruction programme, the first thing that needs to be taken into account is the form of the instruction. Currently, there have been few studies investigating the effectiveness of phonics instruction programmes in L2, yet a robust line of studies have looked into different types of phonics instruction programmes in L1, given that there has been a general consensus that decoding is a key element in developing children's reading proficiency in languages with alphabetic writing systems (e.g., Comeau, Cormier, Grandmaison & Lacroix 1999, Droop & Verhoeven, 2003). Hence, different types of L1 phonics instruction programmes are briefly reviewed as a starting point.

Currently, many different L1 phonics instruction programmes are being used, which mainly fall into four categories (Slavin et al., 2008). The first category is the reading curricula programme, using textbooks that provide phonics and phonemic practice in the context of engaging stories. In this kind of programme, teachers do not give

explicit lessons on phonics knowledge; such knowledge is encompassed in the structured reading materials. The second category is the instructional technology programme, including supplementary computer-assisted instruction programmes in which students are sent to computer labs for additional exercises, as well as using embedded multimedia in regular lessons. The third category is the instructional process programme, which is characterized by providing teachers with professional development to implement specific instructional methods to teach phonics and phonemic awareness skills. Teachers then provide explicit and systematic instruction on phonics and phonemic awareness in their classes. The fourth category is the combination of reading curricula and instructional process programmes, in which teachers' lessons on phonics and phonemic awareness are further strengthened by structured reading materials. As can be seen, two key differences among the four types of instruction programme are: (a) whether explicit instruction in phonics is provided, and (b) whether structured reading materials are used.

Slavin et al. (2009) conducted a systematic review of the effectiveness of the four types of programmes with beginning readers in kindergarten and first grade in various native and non-native English speaking countries in Europe and North America. Criteria for inclusion of studies included the use of randomised or matched control groups, a study duration of at least 12 weeks and valid achievement measures independent of the experimental treatment. Most of the studies that were included focused on L1 English speakers. It was found that that the 7 reading curricula

programmes had a mean effect size of $d = 0.23$ for decoding measures, and a mean effect size of $d = 0.09$ for reading measures. Similar mean effect sizes were found for the 13 instructional technology programmes. In contrast, the 17 instructional process programmes yielded very positive results, where a mean effect size of $d = 0.47$ for decoding measures and a mean effect size of $d = 0.30$ for reading comprehension measures were found. The 23 combined reading curricula and instructional process programmes also revealed good results, with a mean effect size of $d = 0.33$ for the decoding measures of a mean effect size of $d = 0.27$ for comprehension measures. The results demonstrate that all the four types of instruction programmes had positive effects on promoting young native and non-native children's decoding and reading proficiency, but the instructional process programmes, featuring explicit phonics instruction, seemed to have the strongest evidence of effectiveness. Though the overall effects (small to medium, based on Cohen's rules of thumb[1]) were stronger for decoding measures than for comprehension measures, multiyear studies that followed children into Grade 2 or beyond (*e.g.* Livingston & Flaherty, 1997) revealed larger effect sizes in the delayed post-test ($d = 0.34$) than the immediate post-test ($d = 0.27$) for reading comprehension. Thus, it was argued that positive effects of phonics instruction on reading proficiency were more likely to show later (because initial gains in decoding proficiency then led to improved reading comprehension in the longer term). Another systematic review conducted by Slavin et al. (2008) has also compared the effectiveness of these four types of instruction programmes in middle

[1] Cohen (1992) suggests that $d = 0.2$ be considered as a small effect size, $d = .05$ a medium effect size and $d = .08$ a large effect size

and high schools, the results of which generally conformed to those of the above study.

Taken together, these two systematic reviews provide convincing evidence for the effectiveness of phonics instruction on promoting L1 decoding and reading proficiency, and the instructional-process programmes featuring explicit instruction seem to produce the best results. Similar results were also reported in the National Reading Panel Report (2000) conducted in the US context, where systematic phonics instruction programmes were found to be effective in promoting children's ability to decode regularly spelled real words ($d$ = .67) and pseudo words ($d$ = .60), as well as irregularly spelled real words ($d$ = .40); while different systematic phonics instruction programmes was also found to be effective in promoting children's reading comprehension, with effect sizes varying from $d$ = .23 to $d$ = .68.

The Rose Review (2006: 20) conducted in the UK context also recommended providing children with systematic phonics instruction, and highlighted the four important aspects of high-quality phonics instruction, which are:

(a) [to teach] grapheme/phoneme (letter/sound) correspondences (the alphabetic principle) in a clearly defined, incremental sequence;

(b) to apply the highly important skill of blending (synthesising) phonemes in order, all through a word to read it;

(c) to apply the skills of segmenting words into their constituent phonemes to spell;

(d) [to teach] that blending and segmenting are reversible processes.

Now let us take a look at the existing literature on L2 phonics instruction programmes. These studies, though limited in number, seem to provide some support for the effectiveness of phonics instruction in promoting L2 decoding proficiency. For instance, Sturm (2013) conducted a 15-week instruction programme on L2 French phonetic and phonics for a group of 22 native English-speaking students in a U.S. university. The instruction was conducted by the researcher herself, featuring the learning of IPA symbols representing French phonemes and the instructing of French GPCs. Other aspects of pronunciation were also instructed, including accentuation and intonation. Useful skills for L2 decoding, such as syllabification, were also instructed and practised repeatedly over the course of the instruction programme. After class, participants were asked to record their reading of a text, and were given feedback on their pronunciations. At the beginning and the end of the instruction programme, participants' decoding proficiency was measured by reading the same assigned text, and the percentage of correctly pronounced syllables were counted as scores. The results showed that the intervention participants made significantly more improvements at the end of the term compared to the control groups who did not participate in the instruction programme, suggesting that phonics instruction, together with instruction in phonetics and other pronunciation skills, can be helpful in promoting L2 French pronunciation.

Some limitations of this study should still be acknowledged. The first one is that the test instrument may lack ecological validity, given that the same text made up of real

French words was used twice to measure the progress of participants' pronunciation. As there is no control of vocabulary knowledge, there is no way to know if participants' better pronunciation at the end of the instruction was a result of the instruction programme, or was simply because they knew more French vocabulary at the time. The second one is that the study is relatively small-scale, making it difficult for the results to be reliably generalised. The third one is that the intervention was heterogeneous; as a result, it is difficult to know which of the specific elements really made the difference, and to know whether the phonics instruction specifically made any difference.

Woore (2011) conducted an intervention study investigating the effectiveness of two phonics instruction programmes on French decoding for 186 secondary school students in the UK. Intervention participants were either assigned to the 'poem approach' instruction or the 'phonics approach' instruction, with the difference being that the poem approach instruction additionally included structured reading materials (poems containing target GPCs). In both intervention programmes, 55 challenging French GPCs were instructed over 19 weeks, with a focus on explaining the different realisation of these graphemes in English and French, and opportunities to practise decoding unfamiliar French words containing these graphemes. Intervention participants also received form-focused feedback from their teachers. It was found that both groups of intervention participants made significantly more progress in French decoding compared to the control group, judging by the increases in the

number of correctly decoded graphemes, though the magnitude of progress appeared to be small. The analysis of participants' actual realisation of individual graphemes showed that the intervention participants tended to move away from L1-based pronunciations and towards pronunciations that were more target-like, indicating that some conscious reasoning may have developed as a result of the instruction programmes.

A recent, large-scale intervention programme, Foreign Language Education-Unlocking Reading (FLEUR) conducted by Woore, Graham, Porter, Courtney and Savory (2018) examined Year 7 students' French reading comprehension, decoding proficiency, vocabulary knowledge and overall self-efficacy for reading in the UK. 36 intact classes consisting of approximately 900 students in various parts of the country took part in the project. The classes were randomly assigned to receive phonics instruction, strategy instruction or text-only instruction. In the phonics instruction group, participants received instruction on a set of key GPCs, and worked with structured reading materials that had been designed to exemplify these GPCs. In the strategy instruction programme, participants received instruction on a checklist of 8 strategies to facilitate their understanding of challenging texts, and worked with the same reading materials as the phonics instruction group. In the text-only instruction, participants worked with the same reading materials as the other two groups, but received no explicit instruction in either phonics or strategies. The instruction lasted for at least 16 weeks. Participants' French reading comprehension, decoding

proficiency, vocabulary knowledge and overall self-efficacy for reading were measured immediately before (t1), immediately after (t2) and six months after (t3) experiencing the instruction. There was considerable attrition at t3, leading the authors to express greater confidence in the data obtained at t1 and t2, which is therefore focussed on here.

The results showed that, in terms of reading comprehension, all three groups made significant progress over the course of instruction, though there was no significant advantage for any group(s) over the others. In terms of phonological decoding, all three groups made significant progress, but there was a significant advantage for the phonics instruction group over the other two groups. In terms of vocabulary knowledge, all three groups made significant progress, but there was an advantage for both the phonics instruction group and the strategy instruction group (though most clearly for the former) over the text-only instruction group over the course of instruction. In terms of overall self-efficacy for reading, all three groups became more confident over the course of intervention. The results of this large-scale intervention study support the value of phonics instruction for improving participants' decoding proficiency and vocabulary knowledge, even though there was no evidence that it improved L2 reading comprehension more than alternative approaches.

The above studies have provided encouraging evidence of the effectiveness of L2 phonics instruction programmes in promoting L2 decoding proficiency and other

aspects of L2 learning, and have shed light on how L2 phonics instruction might be conducted effectively. Echoing the recommendations of the Rose Review (2006), as mentioned earlier, L2 GPC knowledge is most effectively acquired when it is explicitly instructed in a clear defined and incremental sequence. Moreover, exercises that promote intraword analysis, such as word segmentation and phoneme blending, can be useful in helping learners better decode L2 words. Moreover, form-focused feedback on learners' interlanguage can also be helpful in promoting learners' pronunciation.

Similar studies investigating the effectiveness of other L2 phonics instruction programmes for English speakers, such as Lord's (2005) L2 Spanish phonics instruction programme has also reported positive results. One common feature of these studies, despite the differences in target L2, is that the learners were moving from a deep orthography (English) to a shallower orthography (French, Spanish). According to Katz and Frost's (1992) Orthographic Depth Hypothesis, moving from a deep orthography to a shallower orthography is hypothesized to be easier than the other way around. What, then, about moving from one orthography to a typologically different orthography? Liaw's (2003) study investigated the effects of an English phonics instruction programme with a class of Taiwanese elementary school students. The phonics instruction programme combined the instruction of GPC knowledge and structured reading. 'Jolly Phonics', a phonics instruction programme widely used in the U.K. for young children, was chosen as the instruction material. In class, some

activities promoting intraword analysis was conducted, including the blending, segmentation and deletion of sounds. Each class lasted for 80 minutes, and the instruction lasted for three months. Participants' phonics knowledge, vocabulary and reading comprehension were assessed at the end of each month to evaluate their progress. It was found that participants made progress in all three measures, leading the author to conclude that the phonics instruction programme was effective in promoting both L2 English phonics knowledge and other L2 skills.

Liaw (2013)'s study provides an interesting starting point for the research into English phonics instruction for learners from a typologically different L1 background. However, some obvious limitations need to be acknowledged. Firstly, without a control group, any progress reported is open to question, as there is no way to know if the progress was indeed the result of the phonics instruction programme, or simply because the participants just got better with more time learning English. Secondly, all the results were reported at a descriptive level; in other words, no statistical analysis was conducted. Thirdly, the test instruments also lacked validity, especially the vocabulary measure which only used five words as test items. Hence, the promising results reported here need to be interpreted with caution. With the lack of research on the effects of English phonics instruction with learners from a typologically different L1 background, it remains unclear whether phonics instruction is useful for such learners.

## 2.4 L2 English vocabulary acquisition in the Chinese context

The previous sections have explored the importance of decoding on various aspects of L2 learning, and the need for explicit instruction of L2 phonics knowledge. This section discusses another important and enduring topic in L2 learning, namely vocabulary acquisition, with a focus on Chinese EFL learners.

The importance of vocabulary knowledge in L2 learning is well acknowledged (Nation, 1990). All aspects of language use, including listening, speaking, reading and writing, are built on the foundation of vocabulary knowledge (Schmitt, 2010). In the context of China, EFL learners often attach a great deal of importance to English vocabulary learning (Gu & Johnson, 1996). Chinese students typically regard a good English learner as someone who has a large vocabulary, and some students even attempt to memorise a whole English dictionary so that they can understand every English word coming their way (Evita, 2014).

Similar to other aspects of L2 learning, vocabulary acquisition in L2 is also influenced by learners' L1 (Schmitt, 2010). This section firstly reviews research into L1 influence on L2 vocabulary acquisition, with a focus on Chinese EFL learners. Then, this section moves on to explore why decoding proficiency is of crucial importance for L2 vocabulary acquisition, and whether phonics instruction can

promote L2 vocabulary acquisition.

### 2.4.1 L1 influence on L2 vocabulary acquisition

As discussed previously, the processing mechanisms associated with Chinese have an important impact on Chinese EFL learners' recognition of written English input. It is natural to question whether such mechanisms also influence their learning of English vocabulary. Current research evidence seems to provide an affirmative answer to this question. For instance, Wang and Geva (2003a) compared Chinese ESL children's lexical and visual-orthographic processing in English spelling with native English-speaking children's in Canada. In Grade 2 these children took three tests, including a real-word spelling test where they heard some familiar English words and wrote them down, a pseudo-word spelling test in which they heard some one-syllable, four-letter orthographically legitimate pseudo words (e.g. stiv) and spelt them, and a pseudoword spelling test in which they were visually shown both orthographically legitimate, pronounceable pseudo-words (e.g. geth) and orthographically illegitimate, unpronounceable letter strings (e.g. pcth) and did an immediate recall of the stimuli. It was found that the Chinese group performed much worse in the pseudo-word spelling test than the native group, though such difference was not found in the real-word spelling test, suggesting that Chinese children were less capable of using phonics knowledge to spell words. Moreover, it was also found

that the Chinese group significantly outperformed the native group in the confrontation spelling test on both orthographically legitimate and illegitimate strings, and the difference between their spelling performance on legitimate and illegitimate strings was also significantly smaller than that of the native group. This suggests that the Chinese ESL learners may be more inclined to use visual memory to learn new words, while phonological decoding, which is of a frequently adopted approach by native speakers of English, seems to be less important for them.

Similar findings are also reported in studies of advanced Chinese EFL learners. Li's (2012) exploratory study examined the strategies that advanced Chinese EFL learners used in a paired-associate vocabulary memorisation task. The interview conducted immediately after the memorisation task revealed five types of strategies. The most commonly used strategy, reported by 38 out of the 60 participants, was 'spelling words in one's mind / out loud', followed by 'looking at words and imprinting them in one's mind', which was reported by 20 participants. 13 participants reported that they identified special orthographic features to facilitate their memory. They made comments like 'There were two '*o*'s in the word *stooge*, which look like two eyes, so the word must have some connection with people'; 'The word *sulcus* starts and ends with *s* and the word also means something like a curve'; 'The word *burlap* starts with *b* and ends with *p* and they both resemble bags' (p. 32). 10 participants associated the stimuli with other words that look similar, with comments such as 'The word *stooge* looks like *stage* and it also has something to do with stage and acting'; 'The word

*sirrah* looks like *sir* and their meanings are also similar'. Only 8 participants reported using the strategy of 'subvocalizing words in the mind / pronouncing words'. It can be seen that among the five strategies reported, four were visual strategies and only one was phonological, which was also the least frequently used one. Though the sample size of the study was small and the results may be difficult to generalise, the fact that these advanced Chinese learners of English (with IELTS scores of 7 or above) overwhelmingly preferred visual strategies again suggests that Chinese EFL learners may be heavily influenced by Chinese-tuned processing mechanisms when learning new English words.

The processing mechanisms of Chinese may not only influence Chinese students' learning strategies of English words, but may also have some impact on their beliefs regarding vocabulary learning, As mentioned in section 2.2.1, unlike English where most vocabulary has its unique phonological form, many Chinese characters might share the same segmental phonology, but each has a different orthographic form. This feature of the Chinese writing system may potentially lead Chinese native speakers' to attach more importance to learning the orthographic forms of new words, as they are unique. For instance, Ma (2009) asked 109 Chinese university students to rate 55 statements about English vocabulary learning using a scale of 1 to 5, 1 being the least likely to happen in the course of their daily studies and 5 being the most. It was found that at a descriptive level, the participants attached more importance to spelling (4.36) than pronunciation (4.28). In addition, visual strategies were frequently used by the

participants to learn English words, including visualising the word form mentally (3.78), writing down the word several times (3.54) and looking at the word several times (3.34). In contrast, sound-based strategies, such as saying the word aloud several times (3.28) and listening to the sound recording of the words (2.71) were less favoured by the participants.

### 2.4.2 Phonics knowledge and vocabulary acquisition in L2: bridging the gap

The lack of research on L2 phonics instruction makes it difficult to draw a convincing conclusion about the relationship between decoding proficiency and vocabulary knowledge in L2. However, research on L1 phonics instruction has identified the positive role of decoding proficiency in promoting spelling attainment for native English-speaking children. For instance, Johnson and Watson (2005) conducted a seven-year longitudinal study which followed around 300 children from Grade 1 to Grade 7 in Scotland. In Grade 1, children in the experimental group entered a one-year phonics instruction programme, in which they received explicit phonics lessons from their teachers. At the end of the programme, the children in the experimental group were 7 months ahead of the control group in reading comprehension, and 8 to 9 months ahead in spelling. In Grade 7, the children who received phonics instruction were reading 3.5 months ahead of the control group, and were spelling 1 year 8 months ahead. It is of interest to note that the intervention

group's advantage in spelling performance enlarged six years after the instruction in comparison to immediately after the instruction, suggesting that the effects of phonics instruction on spelling attainment not only stood the test of time, but also seemed to be cumulative. In other words, one may argue that there is a Matthew effect of phonics instruction on spelling attainment. Similar results were also reported in Ball and Blachman's (1991) study of kindergarten children taking part in an English phonics instruction programme in the U.S.

In China, phonology-related aspects of English learning, including both pronunciation and phonics knowledge, generally receive less attention from learners and teachers compared to other aspects, such as grammar and vocabulary. There are several reasons for this. Firstly, in primary school, middle school and high school, English lessons are usually instructed in big classes consisting of 50 or even more students. It is obviously unfeasible for every student to practice speaking and receive feedback from the teacher in class. As a result, the Communicative Language Teaching approach (Thompson, 1996), which is advocated by many researchers, is not always favoured by Chinese teachers and learners. For instance, some participants in Rao's (2002) study of Chinese learners' perception of communicative activities in EFL classrooms reported feeling frustrated when communicative activities were going on in class, as 'there is chaos when we are asked to interact with each other' (p.93). Secondly, given the vast differences in access to educational resources and in instruction quality between different regions in China, English teachers do not always

have the ability to conduct pronunciation and phonics instruction in class, as they may have difficulty in correctly pronouncing English words themselves. Thirdly, the majority of English exams in China, including the National College Entrance Exam ('gaokao', 高考), which is considered as the single most important exam for Chinese students (Shi, 2006), focus on grammar and vocabulary. In contrast, pronunciation is seldom tested, unless students intend to become English majors in college.

Though great importance is often attached to vocabulary learning and less focus is given to phonology and phonics, inadequate understanding of English phonology and phonics may actually affect Chinese learners' acquisition of English vocabulary. Take the aforementioned study by Li (2012) as an example. The overwhelming dominance of visual strategies may not only be the result of influence of Chinese orthographic processing mechanisms, but may also indicate participants' lack of systematic English phonics knowledge. It is noteworthy that the 8 participants who reported using phonological strategies were also the most proficient decoders among all the participants, while those who were less proficient in decoding made comments such as 'My guessing of the pronunciation is probably wrong anyway, so I don't dare to memorise new words based on this' (p. 45). This echoes the findings reported by one of the most oft-cited studies of Chinese EFL learners' vocabulary learning strategies, which was conducted by Gu and Johnson (1996). In this large-scale study of 850 university EFL learners in China, the relationship between beliefs about vocabulary learning and self-reported vocabulary learning strategies was explored. It was found

that the participants generally did not have a positive attitude towards visual memorisation. In fact, they reported using more oral repetition than visual repetition as a rehearsal strategy. However, when asked about recall strategies, visual recall (or 'visual recording' as used by the researchers) was still preferred over auditory recall (or 'auditory recording'), as most participants found it to be a more reliable way to spell English words. Such contrast is interesting: the participants did not particularly favour visual strategies in memorising new words, but still predominantly used them to spell words.

It should also be borne in mind that Gu and Johnson (1996)'s study was conducted more than two decades ago, at a time when Chinese students' English learning opportunities were very much confined to studying textbooks and word lists. In these scenarios, pronunciations of the words are often provided in IPA form, so that learners do not necessarily have to decode the words in order to obtain correct pronunciations. This may also explain why the participants in Gu and Johnson (1996) reported using more oral repetition than visual repetition as a rehearsal strategy, as the pronunciations were likely available to them. In contrast, the participants in Li (2012) predominantly used visual strategies to memorise the words, probably because their pronunciations were not readily available and had to be obtained through decoding. From a different perspective, the finding that the participants in Gu and Johnson (1996) did not favour visual strategies in memorising new words but still used them to spell new words indicates that they may encounter difficulty in mapping their

(possibly readily available) pronunciations to their orthographic forms. This may potentially indicate a lack of understanding of phonics knowledge.

In the current context, learning phonics is even more important, because with access to ever-growing English input from different resources in China, it is not always the case that the IPA form or the phonological form of words are provided together with the orthographic form. For instance, learners can pick up new English vocabulary from reading magazines, listening to songs, watching movies or browsing the internet, in some of which cases the orthographic and the phonological information of new words may not be available together. The lack of adequate phonics knowledge will hinder learners from picking up new vocabulary in these scenarios.

Taken together, research reviewed in this section has demonstrated that Chinese EFL learners seem to depend more on visual strategies in learning English vocabulary. However, their heavy reliance on visual memory may be a reluctant choice, as they do not have enough English phonics knowledge which can be reliably used to learn new words. It is therefore worth exploring whether a systematic English phonics instruction programme can be of help in addressing this problem. It is also worth mentioning that the above-mentioned studies on Chinese EFL learners' vocabulary learning, like other similar research, seems to focus more on the spelling aspect of vocabulary learning. It is also worth exploring whether phonics instruction can be of help in promoting different aspects of vocabulary learning, including the learning of

both receptive and productive, and both phonological and orthographic knowledge.

## 2.5 Summary

Research reviewed in this chapter has demonstrated that the importance of phonological decoding in L2 learning is supported by both theoretical models and empirical evidence. Reliable phonological representations of written input have been found to influence the outcomes of reading comprehension, vocabulary learning, and also learners' general attitude and motivation towards learning the L2. In regard to vocabulary acquisition, adequate phonics knowledge not only better prepares the phonological loop to store and rehearse new pronunciations, but also frees up more resources for the visuospatial sketchpad to deal with the orthographic characteristics of new words. In learning new words from written input, grapheme-to-phoneme mapping generates the pronunciation in memory, the correctness of which determines to a large extent whether the word can be successfully recalled orally and recognised aurally. In spelling, phoneme-to-grapheme mapping can also be used to spell out the phonological information stored in memory. Both of these operations require systematic phonics knowledge, which L2 learners, and Chinese EFL learners in particular, often lack.

Chinese EFL learners' poor decoding of English words may in part be attributed to

the characteristics of their L1. Under the influence of the processing mechanisms associated with their predominant, morphemic L1 writing system, Chinese EFL learners predominantly rely on visual processing in the reading of English words, and are less sensitive to intraword structures. However, when they try to analyse the pronunciations of English words, their Pinyin phonics knowledge can also be automatically triggered and interfere with decoding, as Pinyin shares the same alphabet as English. However, as current research evidence on Chinese EFL learns' English decoding is limited, it remains unclear whether, and in what ways, the two completely different systems, Chinese characters and Pinyin, may interactively influence English decoding.

Given the importance of phonics knowledge in L2 learning and Chinese EFL learners' lack of such knowledge, it is of important pedagogical value to consider how English phonics knowledge can be effectively taught to Chinese learners. Existing L1 and L2 phonics instruction programmes provide possible templates for the design of a systematic English phonics instruction programme for Chinese learners. L2-specific phonotactic features should be given focus to, and the production of which needs to be practised. GPC knowledge should be explicitly instructed in a clearly defined and incremental sequence. Exercises such as word segmentation and phoneme blending can help promote learners' intraword awareness, and may be of help in L2 decoding. Moreover, feedback on participants' pronunciations is also useful.

Currently, there is no research evidence on the effects of systematic English phonics instruction on Chinese EFL learners. This study aims to address this gap in the literature by exploring the effects of such instruction on Chinese university-level EFL learners. It is worth exploring whether, through systematic English phonics instruction, Chinese learners' English decoding proficiency can be promoted, and whether this can in turn result in better performance in vocabulary learning tasks.

## 2.6 Research Questions

To address the gap in literature as discussed in this chapter, the current study aims to answer an overarching research question:

> *What are the effects of a systematic programme of L2 English phonics instruction on (a) English decoding proficiency and (b) English vocabulary learning, amongst Chinese university-level EFL learners?*

This overarching research question is addressed from several different aspects. The most direct approach to address this research question is to compare participants' scores/accuracy percentages achieved in a phonological decoding test before and after the instruction, in order to examine whether any progress has been made in this area over the course of the instruction. However, this level of analysis alone would not

investigate whether the phonics instruction is more effective for some GPCs than others, which could potentially shed light on the design of future instruction programmes for Chinese EFL learners. This seems to be of particular importance in the current context, given the lack of such a systematic instruction programme specifically designed for Chinese EFL learners, and growing interest in phonics instruction in China over the past few years. Therefore, a more nuanced line of analysis is called for, to examine the effects of the instruction programme on different GPCs. Moreover, given the lack of qualitative research on Chinese EFL learners' English decoding, it is also of interest to examine the 'product' of participants' decoding, which may provide a more comprehensive picture of the features and characteristics of Chinese EFL learners' English decoding. Finally, the current study also investigates whether systematic phonics instruction can help promote various aspects of participants' intentional English vocabulary learning, including the learning of both receptive and productive, both phonological and orthographic knowledge.

Based on this, the following four specific research questions are proposed:

*RQ 1: Does a programme of systematic phonics instruction lead to improvements in Chinese university EFL learners' English decoding?*

*RQ2: Is the programme of phonics instruction more effective for some GPCs than others?*

*RQ3: What are the features and problems of Chinese university EFL learners'*

*English decoding at each time point?*

*RQ4: Do participants who have followed the programme of phonics instruction also show gains in their ability to recall (productive knowledge) and recognise (receptive knowledge) new English words?*

# Chapter 3. Research Design

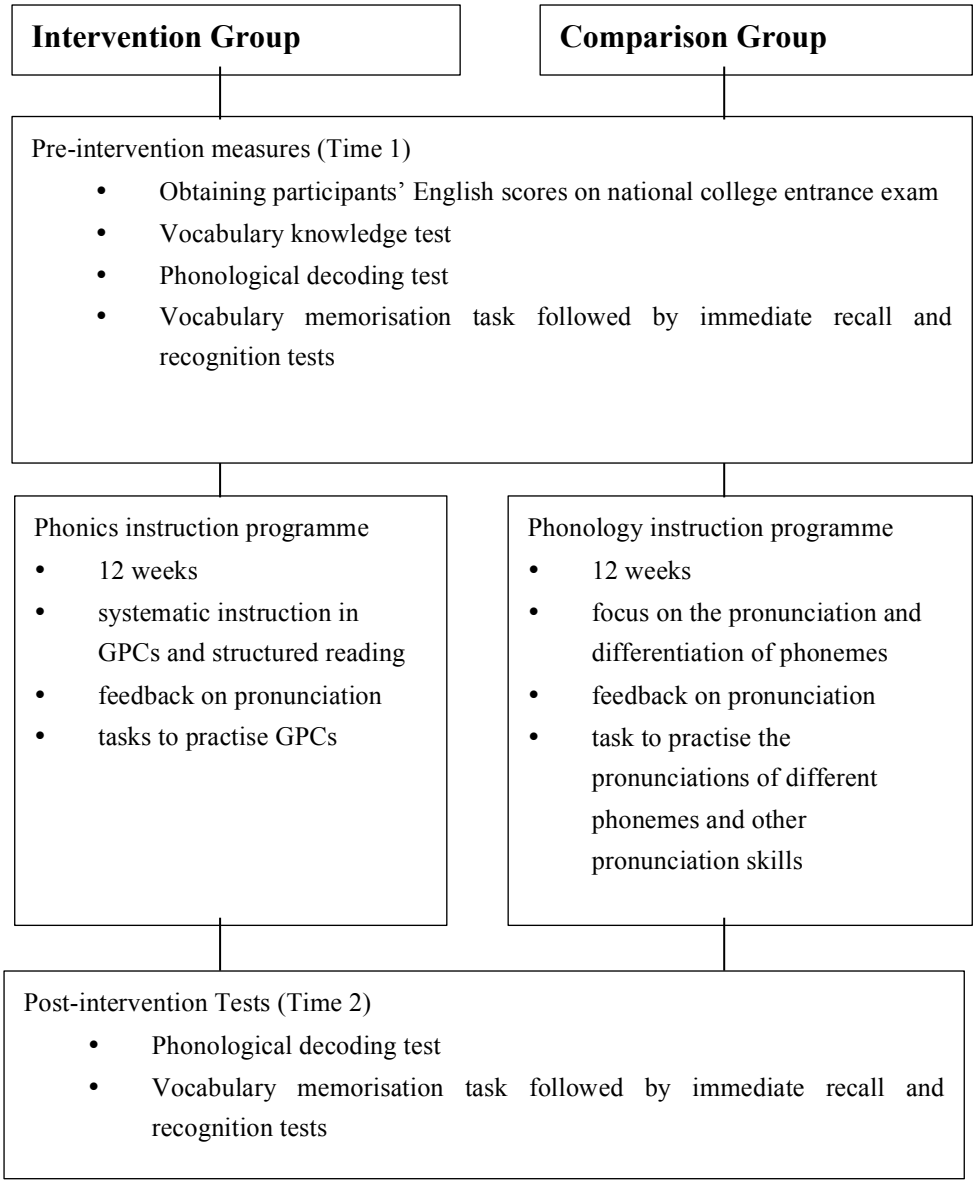
## 3.1 Overview

The purpose of this study was fourfold. Firstly, as discussed in Chapter 2, empirical evidence suggests that Chinese EFL learners generally perform less well in English decoding compared to their proficiency-matched counterparts with an alphabetic L1. Therefore, this study set out to examine whether the English decoding proficiency of Chinese EFL learners at university level can be promoted by a programme of systematic instruction. A quasi-experimental pre/post design was adopted, in which the intervention participants received a twelve-week systematic phonics instruction programme covering 101 English graphemes, while the comparison groups received a twelve-week phonology instruction programme with no focus on GPCs. Secondly, the characteristics and problems of participants' decoding before and after the instruction programme (henceforth t1 and t2) were also analysed, in order to pinpoint the strengths and weaknesses of participants' decoding at each time point, and to examine whether participants' decoding and the nature of their errors change as a result of the instruction programme (beyond any mean improvements in their overall decoding scores). Thirdly, the characteristics and problems of participants' decoding at t1 and t2 were also analysed, in order to pinpoint the strengths and weaknesses of participants' decoding at each time point, and to examine whether participants' decoding and the nature of their errors change as a result of the instruction

programmes. Fourthly, as there has been no experimental evidence demonstrating any causal relationship between L2 decoding and vocabulary learning, another key aim of this study is to explore whether such a causal linkage might exist. In order to examine this, the results of a vocabulary memorisation task followed by four recall and recognition tests at t1 and t2, namely an oral recall test, a written recall test, an aural recognition test and a written recognition test, were compared, in order to examine whether the instruction programme led to significant improvements in participants' ability to memorise new English vocabulary.

Figure 3.1 provides an overview of the research design.

*Figure 3.1. Overview of the research design*[2]




[2] Given that the end of the instruction programme was already the penultimate week of the academic term, a delayed post-test was not conducted.

This chapter is divided into 6 sections. Section 3.2 describes general considerations of sampling and the participants in this study. Section 3.3 discusses the content of the phonics instruction programme for the intervention participants and the phonology instruction programme for the comparison groups. The format of the two programmes is presented in section 3.4. Section 3.5 describes the data collection procedure. Finally, section 3.6 addresses ethical issues in this study.

## 3.2 Participants

This section is divided into three sub-sections. Section 3.2.1 discusses general considerations of sampling. Section 3.2.2 presents the initial plans for sampling, the problems encountered and the final approach taken. The final sample used in the study is described in section 3.2.3.

### 3.2.1 General considerations of sampling

This study targets Chinese university-level students for two reasons. Firstly, as previous research suggests, Chinese university-level EFL learners seem to encounter great difficulty in decoding English words, even though they have more than ten years of English learning experience (e.g. Hamada & Koda, 2008). As a result, university-level students are deemed to be appropriate participants for an intervention study whose aim is to promote English decoding proficiency. Secondly, as Chinese schools usually have a very full curriculum, it is hardly possible to conduct a long-term intervention study in primary school, middle school or high school,

considering that these students already have too many classes to attend and are constantly under high pressure relating to national college entrance examinations. University students, in contrast, have relatively fewer classes to attend and more flexible schedules. Hence, they are more likely to be open to the idea of participating in a long-term intervention study.

Currently, Chinese universities are divided into three categories based on admission criteria, namely first tier universities (*yi ben*, 一本), second tier universities (*er ben*, 二本) and third tier universities (*san ben*, 三本) (Min, 2004). First tier universities are the high-level universities which have the highest admission standards, followed by second tier universities and then third tier universities. All the universities admit students through the national college entrance examinations (*gaokao*, 高考), after which the first tier universities recruit students first, second tier universities second and third tier universities last. As a result, the three types of universities differ considerably in terms of students' academic performance.

The current intervention study was conducted in Wuhan, the capital city of Hubei Province and the largest city in central China with a population of roughly 10 million. Wuhan was chosen for two reasons. Firstly, as Wuhan is the researcher's hometown, it was convenient to go to the universities and collect data. Secondly, Wuhan is also the educational centre of central China with the most universities. Currently, there are 45 universities in the city of Wuhan, including 10 first-tier universities, 12 second-tier

universities and 23 third-tier universities (Higher Education Institutions, 2018).

Ideally, it would have been desirable to deliver the phonics instruction programme in as many classes as possible, randomly drawn from the 45 Wuhan higher education institutions, in the interest of generalisability. However this was unfortunately unattainable for the current study, due to resource constraints. Sample size was also constrained by other aspects of the research design. The first was related to the delivery of the instruction programme. Though it would be ecologically valid for teachers in the participating schools to deliver the instruction programmes, as in many other intervention studies (e.g. Woore, 2011), this was not achievable for the current study for reasons that will be discussed in section 3.2.2. As a result, all intervention groups and comparison groups were taught by the researcher herself for 12 weeks. As considerable time and energy had to be invested in the implementation of the intervention, the possible number of participating classes was limited to a great extent.

Secondly, the sample size of the current study was restricted by the nature of the decoding test. Though pen-and-paper decoding tests can be administered to a large number of participants quickly as in the case of Erler and Macaro (2011)'s study of the relationship between decoding ability in French and motivation for learning the language (N = 1735), the current study used a reading aloud test that was administered individually so that participants' pronunciation could be analysed and

compared. Clearly, much more time and labour was needed to both conduct the test and analyse the test results, and this in turn constrained the sample size of the study.

The researcher approached 7 universities to which she has some connection in the city of Wuhan, including two first tier universities, three second tier universities and two third tier universities. Only one first tier university, one second tier university and one third tier university agreed to participate. Though the data do not allow the findings to be generalised statistically to the wider population of Chinese university EFL learners, it is nonetheless large enough to conduct statistical comparisons between the groups (particularly given that participants within each university were allocated to both the intervention and comparison groups in equal number). Moreover, though the participating universities are not representative of the universities nationwide, they did include participants with varying levels of English proficiency (see section 4.1).

### 3.2.2 Initial plans for sampling and problems encountered in reality

The initial plan for this study was to recruit first-year non-English major students as participants. The main reason for this was that non-English majors do not receive specific training in English phonology; hence, there would have been reduced possibility of an interaction between the contents of their instruction programme and what they were being taught in their regular classes. In addition, the original plan was to conduct the intervention sessions outside their English classes, which would ensure

that the participants did not miss out on their existing English learning programme. However, none of the three universities that agreed to participate would allow the researcher to conduct a long-term instruction programme outside students' existing curriculum. The reasons were various. Firstly, it was difficult to arrange a fixed time each week for the intervention, as the students in a given English class took different content classes and had different schedules. Secondly, it was against the regulations of the local ministry of education to arrange classes at the weekend, even if the students voluntarily participate in these classes. Thirdly, as the students were not English majors, few of them had the incentive to follow a long-term instruction programme on English decoding in their own time, and thus participation could not be guaranteed.

Therefore, an alternative solution was proposed, which was to recruit first-year English majors as participants. As first-year English majors have a phonology class whose primary aim is to promote English pronunciation, the instruction programme was implemented during this scheduled class. Half the participants (the comparison group) would follow the regular curriculum of the phonology class, whereas the other half (intervention group) would instead receive the explicit phonics programme.

There were some advantages to this plan. Firstly, as the participants did not have to take extracurricular classes, the organisation was much easier and participation could be guaranteed. Secondly, as the participants were English majors, they were more

motivated to develop systematic knowledge about English GPCs, which could benefit their future study. However, some potential issues arising from this design needed to be taken into account.

Firstly, as the participants were all English majors, they had considerably more exposure to English than other non-English major students. English major students spend approximately 40 hours each week in developing various aspects of skills in English including listening, speaking, writing and reading. In contrast, non-English major students have only 3 to 4 hours' English classes each week. As a result, the English major participants in this study would clearly have more opportunities to practise the English GPCs knowledge they would learn in the instruction programme, compared to the non-English majors in the original plan. As a result, conclusions concerning the value of explicit phonics instruction for non-English majors cannot necessarily be drawn from this study, as the opportunities for practice afforded during the instruction programme for English majors may have had an effect on its effectiveness. However, this does not detract from the value of the study, since there are many such English majors studying EFL in Chinese universities. It simply means that any pedagogical implications of the study will relate to a different population of learner than was originally intended.

Secondly, as the comparison groups would receive phonology classes, it is important that they should not receive any explicit instruction in GPC knowledge, as this would

introduce a problem of contamination between the programmes, thus reducing the ability to draw conclusions concerning the effectiveness of the phonics instruction (because the difference between the intervention and comparison programmes would be diluted). A detailed discussion of the textbooks and the class activities in the two conditions is presented in section 4.3.

Thirdly, as the phonics instruction programme would take place in participants' existing phonology classes, randomly assigning participants to the intervention or the comparison group, which is deemed as the gold standard for conducting a true experiment by some researchers (e.g. Torgerson, 2002), was not possible, as the students were already grouped into different classes upon entering the university. Instead, whole intact classes are randomly assigned to intervention or comparison conditions, as in Woore (2011). However, a series of tests were conducted in order to check for confounding variables, and these are discussed in section 3.5. Further, note that the researcher conducted the teaching in both the intervention and comparison classes, thus removing any possibility of teacher effects as a confounding variable. Finally, readers are reminded that there were intervention and comparison groups within each university, which is a stronger design than allocating each condition to a different university (because in the latter case, the comparison between experimental groups would have been confounded by possible institution-level effects).

### 3.2.3 Describing the sample

The three universities participating in this study were one first tier university, one second tier university and one third tier university (henceforth University A, University B and University C). All the participants were first-language Mandarin speakers who had had been in Mandarin-medium education through high school in China. None of the participants had family members who were native speakers of English, as revealed by a questionnaire conducted before the data collection.

As the universities are very different from each other, four kinds of information are presented to describe them, namely the nature and size of the university, the characteristics of the student population, students' academic performance and the class arrangement of the university. All the information was obtained from the universities' websites.

In terms of the nature of the university, University A is a national high-level university funded by the Ministry of Education of China, while University B and University C are both funded by the local government of Wuhan. The three universities also vary in size: University A has roughly 36,000 students on roll, University B has approximately 18,000 students and University C has 11,000. However, all three universities recruit four classes of English majors each year, with roughly 20 to 25 students in each class. This represents the size of a typical class of English majors in Chinese universities based on the researcher's personal knowledge

as someone who studied English at university-level in the past. However, as the phonology class in University B is conducted in a 'big class', i.e. two classes take the class together, the number of participants in University B was twice that in University A or University C.

As for the characteristics of the student populations, most of the participants in the three universities come from central China, comprising Hubei, Hunan, Anhui, and Henan provinces. These provinces are generally considered to rank middle in terms of economic development in China, behind some southern and eastern provinces but better than the western provinces (Jian, Sachs & Warner, 1996). Given the economic discrepancy among different regions in China, educational resources also vary widely (Zhang & Kanbur, 2005). As a result, the English teaching quality in the central region has been claimed to be better than in the western region and worse than in the eastern and southern regions (Hu, 2005). Similarly, an inequality of educational resources is observed between big cities and small towns, with students in big cities having earlier access and more exposure to English compared to those in rural areas (Mok, Wong & Zhang, 2009). It should be noted that though most participants in the three universities come from the same region in China, the percentage of local participants, *i.e.* participants who come from Wuhan, the biggest city in central China in University B (35%) was higher than in University A (10%) or University C (20%), whereas the rest of the participants are mostly from small towns and rural areas.

As mentioned in section 3.2.1, the participants in the three universities also differed in terms of their academic performance, as measured by the national college entrance examinations. The participants in University A achieved the highest mean score, followed by the participants in University B and then University C.

Finally, the three universities also differ in terms of class arrangement. Participants in University A and B start to learn a second foreign language in their second year of university, which is the case for most English majors in China. In contrast, University C participants begin to learn a second foreign language in the first year. This is because in University C and some other tier three universities, most classes are completed in the first three years of university, so that the students can find internships in the last year, which promotes their competitiveness in the job market after graduation. As a result, the intervention participants in University C were learning French at the time of the instruction programme, and the comparison groups in this university were learning Japanese. Though clearly it would have been better to recruit two classes that were both learning the same foreign language, to eliminate this potential confounding variable, this unfortunately could not be achieved due to timetabling difficulties.

In summary the participants in the three universities cannot be said to represent the wider population of university English majors in China, as they are mainly from the same region. However, the participants in University A and B are arguably not

untypical of other first year English majors in central China. University C participants have the unusual feature of learning a second foreign language in their first year. Table 3.1 shows the number of participants in each class.

*Table 3.1 Number of participants in each class*

| | |
|---|---|
| University A | Intervention (N=24) |
| | Comparison (N=21) |
| University B | Intervention (N=47) |
| | Comparison (N=46) |
| University C | Intervention (N=23) |
| | Comparison (N=19) |

## 3.3 Intervention design

This section describes the programme of systematic phonics instruction created for the intervention group in this study, and the phonology course that the comparison group took.

### 3.3.1 Contents for the intervention participants

The GPCs to be included in the phonics instruction programme were firstly considered. Though the pilot study already showed that the consonant graphemes, most of which also exist in Pinyin and have fewer variations in decoding, pose less of a challenge to Chinese EFL learners than the vowel graphemes, it was still decided that a relatively comprehensive picture of English graphemes should be presented to the participants, rather than only the challenging ones as in the phonics instruction programmes of some other intervention studies (e.g. Woore, 2011). This is mainly

because in Woore's (2011) study, the French GPC instruction was integrated into participants' existing French class, and the participants received 10- to 15- minute instruction sessions over the course of 38 lessons. In contrast, the phonics instruction sessions in the current study were not integrated into participants' phonology class; rather, the phonics instruction itself replaced the phonology course for the intervention participants. As a result, it was determined that the phonics instruction programme should cover all the English phonemes and the corresponding graphemes, so that the participants in the intervention group did not miss out on any pronunciation instruction that the comparison groups would receive. A phonics instruction programme which provided systematic knowledge of English GPCs was thus needed.

Clearly, it would have been ideal if a phonics instruction programme specifically targeting Chinese EFL learners could have been used. However, English phonics programmes that the researcher could find are mostly designed for native speakers of English, especially for young children, in order to promote their reading and literacy skills in the early years of school. As a result, most of these programmes have the following characteristics in common.

Firstly, as the activities in these programmes are designed for young children in the early years of primary school, it is important they should not only be informative but also interesting and fun. For instance, the widely used phonics programme in the UK,

'*Jolly Phonics*' (Lloyd, 1992) introduces 'Inky Mouse and her friends, Bee and Snake' and engages children in multi-sensory activities such as colouring and practising writing graphemes. Though these child-centred learning activities are well suited for young learners and are effective in promoting young native English speakers' literacy skills (Slavin et al., 2009), they are less suitable for the participants in the current study, who were 17 to 18 years old at the time of the instruction. Moreover, as this kind of activity is not commonly seen in university classes in China, there is the danger of the participants realising the novelty of the class activities and being aware that they are taking part in an experiment. The possibility of Hawthorne effects might then potentially jeopardise the validity of the research.

Secondly, the structured reading materials in most of these programmes are likewise designed to suit the age and cognitive maturity of children aged between 4 and 7. As a result, the choice of vocabulary and the design of stories in these textbooks are less suitable for the participants in the current study, who were first year university students. The participants might find the reading materials juvenile, which would pose the risk of weakening the effects of the instruction programme, as the participants might be demotivated.

Thirdly, the pilot study showed that first year university students correctly decoded roughly 10 words out of the 28 pseudo words in the *Woodcock Reading Mastery Tests* (Woodcock, 2011). According to the data provided by the *Woodcock Reading*

*Mastery Tests* handbook, these results reflect the decoding proficiency of a lower-achieving Year 5 or Year 6 native English student (aged 9-11 years). As a result, the content of the phonics instruction programmes whose target audience is mostly Year 1 to Year 3 students might be too easy for the participants in the current study.

Taking these issues into consideration, the programme chosen for the phonics instruction in the current study was *Read Write Inc: Fresh Star*t (Miskin, 2011). This was for five reasons.

Firstly, Miskin (2011) was originally designed for struggling native English readers aged between 9 and 13 (Years 4 to 8) in Britain. Hence, this instruction programme suits the baseline decoding level of the participants in the current study.

Secondly, compared to other phonics instruction programmes created for children in the early year of primary school, the age gap between the target audience of Miskin (2011) and the participants in the current study is smaller. As the target audience of this programme is older children, the learning activities, such as reading graphemes, reading words, counting syllables and spelling words, are also commonly used teaching techniques in first year English majors' phonology class. As a result, the learning activities are deemed to be age-appropriate for the participants. Moreover, the structured reading in this programme is also better suited to older students so that

the participants in the current study would not feel that the stories are too juvenile. This was checked by an informal interview conducted after the pilot teaching, in which the participants were asked if they found the materials very different from those they used in other classes such as listening comprehension and intensive reading. The participants in the pilot study did not report anything unusual about these reading materials.

Thirdly, though Miskin (2011) is not tailored for Chinese university EFL learners, many activities in the instruction programme are well suited to address the problems experienced by Chinese students as mentioned in section 2.2. For instance, the activity of counting the number of graphemes in a word encourages the participants to form the habit of analysing intraword structure, which helps them to curb the tendency to see English words as inseparable wholes like Chinese characters (section 2.2) and lays the groundwork for proficient decoding. Moreover, this instruction programme also includes spelling tasks that help the students consolidate the newly learned GPCs; in these, the students are asked to spell a new word with the GPCs in which they have received instruction.

Fourthly, this instruction programme provides instruction in all 44 English phonemes and their spellings. As mentioned earlier in this section, the phonics instruction programme itself constituted the intervention participants' phonology class, the original aim of which was to help first year English majors distinguish and pronounce

the 44 English phonemes. As all the phonemes are covered in this instruction programme, the participants in the intervention group would not miss out on any pronunciation practice that the comparison groups would receive, the absence of which could pose a threat to their future study as English majors.

Finally, though Miskin (2011) is a relatively new phonics instruction programme, it has already received much positive feedback in terms of the effectiveness of promoting students' English GPC knowledge. For instance, Cox (2011) observed that after this training programme, the pass rate for the phonics screening check[3] rose from 67% to 80% in her school (Brompton-Westbrook Primary School, Kent). Moreover, this instruction programme has been used in Hong Kong with international students coming from more than 40 countries (including China), and has yielded promising results (Berry, 2015).

In addition, Miskin (2011) was compared with other phonics instruction programmes on the market, such as *Jolly Phonics* (Lloyd, 1992), *Success for All* (Wasik & Madden, 1995) and *PALS Reading* (Fuchs, Fuchs, Thompson, AlOtaiba, Yen & Yang, 2000). It was found that the format of these programmes is very similar, in that all include explicit instruction in English GPCs, with supplementary structured reading to help students to consolidate the phonics knowledge. In other words, the instruction programme chosen for the current study did not fail to cover any content that other

[3] The phonic screening check is a compulsory test for Year 1 pupils in Britain to confirm whether they have achieved a specified level of phonological decoding proficiency.

programmes provide.

Miskin (2011) consists of two kinds of lesson, which are called the 'phonics lessons' and the 'module lessons' respectively. The phonics lessons provide explicit instruction in the relationship between the 44 English phonemes and their corresponding graphemes. The module lessons provide structured reading materials to help the students consolidate the GPC knowledge presented in the phonics lessons. A detailed description of the two kinds of lesson is presented in the next two sections.

#### 3.3.1.1 The phonics lessons

44 phonemes and 101 corresponding graphemes, and 27 common word endings (e.g. *–tion*, *-al*), were explicitly instructed in the phonics lessons. All the graphemes and common word endings covered in the instruction programme are listed in Appendix 1.

All 44 English phonemes were covered in the instruction programme. Though the relationship between most phonemes and the corresponding graphemes was clearly explained, it should be noted that the schwa (/ə/) and its corresponding graphemes was not directly instructed; rather, the schwa was taught in combination with other phonemes in the common word endings. This is mainly because the spelling variations of the schwa are numerous and the rules are more difficult to summarise compared to other phonemes. Zhou (2002) lists the 13 graphemes that can be decoded

as the schwa, namely <a> <ar> <e> <er> <i> <o> <or> <oar> <ou> <u> <ur> <ure> and <our>. However, except for the cases of the common word endings such as <-er> <-or> and <-our>, it is very difficult to summarise the rules of the spellings of the schwa (Zhou, 2002). Another reason why the schwa was instructed together with common word endings was because the schwa is generally found in unstressed syllables, and many common word endings (e.g. *-el*, *-il*, *-al*, *-le*, *-ence*, *-ance*, *-ture*, *-our*) covered are unstressed syllables with the schwa in it.

It should also be noted that though the instruction programme presented a relatively comprehensive picture of English GPCs, it clearly did not, and could not possibly, cover all the probabilities of GPCs in the language, as English is an orthographically deep language (Geva & Siegel, 2000). This is especially true for the vowel graphemes, as the variations in their pronunciation are numerous (e.g. the grapheme <a> is pronounced as /æ/ in *bad* and /ə/ in *along*) compared to the consonant graphemes. As a result, only two types of vowel GPCs were presented in the instruction programme. Firstly, when the GPCs have well-established, canonical rules, these rules were taught to the participants. For instance, the correspondences between grapheme <a> and only two phonemes /æ/ and /eɪ/ were instructed, because in closed syllables <a> is pronounced as /æ/ and in open syllables it is decoded as /eɪ/. Though <a> can also be decoded as other phonemes, such as /ɑː/ in *after*, /ɒ/ in *was*, /ɔː/ in *all*, /e/ in *many*, /ɪ/ in *manager* and /eə/ in *various* (Gontijo *et al*., 2003), these possible realisations were not covered in the instruction programme, as the rules are less clear. Secondly, when a

grapheme can be decoded in several different ways, only its most common pronunciation was taught to the participants. For instance, the grapheme <ou> was instructed with only one corresponding phoneme /aʊ/, as this is its most common pronunciation with a probability of .36, while its other possible pronunciation, such as /u:/ in *group* (*p* = .22), /ʊ/ in *should* (*p* = .17), /ɔ:/ in *four* (*p* = .07), /ʌ/ in *country* (*p* = .06), /e/ in *tenuous* (*p* = .01), /əʊ/ in *soul* (*p* = .01), /ɜ:/ in *journal* (*p* = .004), and /ɒ/ in *cough* (*p* = .001) (Gontijo *et a*l., 2003) were not taught, as they are less common compared to /aʊ/ in English words. In other words, some simplification of the true picture was accepted in the phonics instruction programme so that the most important GPC rules could be taught more easily.

In fact, in the first session of the phonics instruction programme, the intervention participants were explicitly told that English is very different from Pinyin, as the correspondences between graphemes and phonemes are not fixed; that though many correspondence rules were to be learned in the instruction programme, some exceptions were inevitable. This is based on two observations from the pilot teaching stage. The first is that when the participants were told that there exist correspondence rules in English (as is the case in Pinyin), some of them took it for granted that these are the same 'one-to-one' fixed rules as in Pinyin and applied these rules universally without considering the orthographic context. An example is that in the first week of the pilot teaching, the participants were taught that <a> is decoded as /æ/ in closed syllables, with many examples such as *hang*, *sank*, *match*, *hatch* and *fang*. Then the

participants were provided with a structured reading material called *The thing from the black planet* with many words that contain the grapheme <a> (pronounced as /æ/) to consolidate this knowledge. However, when the participants were asked to read the article, some of them mechanically decoded <a> in all words as /æ/, even in words such as 'was' 'across' and 'along' which they already knew. The second observation is that some participants approached the researcher after class and expressed their confusion about the GPCs in English. For instance, some participants asked why the grapheme <a> is not decoded as /æ/ in some closed syllables, such as *was* and *journal*. As a result, it was determined that the participants should be explicitly told about the differences between the GPCs rules in English and in Pinyin at the beginning of the instruction programme, in order to discourage the participants from mechanically applying the rules to all English words, and prevent demotivating the participants when they realise that there are some exceptions to these rules.

The GPCs taught in each week of the instruction programme is presented in Appendix 2. It can be seen that, similarly to many phonology textbooks for English majors in China, the instruction programme teaches the six stops /b/ /p/ /t/ /d/ /k/ /g/, which are easy for Chinese speakers to pronounce as they are also phonemes in the Chinese phonology, and their corresponding graphemes, which map onto these sounds with relative consistency in the first week. Then the other consonant graphemes were taught from week 2 to week 9. It can be seen that for the phonemes that correspond to several different graphemes, the more common spellings were taught first, while the

other less frequently seen spellings were taught later. For instance, the phoneme /ʃ/ was firstly taught with the grapheme <sh> in week 2, and was not instructed with the graphemes <ti> and <ci> (as in words such as 'ambition' and 'delicious') until week 9. The vowel graphemes were instructed from week 2 to week 12, with the five vowel letters (<a> <e> <i> <o> <u>) and their canonical pronunciations in closed syllables taught first. Then the other vowel graphemes were taught from week 3 to week 10. Similarly to the consonants, the most common spellings of the vowel phonemes that correspond to several different graphemes was taught first, and the other less common spellings were taught later in the instruction programme. For instance, the phoneme /eɪ/ was firstly instructed with the grapheme <ay> in week 3, then taught with the grapheme <a-e> in week 7, and finally with the graphemes <ai> <ey> <aigh> and <eigh> in week 10. Some common word endings were taught in the last two weeks of the instruction programme.

It can be seen that the number of target graphemes in the instruction programme is very large. Though Miskin (2011) was originally designed to be instructed over 33 weeks, in which 3-4 graphemes are taught each week, this was unfortunately not attainable in the current study, considering the length of the instruction programme. Instead, the participants were instructed on average in 10 graphemes each week. This was argued against by Woore (2011), on the basis that teaching a large set of graphemes as a 'one off' may not guarantee that they are learned. Erler (2003) shares similar observations in her interview of Year 7 English learners of French. On this

basis, Woore (2011: 138) chose only 55 graphemes that were deemed to pose challenges to Year 7 English learners of French to include in his intervention study out of the total 122 GPCs. These concerns are clearly reasonable. However, the participants in the current study were older (17 to 18 years old) compared to those in Woore (2011) and Erler (2003), where they were 11 to 12 years old. Therefore, the participants in this study were more cognitively mature and should be able to process more information in a single class session. Given that they were English majors, they were also likely to have higher levels of motivation for learning to decode the foreign language. The appropriateness of the instruction programme schedule was checked by five weeks of pilot teaching prior to the intervention study, in which the contents of week 1 to week 5 of the instruction programme were taught to one non-participating class in each university. Observations based on the interaction with participants in class and informal interviews with some participants after class confirmed that the content of each week was not overwhelming for the participants. Moreover, It can also be seen in Appendix 2 that some graphemes in the programme were repeated in several weeks. For instance, the grapheme <se> was firstly taught in week 5, but was repeated in week 7 and week 8 with other graphemes that share the same pronunciation, namely <c> and <ce>. This took into account the suggestions made by Woore (2011) and Erler (2003), that the graphemes that were learned in the previous week were also reviewed at the beginning of the class in the following week, in order to examine whether the participants had any problems with the previously learned graphemes.

Another issue of great importance is that as the participants in this study were non-native speakers of English, it cannot be assumed that they had a reliable English phonological system at their disposal, like the native speakers for whom the instruction programme was initially designed. Though the participants had roughly 10 years of English learning experience at the time of the instruction programme and had probably encountered all the English phonemes in many different words, they had never been explicitly instructed about the English phonological system prior to the intervention. Unlike for L1 beginner-readers, the participants may be learning both the phoneme knowledge and the GPCs knowledge concurrently. This development of their phonological knowledge can be argued as an added benefit (by-product) of their learning to decode.

The importance of understanding the L2 phonological system for learners of the language is observed in many studies. For instance, Walter (2008) points out the crucial role of reliable L2 phonological inventories in L2 reading in her study of French learners of English, as weak representations of English phonemes resulted in poor comprehension of the language, even the written language. This may well be true for vocabulary learning, as an incomplete phonological repertoire might prevent the learner from correctly decoding new words in the first place, and the faulty grapheme-to-phoneme conversion might result in activation of the wrong grapheme when recalling words. For instance, if a Chinese learner of English has a poor representation of the English phoneme /θ/, which does not exist in Chinese, he/she

might confuse it with phoneme /s/ and decode the word *think* as /sɪŋk/, which can lead to the wrong recall of the word's written form as *sink*. As a result, when a new grapheme was introduced in the instruction programme, its pronunciation rules and the comparison with other similar-sounding phonemes in both English and Pinyin were also provided to the participants. This was to strengthen participants' understanding of the English phonological system and help them develop their knowledge of the English phonemes to which the target graphemes correspond. Given that the comparison groups were also learning the phonetic symbols of the 44 phonemes in the International Phonetic Alphabet (IPA) (see section 3.3.2), the IPA symbols were also provided when a new grapheme was introduced. Many participants already possessed some knowledge of the IPA symbols of English phonemes before the instruction programme. In fact, an informal investigation was conducted in the first class in which the participants were asked to raise hands if they knew all the phonetic symbols, and more than 80% of the participants in each university did so. In order to examine participants' command of the 44 phonemes and their IPA symbols, the intervention and comparison groups' scores in the phonology class at the end of the intervention as examined by each university[4] were compared. The results showed that the participants in the two conditions in each university did not differ significantly in terms of the test results[5], indicating that the phonics instruction programme was as successful as the phonology instruction programme (see section

[4] Though each university conducted their own exam for the phonology class, the contents of the tests were very similar, which include the production and recognition of English phonemes, production and recognition of IPA symbols, and reading a short paragraph in English.

[5] University A: $F(1, 43) = .63$, $p = .59$, University B: $F(1, 91) = 1.58$, $p = .27$; University C: $F(1, 40) = 1.68$, $p = .29$

3.3.2) in promoting English phonological knowledge and IPA knowledge.

In summary, the phonics lessons provided systematic instruction in English GPCs. Considering that the participants are non-native speakers of English, the differences between the GPCs in Pinyin and in English were explicitly explained, and the pronunciation and instruction also focused on the differentiation of English phonemes.

#### 3.3.1.2 The module lessons

After each phonics lesson, the intervention participants received a corresponding practice lesson called the module lesson in the Miskin (2011) programme. The module lesson provided structured reading material focusing on the graphemes taught in the previous phonics lesson. Each module lessons consisted of one article, which was roughly 300 to 400 words in length. In each week of the instruction programme except for the first week, the participants were given three articles each week. The participants were asked to firstly read the articles out loud by themselves, and the researcher would walk around in the classroom to ensure that they were reading aloud. After that, one participant would be randomly asked to read one paragraph of the article to the whole class, and feedback was given to his/her pronunciations. Though not every one would be selected to read to the whole class given the limited class time (especially for the University B class which had 47 students), the rest of the class still

had the opportunity to learn when their classmate's pronunciations were corrected.

The contents of the module lessons were mostly interesting short stories and some non-fiction articles. In the five weeks of pilot teaching, the participants showed interest in the articles and answered questions regarding them actively.

It should be noted that though the instruction programme was originally designed for Year 4 to Year 8 students in the UK and do not include any difficult vocabulary in the structured reading materials for these learners, a few words were still new to the participants in this study, who were non-native speakers of English. This problem was observed in the pilot teaching, as some participants tried to check the meanings of some words in the dictionary. In order to find out the quantity of new vocabulary in these articles, the participants were asked to report the words that were unfamiliar to them in each article. It was found that each article contained approximately 3 to 5 new words, though the participants also confirmed that the new vocabulary did not inhibit their comprehension of the articles. This echoes the findings in Hseueh-Chao and Nation (2000), where EFL learners who read English texts with less than 2% of unknown vocabulary were able to perform well on reading comprehension. However, in order to prevent the participants from spending time in class checking up meanings of unfamiliar words, the reading materials were provided to the participants a week early, so that they could look up new words in advance.

As the participants who received the instruction programme also had exposure to some new vocabulary, it is of importance to examine whether this constitutes a confounding variable. However, there is no reason to suppose that the participants in the intervention group learned more words from the class compared to those in the comparison group, as the materials for the comparison groups also included new vocabulary. This is further discussed in Section 3.3.2.

### 3.3.2 Contents for the comparison groups

The textbook for the comparison groups in University A and University B was *English Pronunciations and Intonations for Communication* (Wang, 2005), which is one of the most frequently used textbooks for English majors' phonology courses in Chinese universities (Cai, 2004). The textbook for the comparison groups in University C was *Better Pronunciation for Communication* (Liu, 2013), which is a relatively new textbook on English phonology.

The content of these two programmes was very similar. Firstly, the basic concepts of syllables, stress and rhythm were introduced. Secondly, systematic instruction was provided in the 44 phonemes and their corresponding phonetic symbols, with emphasis on the pronunciation of these phonemes and the distinction of similar-sounding phonemes. Thirdly, pronunciation skills, such as linking, intonation and stress patterns were also provided. The only difference was that the skills of

making speeches and public speaking were only included in the programme for University C but not for University A or University B.

The contents of the phonology classes for the comparison groups in Universities A and B, and University C are given in Appendix 3.

It can be seen that the focus of these two phonology instruction programmes was on the pronunciation and discrimination of the 44 English phonemes, which accounted for half of the course time. All the English phonemes were introduced through phonetic symbols in the IPA. Instruction in phonetic symbols is optional in the high school curriculum (Ma, 2007), and many teachers do not teach them, worrying that the introduction of an additional set of written symbols might confuse the learners. However, many participants in the comparison group already possessed some knowledge of the phonetic symbols. This is because the phonetic symbols are often provided in the vocabulary lists in their English textbooks, and many students had been able to work out what the phonetic symbols represent by comparing them to the pronunciations of the words they learned. This was again confirmed by an informal investigation in the first class, in which the participants were asked to raise their hand if they knew all the IPA symbols, and more than 80% of the participants in each university raised their hands.

Phonetic symbols are systemically introduced in English majors' phonology classes for two reasons. Firstly, it is important for the English majors to develop phonemic

awareness, so that they can have a clear understanding of the pronunciation of each phoneme and achieve better pronunciation of English words. Secondly, knowing the phonetic symbols helps students to learn the pronunciations of new words, as they can look up the words in a dictionary and read them according to the phonetic symbols. It should be noted that neither of the two phonology instruction programmes provided instruction on English GPCs.

The teaching in weeks 1 to 5 of the two programmes was piloted. One observation from the pilot teaching was that some comparison groups also tried to look up unfamiliar vocabulary in class. Similarly to those who received phonics instruction during the pilot stage, the participants here were also asked to report any new vocabulary they encountered. It was found that the new vocabulary that the participants reported mostly appeared in the exercises, such as reading the 'lead-in' articles and comparing the minimal pairs. The participants reported roughly 10 new words in each week's class, which was similar to the number of new words encountered weekly by the intervention participants in the phonics instruction programme. As both the intervention and the comparison groups encountered similar numbers of new words in their class each week, there is no reason to suppose that either programme led to the acquisition of more vocabulary (see section 3.5.1.2). In order to prevent the participants in the comparison group from spending time in class looking up the meanings of unfamiliar words, they were also asked to preview the materials a week early.

In summary, the phonology instruction programmes for the comparison groups focused on the pronunciation and discrimination of the 44 English phonemes, and also provided pronunciation tips in English relating to issues such as stress, linking and intonation. No instruction in the GPCs of English orthography was provided.

Based on the previous two sections, it can be seen that one major difference between the phonics instruction programme and the phonology instruction programme is whether English GPC knowledge was explicitly taught. In the phonics instruction programme, the GPC knowledge was explicitly explained to the intervention participants, and the class activities centred on establishing the links between graphemes and phonemes, with the help of the structured reading materials (see appendix 4). In the phonology instruction programme, no explicit instruction on English GPC was given, and the activities centred on the pronunciation of the 44 English phonemes and the distinction of similar-sounding phonemes (see appendix 5), whereas participants' attention was not directed towards the spelling forms of these phonemes. Nonetheless, the comparison participants still had the chance to learn about English GPCs from the phonology instruction programme, given that they could hear the English pronunciations and saw the written words together. Hence, the major difference between the two instructional programmes boils down to the level of explicitness the GPC knowledge was instructed with.

### 3.4 Format

Two randomly selected intact classes from each of the three universities were chosen as participants in this study. One class in each university was randomly assigned to take part in the phonics instruction programme, and the other class participated in the phonology instruction programme. Both the phonics instruction programme for the intervention participants and the phonology instruction programmes for the comparison groups were delivered in 12 lessons spanning 12 weeks. Each lesson lasted 100 minutes, with a 10-minute break after the first 45 minutes. The duration of the instruction programmes was the maximum length of time left in the first term of the university, which is when the phonology course for English majors is scheduled. This also allowed for time to administer the pre- and post-tests and to pilot the teaching of the two programmes and the testing instruments.

This section describes how the phonics instruction programme and the phonology instruction programmes were delivered. As the format of the two phonology instruction programmes for the comparison groups were basically the same, they are discussed together. Figure 3.2 presents an overview of the core activities in each programme. Some of the terms will be explained more fully in the text below.

*Figure 3.2 Outline of activities for the phonics instruction programme and the phonology instruction programme*


It can be seen that there were three core activities in the phonics instruction programme. Firstly, the written form and the phonological form of the new graphemes were presented to the participants. Secondly, the newly learned graphemes

were practiced through various tasks, including reading words containing the target graphemes and thinking of words containing the target graphemes. The participants were also ask to count the number of syllables in the words, segment the words into graphemes and blend them together. Some unfamiliar words, usually the new words in the structured reading materials (the 3 articles as mentioned in 3.3.1.2), were also presented to the students, and the students were asked to decode them using the GPCs they had been taught. Thirdly, structured reading materials were used so that the participants could further consolidate their knowledge of the newly-instructed GPCs when reading the articles aloud.

The phonology instruction programme for the comparison groups had two core activities. The first was the introduction of new phonemes or pronunciation skills, such as linking, intonation and rhythm. A detailed explanation is provided in the participants' textbooks, and the teacher walked the participants through the content. The second was the repeated practice of the newly-taught phonemes or pronunciation skills, including tasks such as reading words and sentences aloud individually and in chorus and listening to the recording and repeating.

In both the phonics instruction programme and the phonology instruction programme, participants were given opportunities to practise the graphemes, phonemes or pronunciation skills that had just been instructed through various tasks, such as reading the articles aloud individually and in chorus, and they were able to receive

explicit feedback on their pronunciation from the teacher.

A key difference between the two programmes is that the audio recording of the content in the textbook by a native speaker of British English was available for the phonology instruction programme, but was absent for the phonics instruction programme. This is because the phonics instruction programme was originally designed to be instructed by native speakers of English, and to be delivered to English native speakers as well; thus, in such a context, a recording of the learning materials is unnecessary. As a recording was unavailable, the teacher, who was the researcher, herself had to model the pronunciation of the graphemes and asked the participants to repeat after her. Though the researcher had been learning English for more than twenty years and had been living and studying in Oxford for four years, it still could not be guaranteed that her pronunciation of the graphemes was exactly the same as that produced by native speakers of English. This limitation of the phonics instruction programme should be acknowledged. However, note that when assessing participants' pronunciations of words (when scoring the decoding test), a lenient approach was taken to foreign accent (see section 3.5.2.3), so that the phonics group was not disadvantaged by the absence of a native spoken model.

Both the phonics instruction programme for the intervention participants and the phonology instruction programme for the comparison groups were delivered by the researcher herself, which effectively rules out teacher effects as a potential

confounding variable. Though the researcher had not previously worked as a full-time teacher, she had tutored university-level Chinese students before and had 4 years of teaching experience.

As this study is an intervention study, it is also of great importance to consider the possibility of Hawthorne effects, i.e. where the participants in the intervention group realize that they are being studied and thus change their behaviour (e.g. Chiesa & Hobbs, 2008). Specifically, Hawthorne effects are demonstrated from three perspectives (Adair, Sharpe & Huynh, 1989). Firstly, participants in the intervention group receive special attention from the researcher. Secondly, participants in the intervention group take part in some novel activities which are not commonly seen in other classes. Thirdly, participants are aware that they are part of an experiment.

First, as the phonics instruction programme and the phonology instruction programme were both delivered by the researcher, this ensures that the participants in both conditions received equal attention from the researcher.

Second, as mentioned previously, the two programmes of instruction were piloted in context for five weeks. The purposes was to receive feedback from the participants as regards to the appropriateness of the class activities and the amount of content in each class, as well as accumulate some teaching experience, as the researcher had not previously worked as a full-time teacher. The pilot of the phonics instruction

programme was conducted in all the three participating universities, and the pilot of the phonology instruction programme was only conducted in University A and University C, as there was a conflict in the schedule in University B. As previously mentioned, the participants who received the phonics instruction programme in the pilot stage did not report anything unusual about their class activities or reading materials, and felt that the amount of content was acceptable for them. This was also true for the participants who received the phonology instruction programme. The pilot teaching was also observed by the teachers who instruct the phonology course at these universities, who also provided valuable advice on how to arrange class activities and manage the class. The teachers also did not find the activities and reading materials in the two programmes unusual. As a result, it can be concluded that there was nothing novel in terms of the format of the two programmes.

Finally, as both the phonics instruction programme and the phonology instruction programme took place in participants' phonology class, there was no reason to suspect that the intervention participants were more aware of taking part in an experimental study compared to the comparison groups. In addition, though the participants were asked to sign a consent form prior to the study, their consent was only related to taking pre- and post-tests, in which both the intervention and comparison groups participated.

In summary, the current study was designed to largely avoid the possibility of

Hawthorne effects. Nonetheless, it is worth noting that though the participants in the two conditions were from different classes and did not take any lessons together, there was still a chance that the participants in the two conditions in the same university would talk to each other about the content of their class. This possibly constitutes as a confounding variable. However, during the course of the intervention, the researcher did not see or hear any evidence of participants doing this.

## 3.5 Data collection

This section describes the data collection procedure. Firstly, the collection of baseline information, namely the national college entrance English exam and British Picture Vocabulary Scale scores are presented in section 3.5.1. Then, the phonological decoding test is discussed in section 3.5.2. Finally, the vocabulary memorisation tests are discussed in section 3.5.3.

### 3.5.1 Baseline information on the participants

The results of the two baseline tests served to contextualize participants' decoding scores and to examine whether there was any difference between the intervention and comparison groups which could confound the findings.

#### 3.5.1.1 National college entrance English exam (NCEEE)

Participants' scores in the NCEEE were obtained from the three participating universities as an indicator of their English proficiency. This exam was chosen to evaluate participants' English proficiency for three reasons. Firstly, as this test is taken by all high school students in China at the end of the last month of high school, it located participants' English proficiency level within the wider population of high school graduates in China and will thus facilitate comparisons of the findings with those of future studies. Secondly, this test is a relatively comprehensive measure of English proficiency. In the NCEEE participants' listening comprehension, reading comprehension and writing were examined, though oral speaking was not tested. Clearly, it would have been desirable if participants' scores in each individual section of the test had been obtained, but this was unfortunately not possible. Thirdly, as discussed in section 3.2.1, the participants in this study came from three universities that are distinctly different in terms of students' academic performance. As a result, it is highly likely that the participants in the three universities also differed in regard to their English proficiency. Hence, the examination of the NCEEE scores of participants in each university served to determine whether such differences indeed existed, and if so, whether participants with different levels of English proficiency were affected differently by the phonics instruction programme.

Though the NCEEE scores served as a useful indicator of participants' English proficiency, it should be noted that the test was taken three months prior to the current study. Hence, it is possible that the NCEEE scores did not accurately reflect

participants' English proficiency at the time of the study, as some participants might have participated in English learning classes or have done some self-study before they entered university. As a result, another important indicator of participants' English proficiency, which is vocabulary knowledge, was also measured a week before the current study began.

#### 3.5.1.2 British Picture Vocabulary Scale (BPVS)

Participants' vocabulary knowledge was measured for four reasons. Firstly, research suggests that that participants EFL learners' vocabulary breadth plays an important role in their comprehension of English input (e.g. August et al., 2005). As the phonics instruction programme for the intervention participants and the phonology instruction programme for the comparison groups were both conducted in English, it was crucial to examine whether the intervention and comparison groups were matched in terms of vocabulary breadth. Secondly, as this test was conducted a week prior to the start of the intervention, it served as a complementary and current indicator of participants' English proficiency. Thirdly, it is intuitively true that larger vocabulary might potentially contribute to EFL leaners' English decoding proficiency, as the learners have a larger bank of known words on which to draw for either implicit learning of regularities and patterns in GPCs, or for consciously making analogies to help work out pronunciation of unfamiliar words. The importance of vocabulary breadth in decoding is also documented in many studies. For instance, White et al. (1990)

compared the vocabulary breadth and decoding proficiency of native English speaking students in three different elementary schools in the US, and found that the school with the largest mean vocabulary size also produced best results in decoding English words. Finally, there is a possible relation between participants' existing vocabulary knowledge and their performance on a vocabulary memorisation task, as participants with a large vocabulary might be more attuned to the intraword structure of English words, which in turn may facilitate their memorisation of new English words. For all these reasons, it was important to assess participants' vocabulary size and to check for differences between the groups on this variable.

The BPVS was used to assess participants' vocabulary breadth. This test was chosen because it is a standardized vocabulary measure in the UK (Dunn & Dunn, 2009), and also widely used to examine the vocabulary breadth of EAL and ESL students worldwide. For instance, Beech and Keys (1997) successfully used the BPVS to examine the English vocabulary knowledge of Asian children aged between 7 and 8. Though this test is originally designed for native English speakers up to 16 years old of age, it is also frequently used with adult learners. For instance, Howlin, Goodes, Hutton & Rutter (2004) used the BPVS to examine the vocabulary breadth of English native speakers with autism aged between 29 and 64 in their study. The format of the test, which involved listening to English words and selecting the right picture, was deemed to be appropriate for the participants in the current study, who were only a few years older (17- 18 years old) than the originally intended target of BPVS.

Moreover, the difficulty of the test was deemed to suit the level of the participants in the current study, as they were non-native speakers of English whose vocabulary breadth would probably not exceed that of a young native speaker of English. This was confirmed by the analysis of the BPVS scores in section 4.1, as none of the participants performed at ceiling on the test.

The BPVS was conducted one-to-one in a quiet room. The participants heard an English word and were asked to select one picture (from four choices) that best illustrates the stimulus' meaning. As the test used aural stimuli, a confound with decoding was avoided.

### 3.5.2 Phonological decoding test

Participants' English decoding proficiency was measured at both t1 and t2 in order to examine whether the phonics instruction programme led to significantly greater progress in terms of decoding proficiency. Clearly, it would have been ideal if a standardized English decoding test for Chinese EFL learners could have been used in the current study; however, a manual search of the literature did not reveal such a test. As a result, the current study used a section of the Woodcock Reading Mastery Tests (Woodcock, 2011), which includes 28 pseudo words as the test material. This section describes the decoding test from four perspectives: the content of the test (section 3.5.2.1), the format and administration of the test (section 3.5.2.2), the processing of

the data generated in the test (section 3.5.2.3) and the validity and reliability of the test (section 3.5.2.4).

### 3.5.2.1 Content

The word attack section of the Woodcock Reading Mastery Tests was chosen as the test material as it is a standardised test of English decoding proficiency for native speakers of English. It has been widely used in the US and the UK to measure the decoding proficiency of native speakers of English aged between 4 and 79. Though the test was not originally designed for EFL learners in China, it has been widely used in research involving non-native speakers of English, including Chinese students (e.g. Hamada & Koda, 2010).

In the decoding test, participants were asked to read 28 pseudo words of increasing difficulty. The test started with simple consonant-vowel combinations, and progressed to pseudo words with multiple syllables. The two test forms, which were used as pre- and post-test respectively, are presented in Table 3.2.

*Table 3.2 Stimuli in the phonological decoding test*

| | |
|---|---|
| Form A | bab, op, dee, bim, tay, yee, pog, shum, plip, dud's, whie, bufty, vunhip, knaf, twem, adjex, yeng, laip, zirdn't, straced, cedge, wrey, whumb, knoink, bafmotbem, monglustamer, pnir, ceisminadolt |
| Form B | bab, op, ree, raff, dat, glack, hend, weaf, chur, tayed, ful's, rejune, weat, sess, depine, wrault, throbe, gouch, brecked, darlanker, cigbet, mancingful, squow, cyr, quiles, untroikest, pelnidlum, byrcal |

The two test forms have been used on a wide population of native English speakers, and the results indicate that, on average, the difference in the number of correctly decoded words between the two forms is less than one. In other words, there is reason to consider them to be of equivalent difficulty. However, as the test has not been standardized on the population of Chinese EFL learners, it was unclear whether the two test forms were of similar difficulty for the specific participants in the current study.

One possible solution would have been to use the same test form at both t1 and t2, following Woore (2011: 155), thus eliminating the possibility that the pre- and post-tests were of different difficulty to the participants. However, this was deemed to be less desirable for the current study for two reasons. Firstly, the possible practice effect resulting from using the same test form at both times cannot be neglected, as the time between t1 and t2 was considerably shorter in this study (12 weeks) compared to Woore (2011) (7 months). The participants might memorise some of the test items and discuss with other participants, which could impact their decoding at t2.

More importantly, the counterbalanced administration of the two test forms provided a more comprehensive coverage of English graphemes and phonemes than administering either of them. Altogether, 67 graphemes and the 41 corresponding phonemes appear in the two test forms (the only graphemes lacking were those corresponding to /ð/, /eə/ and /ʊə/). As some graphemes, such as <ar>, <ey> and <dge> appears in only one test form but not the other, using one test form at both times sacrifices the opportunity of investigating development in participants' decoding of these graphemes. As a result, it was decided that the two different test forms would be used as pre- and post-test in this study.

In order to examine whether the two test forms were indeed likely to be of similar difficulty to the participants in the current study, the decoding test was piloted on 10 participants randomly drawn from each of the three universities (and who did not go on to participate in the main study). Clearly it would have been desirable to pilot these test forms on a larger number of participants, but this was not achieved due to the timetabling difficulties. Each participants were asked to read both Form A and Form B of the decoding test, and their pronunciations were scored at the word-level. It was found that the number of correctly decoded words in Form A and Form B was very similar: the participants achieved a mean score of 10.29 (*SD* = 1.57), and a mean score of 10.77 (*SD* = 1.38) in Form B, and there was no significant difference between the two scores as measured by a t-test, $t(29) = -.77$, $p = .45$.. This suggests that the two forms of the decoding test are of similar difficulty for the participants in

this study.

All the graphemes in the test were instructed in the phonics instruction programme except for <pn> in Form A. This is further discussed in section 3.5.2.3.

### 3.5.2.2 Format and administration

A split-block counterbalanced research design was adopted in the decoding test. The participants were randomly assigned to read one of the two test forms at t1, and read the other one at t2. It should be noted that total randomization was deemed inappropriate for the current study, as the participants came from three very different universities and had different levels of English proficiency. As a result, the randomization was conducted based on class: half of the participants in each class took Form A at t1 and Form B at t2, and the other half did the opposite.

The decoding test was conducted individually in a quiet classroom. Before the test began, oral instructions were provided to the participants by the researcher in Chinese. The participants were assured that their performance on the test would not affect their scores in this class in any way. They were also told that it is natural that they did not know any of these words as they are pseudowords, and they should just read them as they thought the words should be pronounced.

The 28 words were presented on A4 paper in large, bold, lower case Times New Roman font. The order of the test items was the same as shown in Table 4.5, which progressed from easy pseudowords to difficult ones. This order is the same as intended by the Woodcock Reading Mastery Tests. Another possibility that was considered was to present the test items in a randomised order so that the participants would not be demoralised, as they got progressively more difficult. This was tried out in the pilot study, where the test items were presented on a computer screen in a totally randomised order. However, it was noticed that two participants became tongue-tied after the word *monglustamer* and *pelnidlum* appeared as the first test item, and failed to decode the much simpler words *bab* and *raf* which appeared next. As a result, the test items were presented in a fixed order in the decoding test. Observations made during the test procedure suggested that the participants were not demoralised by the progressively more difficult test items; in fact, all the participants finished the test.

The participants were asked to read aloud all 28 test items on the paper. It should be noted that the discontinuation rule originally introduced in the Woodcock Reading Mastery Tests, which requires the examiner to stop the test after the examinee wrongly decodes four test items consecutively so as to avoid demoralisation, was not followed here. This was because one main purpose of the current study is to examine the strength and weakness of participants' decoding at the two time points, and reading through the whole list of test items could provide more material for analysis.

Moreover, it was observed in the test process that no participant actually correctly decoded a test item after four or more consecutive wrong answers. This was again confirmed in the analysis of participants' decoding. This is expected, as the test items became progressively harder, which made it hardly possible for the participants to correctly decode a more difficult word after failing to decode the previous four easier ones. As a result, though the original discontinuation rule was not followed, the whole-word level decoding results de facto did not violate this rule.

The participants were asked to read through the word list at their own pace. This was different from the method in many other studies, where the test items appeared only for a fixed and usually short period of time (e.g. 5000 ms in Hamada & Koda (2008); 2500 ms in Hamada & Koda (2010)). Though the importance of rapid decoding in reading comprehension is documented in many studies (e.g. Aarnoutse, Leeuwe, Voeten & Oud, 2001; Parrila, Kirby & Mcquarrie, 2004), a fixed time interval was not imposed in the current study for two reasons. Firstly, as the length of the test items was vastly different, it was unfair to expect the participants to decode *op* and *monglustamer* in the same period of time. Secondly, some researchers point out the importance of conscious processing in order to avoid the automatic triggering of L1 processing mechanisms in decoding L2 words (e.g. Woore, 2011). This is echoed in the findings of Erler (2003), where the English learners of French who took the longest time to decode the test items actually produced the best pronunciations. Similarly, Li (2012) also found in the study of Chinese EFL learners with high

English proficiency that some of the most proficient decoders spent the longest time working out the pronunciations. More importantly, they commented that this was because they needed time to think about how these words should sound in English, without resorting to the Pinyin GPCs in their L1 repertoire. As a result, a time limit was not imposed for fear that it would put unnecessary pressure on the participants and encourage automatic L1 processing of the test items.

Participants' decoding was audio-recorded using a digital recorder with built-in microphone. Given the limited time available for data collection, half of the participants' decoding was recorded by the researcher, and the other half by another teacher at the university, who taught the participants in a different class. In order to ensure anonymity and unbiased scoring, the names of the participants were replaced by identification numbers before analysis.

Given that the end of the instruction programme was already the penultimate week of the academic term, a delayed post-test was not conducted.

#### 3.5.2.3 Processing of the test results

The phonological decoding test was scored at two levels, namely the whole-word level and the grapheme level. Though the Woodcock Reading Mastery Tests are originally scored only at the whole-word level, the view taken in this study is that this

holistic scoring does not precisely reflect participants' knowledge of English GPCs. For instance, if a participant decoded the word 'bufty' as /bju:fti:/, it can seen that the participant only had a problem with the grapheme <u>, but correctly decoded the other four graphemes <b> <f> <t> and <y>. Scoring this pronunciation as zero at the whole-word level overlooks the participant's command of these four graphemes. Moreover, the grapheme-level scoring directly pinpoints the graphemes in which the participants' strength and weakness lie at each time point, and allows for examination of whether these problems have been solved by the phonics instruction programme.

Following Woore (2009), the unit of the grapheme (e.g. <a>) is called a 'grapheme type', and the individual representations thereof in the test items (e.g. the specific occurrences of <a> in words such as 'bad' and 'along') are referred to as 'grapheme tokens'. There were 114 grapheme tokens and 47 grapheme types in Form A, and 121 grapheme tokens and 49 grapheme types in Form B. Altogether, 235 grapheme tokens and 67 grapheme types appeared in the decoding tests. The phonological realizations of the stimuli in the decoding test that were judged to be acceptable are presented in Appendix 4. As mentioned in section 4.5.2.1, all the GPCs appearing in the decoding test were taught in the phonics instruction programme, except for the {<pn> → /n/} in Form A. As a result, the grapheme <pn> was excluded from the analysis.

When processing the data from the decoding test, it was found that most participants pronounced the test items in a way such that the phonemes in their pronunciation

could be clearly mapped onto the graphemes in the test items. As the participants were non-native speakers of English, some degree of Chinese colouring in the decoding was inevitable and thus permitted. For instance, decoding the grapheme <r> as /ʐ/, which is how it is decoded in Pinyin, was deemed acceptable, as it sounds similar to the English phoneme /r/ and there is little likelihood that other English phonemes can be confused with /ʐ/. Similarly, pronouncing the grapheme <sh> with Chinese colouring as /ʂ/ was also considered acceptable. However, decoding the grapheme <th> as /s/ was deemed unacceptable: though /s/ is the phoneme in Pinyin that most closely resembles the English phoneme /θ/, there are other English graphemes that should be decoded as /s/, and failing to make a distinction between these two phonemes can result in poor comprehension and wrong representations of words containing them.

Each test item was scored twice by the researcher. Firstly, the decoding was scored at the grapheme level in which one correctly decoded grapheme token was awarded one point. Then, the decoding was again scored at the whole-word level, in which one point was given only when every grapheme token in a test item was correctly decoded. As a result, two sets of scores were produced for each participant in the decoding test.

As speed is also argued to be an important indicator of decoding proficiency (Wolf, Bowers & Biddle, 2000), the time taken to decode stimuli is measured in many studies, often operationalized as the time between appearance of the stimulus and the

onset of participants' decoding (e.g. Hamada & Koda, 2008). Clearly, it would have been ideal if the reaction times of each individual test item could have been measured; however, this was unfortunately not possible in the current study, as the test items were presented to participants on paper. Instead, the time between the onset of the first stimulus, which is after the examiner said 'start', and the completion of participants' final pronunciation of the last test item was recorded as the overall time taken to decode the test items. Though this is evidently a less sensitive measure of decoding speed, it still provides a rough indication of the time needed to decode the 28 pseudo words. Moreover, the traditional measure of decoding speed has the potential problem of overlooking the fact that many participants may correct their pronunciation after the first attempt. The holistic measure of decoding time, on the contrary, did include participants' repetitions (e.g. /mɒ - mɒn - mɒnglʌstəmə/) and self-corrections (e.g. /mɒnglestəmə - mɒnglʌstəmə/).

#### 3.5.2.4 Validity and reliability

Given the research design of the study, it was unfeasible to ask the participants to take the decoding test again so that test-retest reliability could be calculated. However, the intra-rater and inter-rater reliability of the scoring was examined. 10% of the participants' decoding at t1 and 10% of the decoding at t2 were randomly selected and marked again by the researcher and additionally by a Chinese teacher who teaches the phonology class for first-year English majors in University A. The results

found 100% intra-rater agreement and 99.8% inter-rater agreement between the whole-word level decoding scores marked at the two time points. The agreement between the grapheme-level decoding scores was slightly lower: the intra-rater agreement was 94.3% and inter-rater agreement was 90.5%. However, these still represent very high levels of reliability in the scoring.

### 3.5.3 Vocabulary memorisation task

Another important aim of this study is to examine whether the phonics instruction programme led to better vocabulary learning. This section describes the vocabulary memorisation task followed by four immediate recall and recognition tests from the following aspects: section 3.5.3.1 presents the content of the tests; section 3.5.3.2 explains the format and administration of the tests; section 3.5.3.3 discusses how the test results were processed; and section 3.5.3.4 addresses the validity and reliability of the tests.

#### 3.5.3.1 Content

The first issue considered in designing the vocabulary memorisation tests was whether to use real English words or pseudowords as stimuli. Many studies about vocabulary learning use pseudowords as test items (e.g. Hulstijn, 1992), for the simple reason that participants would not possibly know these words, which effectively rules

out prior knowledge as a confounding variable. However, using pseudowords in vocabulary learning tests is open to criticism. For instance, Papagno et al. (1991) argue against this approach, as they discovered that participants performed better at learning real words than pseudo words, as they were less motivated to 'learn nonsense' (p. 342). Given that the participants in this study already had approximately 10 years of English learning experience, it is highly possible that they would realize that pseudoword stimuli were not real English words and thus become demoralised. In addition, asking participants to learn pseudowords can be ethically problematic, as certain degrees of deception may be involved in the learning process (if participants are not informed that the items are not real words). As a result, it was determined that real English words should be used as test items in this study.

The next issue that needs to be addressed is the number of words in the test. As a manual literature search on explicit L2 vocabulary learning only found a few studies involving Chinese university EFL learners as participants, only two studies were consulted. The first one was Hamada & Koda (2008), in which 16 one-syllable pseudo words were presented to Chinese participants studying for an undergraduate or postgraduate degree in a university in the US to learn. The second was Li (2012), in which Chinese participants studying for a Master's or DPhil degree in the University of Oxford were asked to learn 21 two-syllable six-letter real English words. Considering that the participants in the current study were of lower English proficiency compared to those in the two studies who passed TOEFL and IELTS and studied in an English-speaking country, it was determined that a list of 10

two-syllable six-letter English words would constitute the stimuli in the vocabulary memorisation tests. Some English teachers in the three participating universities were also consulted, and agreed that 10 stimuli should be cognitively challenging for the participant without overwhelming them.

All the stimuli in the vocabulary learning tests were chosen from the least frequently used words, as determined by the British National Corpus (Leech & Rayson, 2014). This was to ensure that the participants did not possess any prior knowledge of the test stimuli. Following Woore (2011), words closely resembling valid L1 (in this case, Pinyin) strings were not included, so that automatic triggering of L1 processing mechanisms could be avoided. For instance, a two-syllable six-letter low-frequency word that was considered was 'lichen', but as it is a valid pinyin string and also the name of a famous Chinese soap opera star, it was decoded as /litʃən/ (as in Pinyin) by many participants in a previous study (Li, 2012) and was thus excluded here.

Table 3.3 presents the stimuli in the vocabulary memorisation tests. Similar to the decoding test, two different sets of words were used at t1 and t2 in order to avoid a practice effect.

*Table 3.3 Stimuli in the vocabulary memorisation tests*

| | |
|---|---|
| Form A | argent, doodah, ploidy, cantor, stooge, augean, burlap, tisane, sulcus, rheumy |
| Form B | maenad, orrery, ruddle, precis, ribose, mayhap, zeugma, turgor, callus, zephyr |

All the graphemes in the stimuli were taught in the phonics instruction programme except for the silent <h> in 'rheumy' in Form A and the grapheme combination <yr> in 'zephyr' in Form B. However, given the difficulty in finding replacements which met the criteria as mentioned before, these two words were still included.

The total number of the graphemes was 48 and 50 respectively in Form A and B. The mean bigram frequency of the stimuli in the two forms was also compared in order to examine whether the stimuli in the two test forms were matched in terms of orthographic regularity. Following Solso and Juel (1980), the bigram frequency of each word was calculated, and the results are shown in 4.4. A two-tailed t-test found that the two forms of stimuli did not differ significantly in terms of bigram frequency, with a mean bigram frequency of 2288.8 for Form A ($SD$ = 1693.7), and a mean bigram frequency of 2810.6 for Form B ($SD$ = 1550.6), $t(9) = -.69$, $p = .52$.

*Table 3.4 Bigram frequency of the stimuli in the vocabulary memorisation tests*

| | |
|---|---|
| Form A | argent (3671), doodah (1405), ploidy (2100), cantor (5798), stooge (3941), augean (1583), burlap (1015), tisane (2223), zephyr (410), rheumy (742) |
| Form B | maenad (2764), orrery (3956), ruddle (4999), precis (1887), ribose (2196), mayhap (1948), zeugma (421), turgor (3350), callus (5176), sulcus (1409) |

The two test forms were piloted on 30 participants randomly drawn from the three universities. The participants confirmed that they did not have any prior knowledge of any of the stimuli.

#### 3.5.3.2 Format and administration

Similarly to the decoding test, a split-block counterbalanced design was adopted in the vocabulary memorisation tests. Half the participants in each within each class were randomly assigned to memorise one of the two sets of words as shown above. in other words, half of the participants in each class took Form A at t1 and Form B at t2, and the other half did the opposite.

As vocabulary learning is a multidimensional construct (Laufer, 2004), it is of interest to the current study to examine whether phonics instruction contributes to different aspects of vocabulary memorisation. As a result, the participants were asked to complete the following four tests:

(a) An oral recall test, in which they saw the Chinese translations and said the English words;

(b) A written recall test, where they saw the Chinese translations and spelt the English words;

(c) An aural recognition test, in which they listened to the audio recordings of the English words and wrote down the Chinese translations;

(d) A written recognition test, in which they saw the English words and wrote down the Chinese translations.

The four tests were designed in the hope of providing a relatively comprehensive picture of participants' recall and recognition of the forms of the target words, in which both productive and receptive, written and oral knowledge were tested. Moreover, comparison between the results of the aural recognition test and the written recognition test might provide insight into whether (and with what success) the participants phonologically decode the written stimuli as part of the memorisation process. For instance, if a participant correctly recognises a word in its written form but not its phonological form, this suggests that he/she may not have correctly decoded the word (or decoded it at all) at the point of memorization. Further, comparison between the oral production test and the written production test allows for analysis of whether, and if so how, participants' decoding of the words contributes to their memorisation of the written forms. For instance, if a participant correctly recalls a word's written form but not its phonological form, this might suggest that he/she did

not decode the word during the memorisation process.

The order of the four tests was as listed above. This was to avoid the possible impact of practice effects as far as possible, so that the answer to one test did not appear in a previous test. This requirement was not fulfilled by the aural recognition test, as the answers, which are the Chinese translations, appeared in the previous written recall test. However, given that the difficulty of the aural recognition test lies in matching the Chinese translations to the English pronunciations rather than producing the Chinese translations themselves, the order of the test was deemed to be appropriate. Moreover, any potential practice effect was equal for all the participants in the study.

Another important issue is the duration of the presentation of the stimuli. In previous studies, the stimuli appeared for a very short period of time (2000 ms in Hamada & Koda (2008), 5000 ms in Li (2012)). The original plan was to present each stimulus for 5000 ms. However, this was deemed to be too short for the 30 participants in the pilot study. Almost every participant in the pilot study only recalled zero to one words in the oral recall and written recall tests, and recognised very few words in the oral recognition and written recognition tests. This was understandable, given that these participants were of lower English proficiency compared to those in the other two studies. Given that 5000 ms was too short, three other durations were piloted using different participants, namely 10,000 ms, 20,000 ms and 30,000 ms. The results found that 10,000 ms was still too short for the participants, producing a floor effect on the

two productive recall tests, while 30,000 ms generated many ceiling scores in the two recognition tests and the written recall test. As a result, a duration of 20,000 ms for each stimulus was adopted in this study.

The vocabulary memorisation task was conducted individually in a quiet classroom. Before the task began, the requirements were explained by the researcher in Chinese. The participants were explicitly told about the four tests they were to take after the memorisation process. Hence, this could be classified as an 'intentional' rather than 'incidental' learning task, according to the terms used by Hulstijn (2003). They were also given the opportunity to ask questions.

The stimuli and their Chinese character translations were presented one by one in large, bold, black, Arial, lower case type on a computer screen. In the presentation, the participants were allowed any strategic behaviours they wished, such as writing down the words on a piece of paper, repeating the words and the Chinese translations out loud or silently. This ensures that that memorisation process mimics the real-word learning process as closely as possible. However, this does introduce potentially confounding variables as the participants' strategic behaviours differed.

The recall and recognition tests were conducted immediately after the presentation of the stimuli. This ensures that the participants did not have time to study the words themselves after the presentation, which could possibly constitute a confounding

variable. Participants' pronunciations in the oral recall test were recorded using a digital recorder with built-in microphone. The written recall test, aural recognition test and the written recognition test were conducted on paper. The stimuli in the aural recognition test were recorded by a female native speaker of British English who comes from London and works as a teacher in University A. Each word was repeated three times. The participants were given ample time to finish the tests.

#### 3.5.3.3 Processing of the test results

The four recall and recognition tests were scored in the same way, that is, each correctly recalled/recognised word was given one point, and otherwise zero points. This consistent scoring system allows direct comparisons among the four recall and recognition tests results. However, it should be noted that this scoring system has its drawbacks. This is because participants' vocabulary knowledge is likely to develop incrementally and their learning of new vocabulary is likely not a simple one-off process (Nation, 1990; Schmitt, 2010); yet the binary scoring system cannot reflect participants' partial vocabulary knowledge in the oral recall and the written recall tests. Initially, the plan was to score the oral recall test results and the written recall test results in the same way as scoring the phonological decoding test results; which is, the results would be scored at both the grapheme level and the whole-word level. However, when processing the results of the oral recall test it was found that the participants did not always pronounce the test items in a way such that the phonemes

could be clearly mapped onto the graphemes in the test items (*e.g.* pronouncing *orrery* as /ɔ:'ri:/), which made accurate grapheme-level scoring difficult. As a result, the more nuanced grapheme-level scoring was not adopted for the results of the oral recall and the written recall tests. However, error analysis was conducted for the results of the written recall test, providing more nuanced analysis of the written recall test results.

Previous studies (e.g. Huang & Hanley, 1994) suggest that Chinese EFL learners rely heavily on visual processing in learning new English vocabulary. Hence, it was expected that the participants would achieve better results in the written recall test compared to the oral recall test; and in the written recognition test than the aural recognition test, at least at t1. It is of interest to the current study to examine whether the phonics instruction programme results in participants becoming more 'balanced learners', who not only focus on orthographic forms but also phonological forms in vocabulary learning. As a result, at both time points, the scores on the written recall and the oral recall tests, and the scores on the aural recognition and the written recognition tests were also compared, in order to examine whether the phonics instruction programme narrowed the gap between the two recall tests as well as that between the two recognition tests.

#### 3.5.3.4 Validity and Reliability

Though the vocabulary memorisation task in this study, which involved presenting new words and their translations for a limited period of time, is widely used in research on vocabulary learning, some researchers criticise it as artificial, as no contextual information is provided (Nagy, 1995). This is clearly a valid point. However, Ma (2009) in her investigation of Chinese university EFL learners' vocabulary learning strategies reveals that learning from a word list, which is similar to the vocabulary memorisation task in this study, is regarded as an important way of learning new English words. Moreover, as discussed in section 2.1, making the form-meaning link is an essential first step of learning new vocabulary, and this was exactly what the vocabulary memorisation tasks intended to explore.

Test-retest reliability was not calculated for the vocabulary tests, as asking the participants to take the vocabulary memorisation tests again would cause practice effects. However, the intra-rater and inter-rater reliability of the marking of the oral recall test was examined. 10% of the participants' responses at t1 and 10% of the responses at t2 were randomly selected and marked again both by the researcher and by a Chinese teacher who teaches the phonology class for first-year English majors in University A. 100% agreement between the recall scores marked at the two time points was observed. As the written recall, aural recognition and written recognition tests were pen-and-paper tests, which made it impossible for the other rater to mark the tests as she was based in China, only intra-rater reliability was checked. 10% of the participants' tests results at t1 and 10% at t2 were randomly selected and marked

again by the researcher. 100% agreement was found for all the three tests.

### 3.6 Ethics

The current study gained ethical approval from the Central University Research Ethics Committee (CUREC) in the University of Oxford. Before data collection, all participants were provided with a consent form explaining the requirements of the tests they were to take. They were also clearly informed that they could stop participating in the tests at any time without incurring any penalty, simply by telling the researcher. However, it is worth noting that the participants were only given the chance to stop participating in the data collection, but not the opportunity to stop taking part in the instruction programmes, as the programmes themselves were part of their curriculum, as determined by their teachers.

All the audio data collected in this study was stored on a password-protected laptop and backed up in an encrypted hard disk. All the hard copies of the tests were stored in a locked room.

At the request of the three participating universities, all the materials for the phonics instruction programme were provided to the participants in the comparison group after the data collection at t2 was completed.

## Chapter 4. Findings I: Phonological Decoding Test - Overall Results

This chapter provides an initial evaluation of the programme of phonics instruction by examining the results of the English phonological decoding test. It compares the phonological decoding test scores of intervention and comparison groups before and after the phonics instruction, which directly sheds light on the effectiveness of the instruction programme. Therefore, this chapter addresses Research Question 1:

*RQ 1: Does a programme of systematic phonics instruction lead to improvements in Chinese university EFL learners' English decoding?*

This chapter is divided into five sections. First, as previously mentioned, learners' English proficiency and vocabulary breadth might potentially influence their ability to decode English words, so Section 4.1 compares the intervention and the comparison group in terms of National College Entrance English Exam (NCEEE) scores and British Picture Vocabulary Scale (BPVS) scores, so as to examine whether intervention and comparison groups were matched with regard to English proficiency and English vocabulary knowledge. Second, this chapter compares the phonological decoding test results of participants in the intervention group, who followed the programme of systematic English phonics instruction, with those in the comparison group, who followed the programme of systematic English pronunciation instruction (but without explicit phonics instruction). Following the scoring system discussed in

section 3.5.2.3, participants' pronunciations were judged at two levels: at the whole word level, in which the numbers of correctly decoded words were counted; and at the individual grapheme level, in which the numbers of correctly decoded graphemes were counted. As the participants were from three different levels of universities, the decoding test results were analysed both overall and at the level of the individual participating universities. In addition, participants' speed of decoding, which serves as an indicator of decoding proficiency (e.g. Hamada & Koda, 2008), was also measured. As discussed in section 3.5.2.3, the time between the onset of the first stimulus and the end of the last stimulus was recorded as a rough measure of the overall time of decoding.

Therefore, three sets of results are presented. Firstly, section 4.2 compares the number of words decoded correctly by the two groups. Secondly, section 4.3 compares the groups in terms of the number of correctly decoded graphemes. Finally, the overall time of decoding of the two groups are compared in section 4.4. Section 4.5 summarises the results.

## 4.1 NCEEE and BPVS scores

### 4.1.1 NCEEE scores

The NCEEE scores of all intervention and comparison groups are summarised in

Table 5.1, and the histograms are presented in Figure 4.1. The NCEEE scores of the intervention and the comparison group were compared using a one-way ANOVA. The assumption of Normality was met as the *z*-scores of skewness and kurtosis were smaller than 1.96[6] (Field, 2005: 139). The assumption of homogeneity of variance was also met based on Levene's test[7]. An ANOVA found that there was no significant difference between the two groups, $F(1, 178) = 1.62$, $p = .21$, indicating that intervention and comparison groups were matched in terms of their NCEEE scores. Hence English proficiency was ruled out as a confounding variable.

*Table 4.1 NCEEE scores (out of a possible 150) for all participants*

| | **Median** | **Mean** | **S.D.** | **Minimum** | **Max** |
|---|---|---|---|---|---|
| **Intervention (N=94)** | 123 | 122.3 | 9.0 | 97 | 146 |
| **Comparison (N=86)** | 120 | 120.6 | 9.8 | 94 | 140 |

[6] *z*-score of skewness = 1.23, and *z*-score of kurtosis = 1.36
[7] $F(1, 178) = .41$, $p = .49$

*Figure 4.1. Histograms of NCEEE scores for all participants*

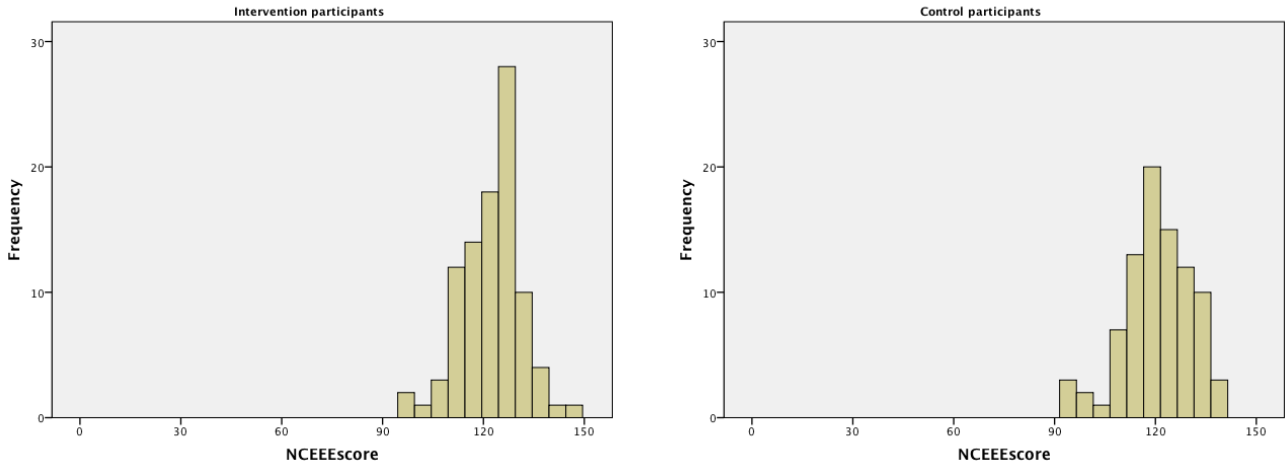



Table 4.2 presents the results of NCEEE scores by university. The assumptions of Normality[8] and homogeneity of variance[9] were met for the data of all three universities. One-way ANOVA tests revealed no significant differences between the intervention group and the comparison group in any of the universities (University A: $F(1, 43) = .73$, $p = .79$, University B: $F(1, 91) = 2.10$, $p = .15$; University C: $F(1, 40) = 1.39$, $p = .25$). This suggests that intervention and comparison groups of each university were matched in terms of their English proficiency.

*Table 4.2 NCEEE scores (out of a possible 150) grouped by university*

| | | Median | Mean | S.D. | Minimum | Max |
|---|---|---|---|---|---|---|
| **University A** | **Intervention (N=24)** | 130 | 130.4 | 6.1 | 120 | 146 |
| | **Comparison (N=21)** | 129 | 129.9 | 5.5 | 119 | 140 |
| **University B** | **Intervention (N=47)** | 123 | 122.5 | 6.4 | 110 | 141 |
| | **Comparison (N=46)** | 119 | 120.5 | 6.7 | 110 | 135 |
| **University C** | **Intervention (N=23)** | 113 | 113.7 | 8.4 | 97 | 126 |
| | **Comparison (N=19)** | 110 | 110.4 | 9.9 | 94 | 126 |

[8] University A: *z*-score of skewness = 1.14, *z*-score of kurtosis = 0.04; University B: *z*-score of skewness = 1.81, *z*-score of kurtosis = -.38; University C: *z*-score of skewness = -1.12, *z*-score of kurtosis = -.1.06

[9] University A: $F(1, 43) = .19$, $p = .67$; University B: $F(1, 91) = .00$, $p = .99$; University C: $F(1, 40) = .79$, $p = .38$

The NCEEE scores of the three participating universities were also compared using a one-way ANOVA. There was a statistically significant difference between groups as determined by the ANOVA, $F(2, 177) = 62.96$, $p < .001$. As the sample sizes of the three universities were different, a Gabriel's post-hoc test was conducted to further examine which universities differed (Field, 2005: 374). It was found that the mean NCEEE score of University A was significantly different from that of University B ($p < .001$) and University C ($p < .001$), and the mean scores of University B and University C were also significantly different ($p < .001$). As the descriptive statistics reveal that the mean NCEEE score of University A was higher than that of University B, which in turn was higher than that of University C, it can be concluded that University A scored significantly higher than University B and University C, and University B significantly outperformed University C in terms of NCEEE scores. These results were in line with that would have been expected, as the three universities were from different tiers.

### 4.1.2 BPVS scores

The results of the BPVS test for all participants are presented in Table 4.3. The histograms are shown in Figure 4.2. The original plan was to analyse the results using a one-way ANOVA test. However, before conducting the test, visual inspection of histograms suggested that the BPVS scores of all participants were not normally distributed, and this was confirmed by statistical analysis[10]. The assumption of

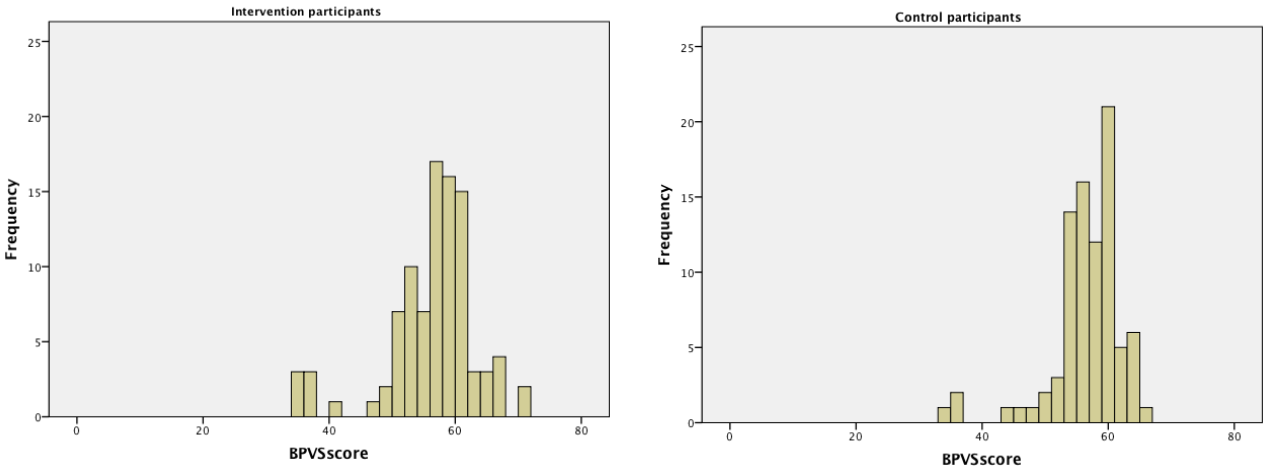
[10] $z$-score of skewness= -7.86, and the $z$-score of kurtosis= 8.80

homogeneity of variance was confirmed, as revealed by a Levene's test[11]. Log transformation was conducted in order to correct this, but the data still showed some degrees of skewness and kurtosis. Hence, the data was primarily analysed using non-parametric tests. However, parametric tests were also conducted as a means of comparison.

*Table 4.3 BPVS scores (out of a possible 70) for all participants*

| | **Median** | **Mean** | **S.D.** | **Minimum** | **Max** |
|---|---|---|---|---|---|
| **Intervention (N=94)** | 57 | 55.8 | 7.2 | 35 | 70 |
| **Comparison (N=86)** | 57 | 56.0 | 5.7 | 34 | 65 |

*Figure 4.2 Histograms of BPVS scores for all participants*




The non-parametric technique adopted was the Mann-Whitney test. The null hypothesis that the distribution of BPVS scores is the same for both groups was retained ($U = 4443.00$, $Z = -.03$, $p = .97$). The parametric test used was a one-way ANOVA, which also found that the difference between intervention and comparison groups was not significant, $F(1, 178) = .05$; $p = .82$. Hence, although the parametric

[11] $F(1, 178) = 2.46$, $p = .12$

tests must be considered with caution, both non-parametric and parametric tests showed no significant difference between the BPVS scores of intervention and comparison groups, indicating that the two groups were matched in terms of their receptive vocabulary knowledge.

The intervention and the comparison group of each university were also compared. Table 4.4 summarises the BPVS scores grouped by university.

*Table 4.4 BPVS scores (out of a possible 70) grouped by university*

| | | **Median** | **Mean** | **S.D.** | **Minimum** | **Max** |
|---|---|---|---|---|---|---|
| **University A** | **Intervention (N=24)** | 56.5 | 56.8 | 9.1 | 35 | 70 |
| | **Comparison (N=21)** | 54.0 | 55.1 | 7.1 | 36 | 65 |
| **University B** | **Intervention (N=47)** | 58.0 | 56.1 | 6.9 | 35 | 66 |
| | **Comparison (N=46)** | 58.0 | 57.2 | 4.6 | 34 | 64 |
| **University C** | **Intervention (N=23)** | 55.0 | 54.1 | 5.2 | 41 | 61 |
| | **Comparison (N=19)** | 55.0 | 54.2 | 6.1 | 35 | 61 |

The assumption of Normality[12] was not met for the BPVS scores of any of the three universities. The assumption of homogeneity of variance was retained for all three universities[13]. As a result, both non-parametric and parametric tests were conducted to examine whether intervention and comparison groups in each university differed in terms of BPVS scores. Note that the parametric tests must be considered with caution in all cases.

For University A, the Mann-Whitney test revealed no significant difference between

[12] University A: *z*-score of skewness = 2.10, *z*-score of kurtosis = 1.53; University B: *z*-score of skewness = -9.42, *z*-score of kurtosis = 13.59; University C: *z*-score of skewness = -3.77, *z*-score of kurtosis = 3.57

[13] University A: $F(1, 43) = .62$, $p = .44$; University B: $F(1, 91) = 2.31$, $p = .13$; University C: $F(1, 40) = .50$, $p = .82$

the two groups, $U = 216.00$, $Z = -.82$, $p = .41$. This was supported by the one-way ANOVA, $F(1, 43) = .48$, $p = .48$.

For University B, the Mann-Whitney test also revealed no significant difference between the two groups, $U = 1282.00$, $Z = -.121$, $p = .90$. This was confirmed by the one-way ANOVA, $F(1, 91) = .79$, $p = .38$.

For University C, the Mann-Whitney test again revealed no significant difference between the two groups, $U = 207.50$, $Z = -.28$, $p = .78$. This was also supported by the one-way ANOVA, $F(1, 40) = .00$, $p = .96$.

Therefore, it can be concluded that the intervention group and the comparison group in each university were matched in terms of their receptive vocabulary knowledge, as demonstrated by both non-parametric and parametric tests.

The BPVS scores of the three universities were also compared using both parametric and non-parametric techniques. A one-way ANOVA revealed no significant differences between the three universities, $F(2, 177) = 2.17$, $p = .12$. However, a Gabriel's post-hoc test revealed that the difference between University B and University C approached significance ($p = .05$). Non-parametric tests were also conducted. A Kruskal-Wallis test found that the differences between the three universities were significant, $H(2) = 9.17$, $p < .05$. Therefore, three Mann-Whitney

tests were used to follow up this finding. A Bonferroni correction was applied; hence all the results were reported at a .0167 level of significance. It was found that the BPVS score of University A did not differ significantly from that of University B ($U$ = 1842.50, $Z$ = -1.14, $p$ = .26) or University C ($U$ = 798, $Z$ = -1.26, $p$ = .21), but University B and University C differed significantly ($U$ = 1291, $Z$ = -3.16, $p$ < .001).

In sum, the intervention and comparison groups did not demonstrate any significant difference in terms of their NCEEE and BPVS scores. Moreover, though there were differences between the universities in terms of participants' NCEEE and BPVS scores, the comparison between the intervention group and the comparison group in each university also revealed no significant difference in the two tests. This suggests that participants' English proficiency and receptive vocabulary knowledge can be excluded as confounding variables in statistical analysis.

### 4.2 Word-level phonological decoding scores

The descriptive statistics for the word-level decoding scores of all intervention and comparison groups at Time 1 and Time 2 are presented in Table 4.5. The histograms are shown in Figure 4.3. The word-level phonological decoding scores were analysed using a two-way mixed factorial ANOVA test with one within subjects variable (time) and one between-subjects variable (condition). The assumptions of the test were checked. Firstly, visual inspection of the histograms and statistical tests suggested that

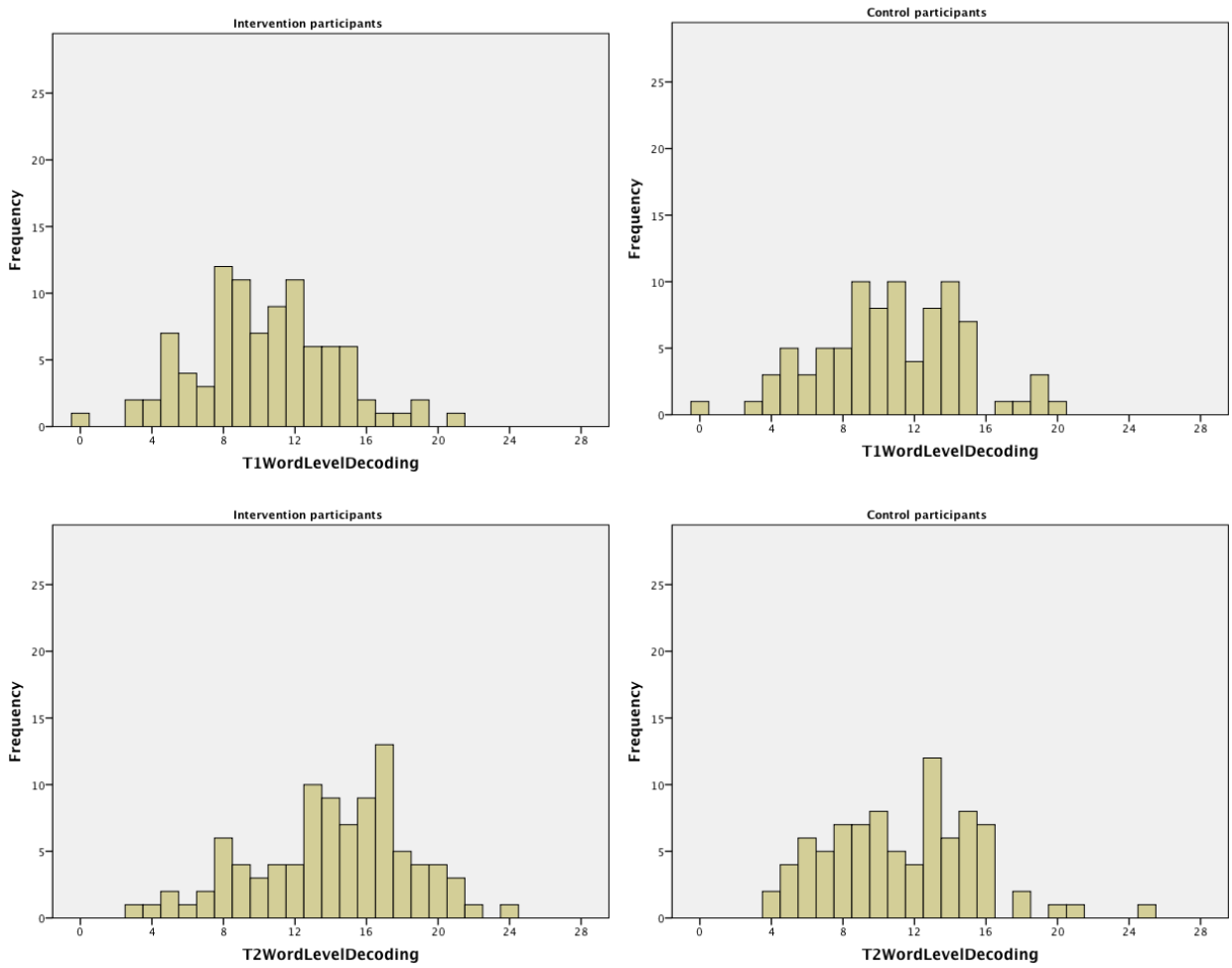
the assumption of Normality was met for both t1 and t2[14]. Secondly, the issue of sphericity was not relevant here as the repeated measures variable had only two levels (t1 and t2) (Field, 2005: 459). Finally, Levene's test confirmed the homogeneity of variances for both levels of the repeated measures variable[15]. The effect size of Pearson's *r* was calculated following Field (2005), where an effect size of .1 is interpreted as a small effect, .3 as a medium effect and .5 as a large effect.

*Table 4.5 Word-level phonological decoding scores for all participants (out of 28)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 10 | 10.3 | 3.9 | 0 | 21 | 14.5 | 14.1 | 4.4 | 3 | 24 |
| **Comparison (N=86)** | 11 | 10.7 | 4.0 | 0 | 20 | 11 | 11.3 | 4.7 | 0 | 25 |

[14] t1 *z*-score of skewness = -.50, *z*-score of kurtosis = .14; t2 *z*-score of skewness = 1.43, *z*-score of kurtosis = .34
[15] At t1, $F(1, 178) = .05$, $p = .83$; at t2, $F(1, 178) = .82$ , $p = .38$

*Figure 4.3 Histograms of word-level phonological decoding scores for all participants (by group) at time 1 and time 2*



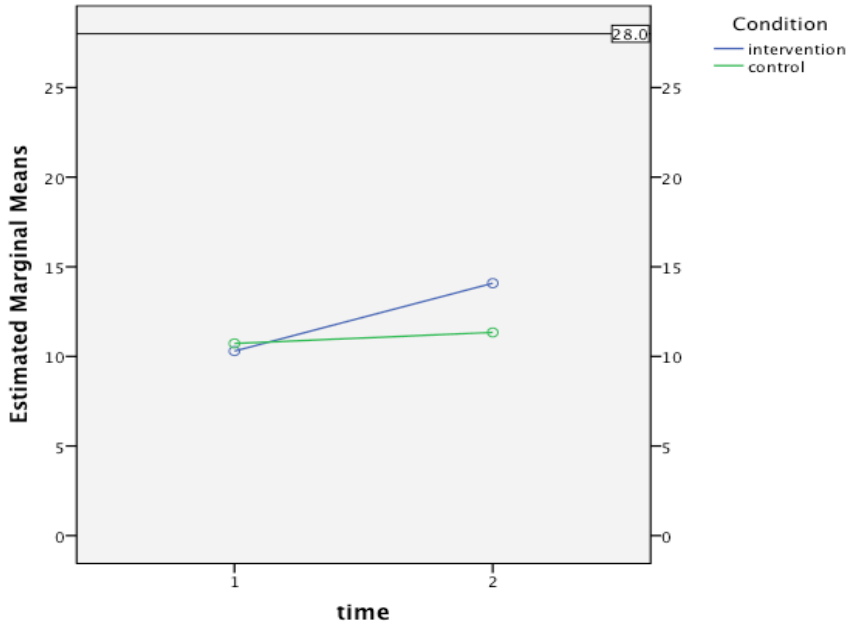

The ANOVA yielded a significant main effect of time with a large effect size, $F(1, 178) = 68.95$, $p < .001$, $r = .53$[16], and also a significant main effect of condition with a small effect size, $F(1, 178) = 4.46$, $p < .005$, $r = .16$, indicating that the word-level decoding scores were significantly higher at t2 than at t1, and for intervention participants than comparison groups.

The interaction between time and condition was also significant, $F(1, 178) = 35.76$, $p$

[16] The effect sizes were measured by Pearson's correlation coefficient throughout different analyses in this chapter and Chapters 5-7 so that the effect sizes of different tests can be comparable to each other (Field, 2005: 448).

< .001, with a medium effect size ($r$ = .41). The interaction graph (Figure 4.4) and descriptive statistics (Table 4.5) were consulted in order to better interpret the ANOVA results. It can be seen that at t1, intervention and comparison groups' scores were roughly the same, while at t2, intervention participants' scores showed an increase, while the comparison groups' scores remained roughly the same. Intervention participants correctly decoded on average three more words out of 28 than comparison groups at t2.

*Figure 4.4. Estimated marginal means of word-level decoding scores for all participants (out of a possible 28)*




The results above demonstrate that participants who received systematic English phonics instruction made significantly more progress in English decoding compared to those in the comparison group in terms of the number of whole words decoded accurately. This provides an affirmative answer to Research Question 1 (*Does a programme of systematic phonics instruction lead to improvements in Chinese*

*university EFL learners' English decoding?*).

The word-level decoding results were also analysed by university. Table 4.6 summarises the descriptive statistics of the three universities. The assumption of Normality was met for the data of all three universities[17]. The assumption of homogeneity of variance was also met for all three universities as revealed by Levene's test[18].

*Table 4.6. Word-level phonological decoding scores by university (out of 28)*

| | | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Uni A** | **Intervention (N=24)** | 10 | 10.1 | 4.5 | 4 | 21 | 14 | 14.8 | 4.2 | 8 | 24 |
| | **Comparison (N=21)** | 9 | 9.6 | 4.1 | 3 | 19 | 9 | 9.4 | 3.5 | 4 | 16 |
| **Uni B** | **Intervention (N=47)** | 11 | 10.9 | 3.7 | 0 | 19 | 16 | 15.2 | 3.6 | 6 | 21 |
| | **Comparison (N=46)** | 11 | 11.2 | 4.1 | 0 | 20 | 13 | 12.5 | 5.6 | 5 | 25 |
| **Uni C** | **Intervention (N=23)** | 9 | 9.3 | 3.4 | 3 | 15 | 11 | 11.1 | 4.7 | 3 | 19 |
| | **Comparison (N=19)** | 11 | 10.8 | 3.7 | 4 | 17 | 11 | 10.7 | 3.3 | 4 | 15 |

Three two-way mixed factorial ANOVA tests with one within-subjects variable (time) and one between-subjects variable (condition) were conducted to analyse the data.

[17] University A: t1 *z*-score of skewness = 1.62, *z*-score of kurtosis = .10, t2 *z*-score of skewness = .99, *z*-score of kurtosis = -.02; University B: t1 *z*-score of skewness = -1.35, *z*-score of kurtosis = 1.30, t2 *z*-score of skewness = -1.93, *z*-score of kurtosis = 0.75; University C: t1 *z*-score of skewness = -0.05, *z*-score of kurtosis = -1.25, t2 *z*-score of skewness = .12, *z*-score of kurtosis = -.94

[18] University A: at t1 $F$ (1, 43) = .04, $p$ = .84, at t2 $F$ (1, 43) = .14, $p$ = .71; University B: at t1 $F$ (1, 94) = .01, $p$ = .94, at t2 $F$ (1, 94) = .02, $p$ = .89; University C: at t1 $F$ (1, 40) = .25, $p$ = .62; at t2 $F$ (1, 40) = 3.75, $p$ = .06). No outlier was detected in all three sets of data

For University A, there was a significant main effect of time, $F(1, 43) = 24.71$, $p < .001$, with a large effect size ($r = .60$), and a significant main effect of condition, $F(1, 43) = 6.35$, $p < .05$, with a medium effect size ($r = .36$). The interaction between time and condition was also significant, $F(1, 43) = 29.09$, $p < .001$, with a large effect size ($r = .64$).

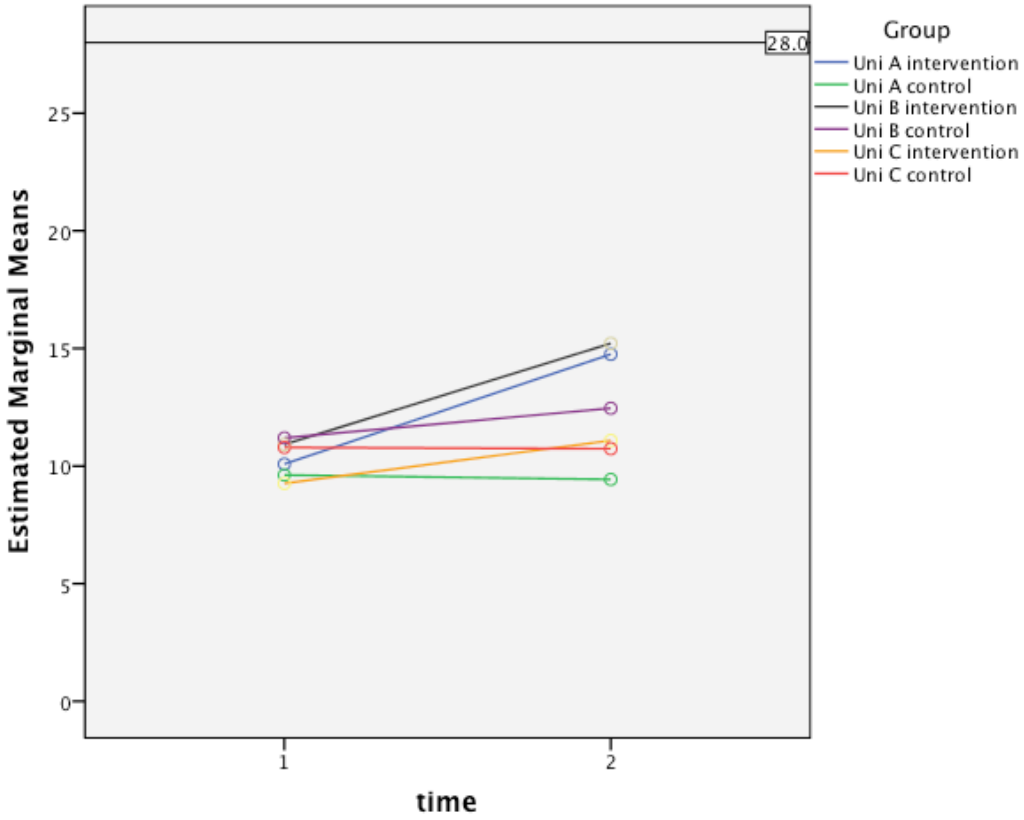
For University B, a significant main effect of time was found, $F(1, 91) = 48.83$, $p < .001$, with a large effect size ($r = .59$). However, there was no significant main effect of condition, $F(1, 91) = 2.99$, $p = .09$. The interaction between time and condition was significant, $F(1, 91) = 14.57$, $p < .001$, with a medium effect size ($r = .37$).

For University C, there was no significant main effect either of time, $F(1, 40) = 3.65$, $p = .06$, or condition, $F(1, 40) = .29$, $p = .59$. The interaction between time and condition was also non-significant, $F(1, 40) = 4.10$, $p = .05$.

The interaction graph of the six groups in the three universities is presented in Figure 4.5. It can be seen that the intervention participants in University A and University B showed good progress and a clear advantage over their counterparts in the comparison groups at t2. In contrast, the intervention participants in University C not only made smaller progress (1.8 words) compared with the intervention groups in University A (4.7 words) and University B (4.3 words) at t2, but also achieved similar scores to the

comparison group in University C at both t1 and t2.

*Figure 4.5. Estimated marginal means of word-level decoding scores of all six groups of participants in the three universities (out of a possible 28)*




One factor that differentiated University C from the other two universities is that the participants in University C were learning second foreign languages when the intervention was occurring (intervention participants were learning French, while comparison groups were learning Japanese). As a result, the effects of the instruction programme may have been compromised by participants' developing knowledge of French GPCs (this will be discussed in Appendix 8, which interested readers may refer to). Given this difference between University C and the other two universities – namely the possible 'interference' of learning another Roman alphabetic writing

system concurrently with the English phonics programme – it is of interest to conduct a further analysis of overall decoding scores at each time point excluding university C: that is, aggregating the scores from universities A and B.

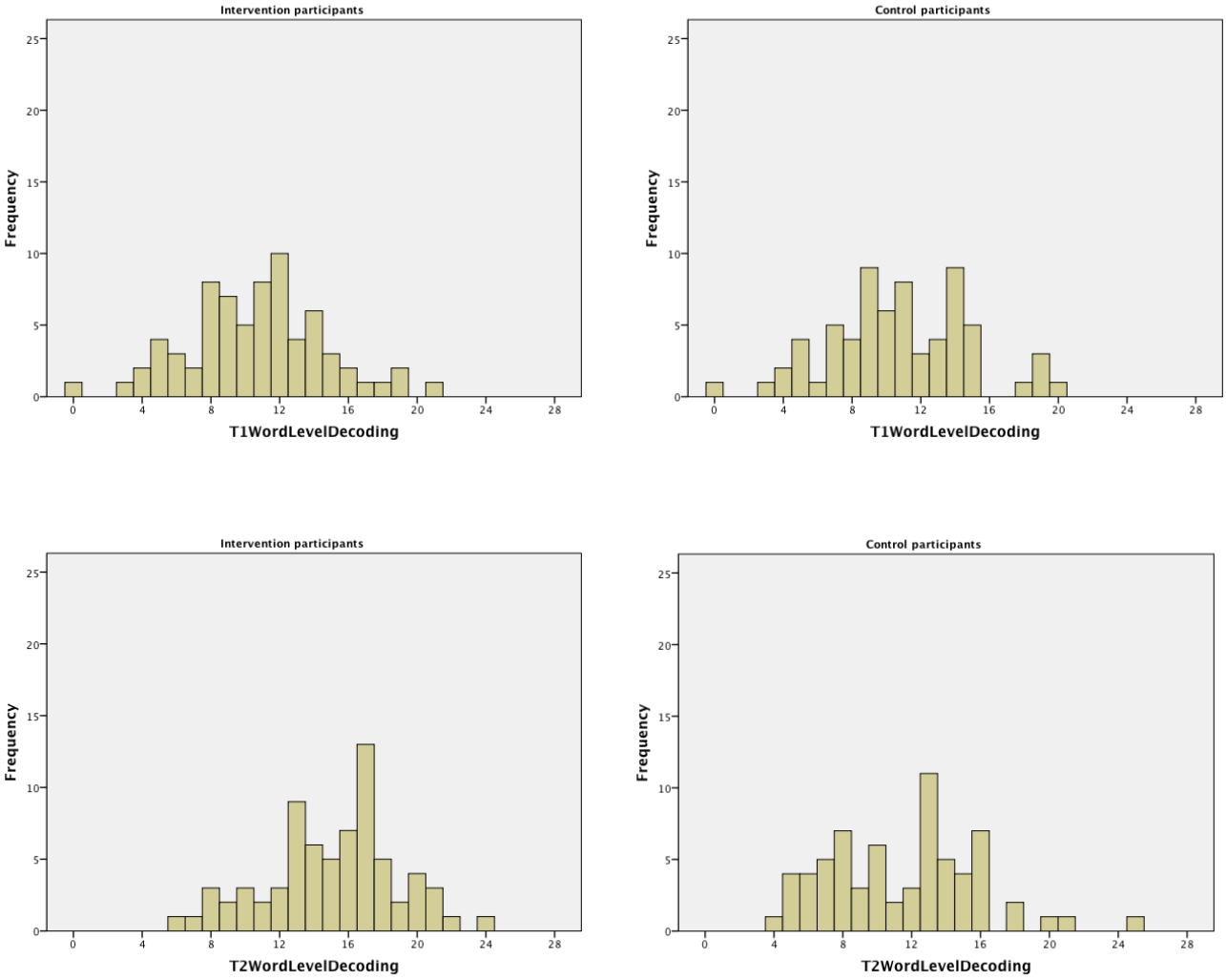
The descriptive statistics for the combined word-level decoding scores of participants in Universities A and B are presented in Table 4.7. The histograms are shown in Figure 4.6. The data was again analysed using a two-way mixed factorial ANOVA with one within-subjects variable (time) and one between-subjects variable (condition). The assumptions of Normality[19] and homogeneity of variance[20] were both confirmed.

*Table 4.7 Word-level phonological decoding scores of participants in Universities A and B combined (out of a possible 28)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 11.0 | 10.6 | 4.0 | 0 | 21 | 16.0 | 15.1 | 3.8 | 6 | 24 |
| **Comparison (N=67)** | 11.0 | 10.0 | 4.1 | 0 | 20 | 12.0 | 11.5 | 4.3 | 4 | 25 |

[19] t1 *z*-score of skewness = .26, *z*-score of kurtosis = .26; t2 *z*-score of skewness = -.08, *z*-score of kurtosis = -1.14
[20] At t1 $F(1, 136) = .06$, $p = .81$; at t2 $F(1, 136) = 1.91$, $p = .17$.

*Figure 4.6 Histograms of word-level phonological decoding scores of participants in Universities A and B (out of a possible 28)*

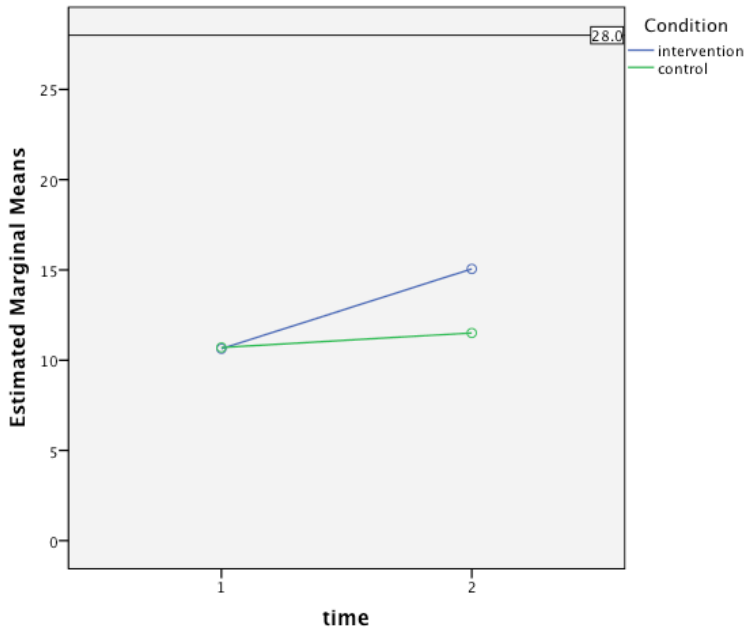



The ANOVA revealed a significant main effect of time with a large effect size, $F(1, 136) = 72.74$, $p < .001$, $r = .59$, as well as a significant main effect of condition with a small effect size, $F(1, 136) = 7.89$, $p < .01$, $r = .23$.

The interaction between time and condition was also significant, $F(1, 136) = 34.80$, $p < .001$, with a medium effect size ($r = .45$). The interaction graph is presented in

Figure 4.7. It can be seen that the intervention group made more progress compared to the comparison groups at t2. However, the exclusion of University C did not have any serious impact on the results: on average, the intervention group still correctly decoded nearly 4 more words than the comparison group at t2 (12% of the total number of words), though the comparison group seemed to make more progress (1.5 words) compared to when University C was included (0.6 words).

*Figure 4.7 Estimated marginal means of word-level phonological decoding scores of participants in Universities A and B (out of a possible 28)*




In summary, the intervention participants in the three universities made significantly more progress in terms of the number of whole words decoded correctly, providing an affirmative answer to Research Question 1. The analysis of the decoding scores of each university reveals that only in University C did intervention participants not make significantly more progress than their comparison counterparts at t2, which (it

could be hypothesized) may have been due to the fact that they were concurrently learning French as well as English at the time of the phonics instruction programme. These languages use the same Roman alphabetic script, but differ in the GPCs that were key to the intervention.

On average, the intervention participants correctly decoded nearly four more words compared to the comparison group by the end of the instruction programme, regardless of whether University C was included or not. By contrast, participants in the comparison condition clearly made less progress at t2, supporting the hypothesis that the phonology instruction programme was less effective than the phonics instruction programme in promoting word-level phonological decoding performance.

### 4.3 Grapheme-level phonological decoding scores

As discussed in section 3.5.2.3, the phonological decoding test was also scored at the grapheme level. The main reason for this is that the grapheme-level score is a more sensitive measure of participants' decoding performance. As a result, this section will explore whether more progress was made in decoding when judged at the grapheme level than at the whole-word level. As was the case in the analyses of the word-level decoding scores, three statistical analyses were performed. Firstly, the grapheme-level scores of the intervention participants in all the three universities were compared to those of the comparison groups, in order to find out whether the participants had made

additional progress beyond that which was detectable at the word level. Secondly, the grapheme-level decoding scores of the intervention group and the comparison group in University C were compared, in order to examine whether the phonics instruction led to significantly greater grapheme-level progress for the intervention participants than their comparison counterparts, given that word-level differences were not detected. Thirdly, the aggregated grapheme-level decoding scores of the intervention and comparison groups in Universities A and B were compared, given the evidence presented above that the effects of the intervention in University C may have been affected by participants' concurrent learning of French.

The grapheme-level decoding scores were converted into an accuracy percentage (i.e. number of correctly decoded grapheme tokens divided by the total number of grapheme tokens) in order to provide a clear picture of participants' decoding performance at each time point. The main reason for this is that the total number of graphemes in the two test forms (randomly assigned to be used at t1 and t2 respectively) are not exactly the same (119 and 121 respectively); as a result, converting the raw scores into an accuracy percentage could afford a more accurate comparison. The assumptions of a two-way mixed factorial ANOVA with one between-subjects variable (condition) and one within-subjects variable (time) were firstly checked. Mauchly's test of sphericity was not considered here, as the repeated measures had only two levels. The assumption of Normality was not met[21]. The

[21] For the data of all three universities combined, t1 *z*-score of skewness = -11.2, *z*-score of kurtosis = 21.65; t2 *z*-score of skewness = -5.18, *z*-score of kurtosis = 2.82. For the data of Universities A and B, t1 *z*-score of

assumption of homogeneity of variance was violated both (a) for the data of all three universities combined and (b) for the data of Universities A and B combined, but confirmed for University C[22]. Hence, the data was primarily analysed using non-parametric tests. However, ANOVA tests were also conducted as a point of comparison.

The descriptive statistics for the grapheme-level decoding scores of all three universities are presented in Table 4.8. The histograms are shown in Figure 4.8.

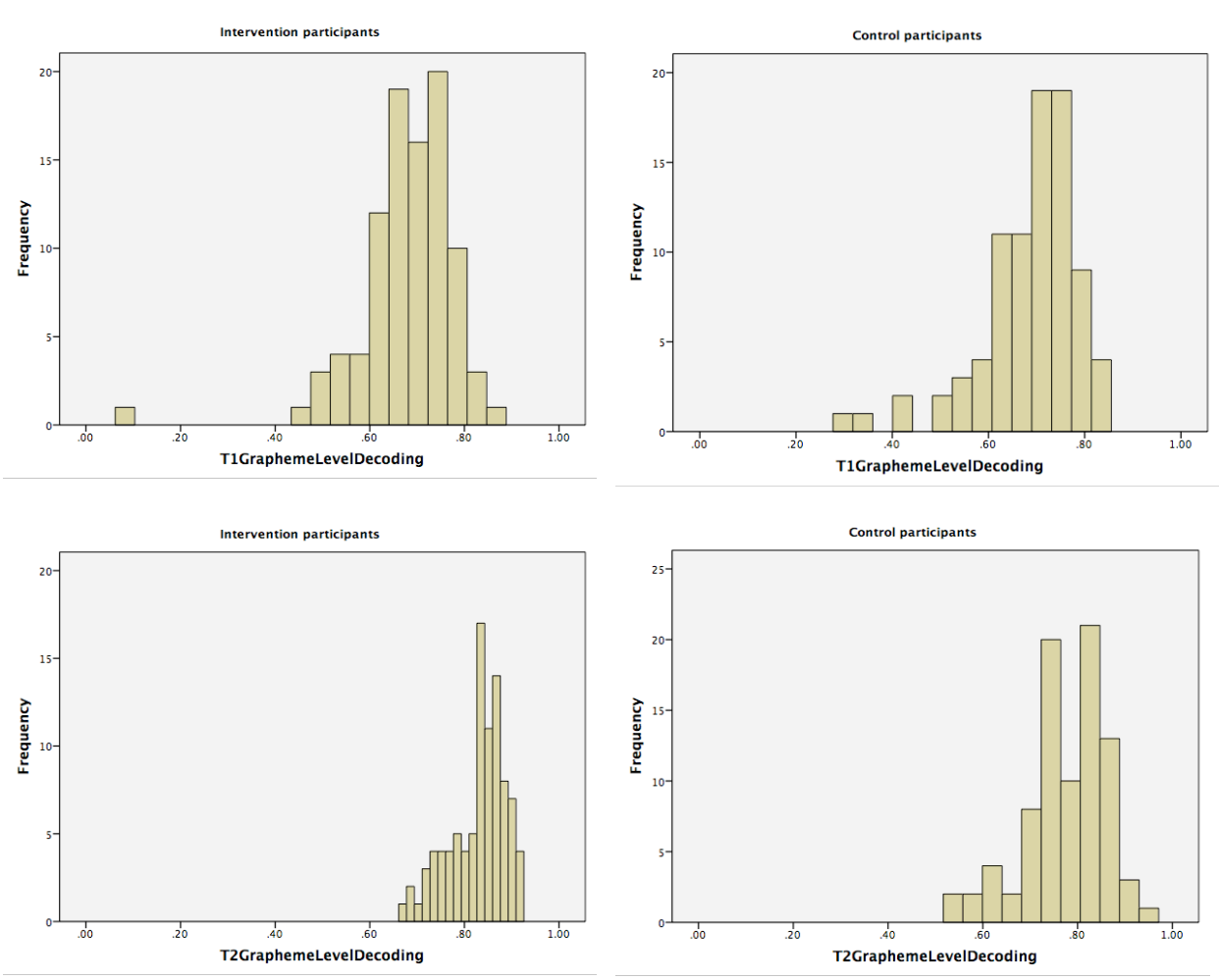
*Table 4.8. Percentage decoding accuracy in participants in all three universities, as measured at the grapheme level*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 69.42% | 68.03% | 9.98% | 8% | 86% | 84.3% | 83.2% | 5.87% | 67% | 93% |
| **Comparison (N=86)** | 71.07% | 68.55% | 10.25% | 30% | 83% | 78.51% | 77.49% | 8.48% | 54% | 96% |

skewness = -4.09, *z*-score of kurtosis = 5.31; t2 *z*-score of skewness = -2.37, *z*-score of kurtosis = .86; for University C, t1 *z*-score of skewness = -4.06, *z*-score of kurtosis = 2.94; t2 *z*-score of skewness = -1.59, *z*-score of kurtosis = .14.

[22] For the data of all three universities combined, t1 $F(1, 178)= .51$, $p = .47$; t1 $F(1, 178)= 8.87$, $p < .001$. For the data of Universities A and B, t1 $F(1, 136) = .03$, $p = .86$; t2 $F(1, 136) = 17.48$, $p < .001$. For University C, t1 $F(1, 40) = 4.08$, $p = .05$; t2 $F(1, 40) = .37$, $p = .55$.

*Figure 4.8 Histograms of percentage decoding accuracy in participants in all three universities, as measured at the grapheme level*




It can be seen that both intervention and comparison groups correctly decoded approximately 70% of the total graphemes at t1. At t2, both intervention and comparison groups made some progress in terms of grapheme-level decoding accuracy, but progress appeared more pronounced for the intervention participants than the comparison groups.

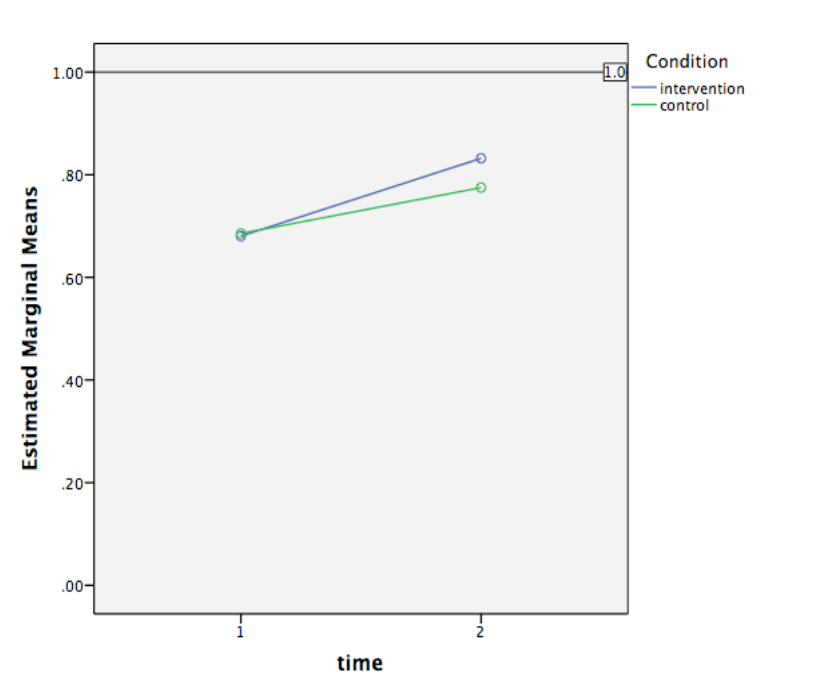
Non-parametric tests were conducted to compare the grapheme-level decoding performance of the intervention and comparison groups (based on the combined data

from all three universities). Firstly, the differences between each group's scores at t1 and t2 were compared using a Wilcoxon Signed Rank test. For both groups, scores at t1 differed significantly from those at t2 (intervention group: $Z = -8.42$, $p < .001$, $r = -.61$; comparison group: $Z = -7.27$, $p < .001$, $r = -.56$). As the mean grapheme-level decoding scores were higher at t2 than t1 for both groups, this confirms significant progress in English decoding by both groups when measured at the grapheme level. Then, a Mann-Whitney test was conducted to examine the differences between the scores of the two groups at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U = 3768$, $Z = -.78$, $p = .43$), but were significantly different at t2, with a small effect size ($U = 2302$, $Z = -4.99$, $p < .001$, $r = .10$).

An ANOVA was also conducted as a point of comparison. There was a significant main effect of time with a large effect size, $F(1, 178) = 339.99$, $p < .001$, $r = .81$, as well as a significant main effect of condition with a medium effect size, $F(1, 178) = 5.21$, $p < .05$, $r = .47$. The interaction between time and condition was also significant, $F(1, 178) = 22.62$, $p < .001$, with a medium effect size ($r = .33$). The interaction graph is shown in Figure 4.9. Although they are to be taken with caution, these results support the findings of the non-parametric tests that the phonics instruction led to significantly greater progress in terms of grapheme-level decoding performance for the intervention participants than the comparison groups in all three universities. From a different perspective, this also offers an affirmative answer to Research

Question 1.

*Figure 4.9 Estimated marginal means of percentage decoding accuracy in participants in all three universities, as measured at the grapheme-level*

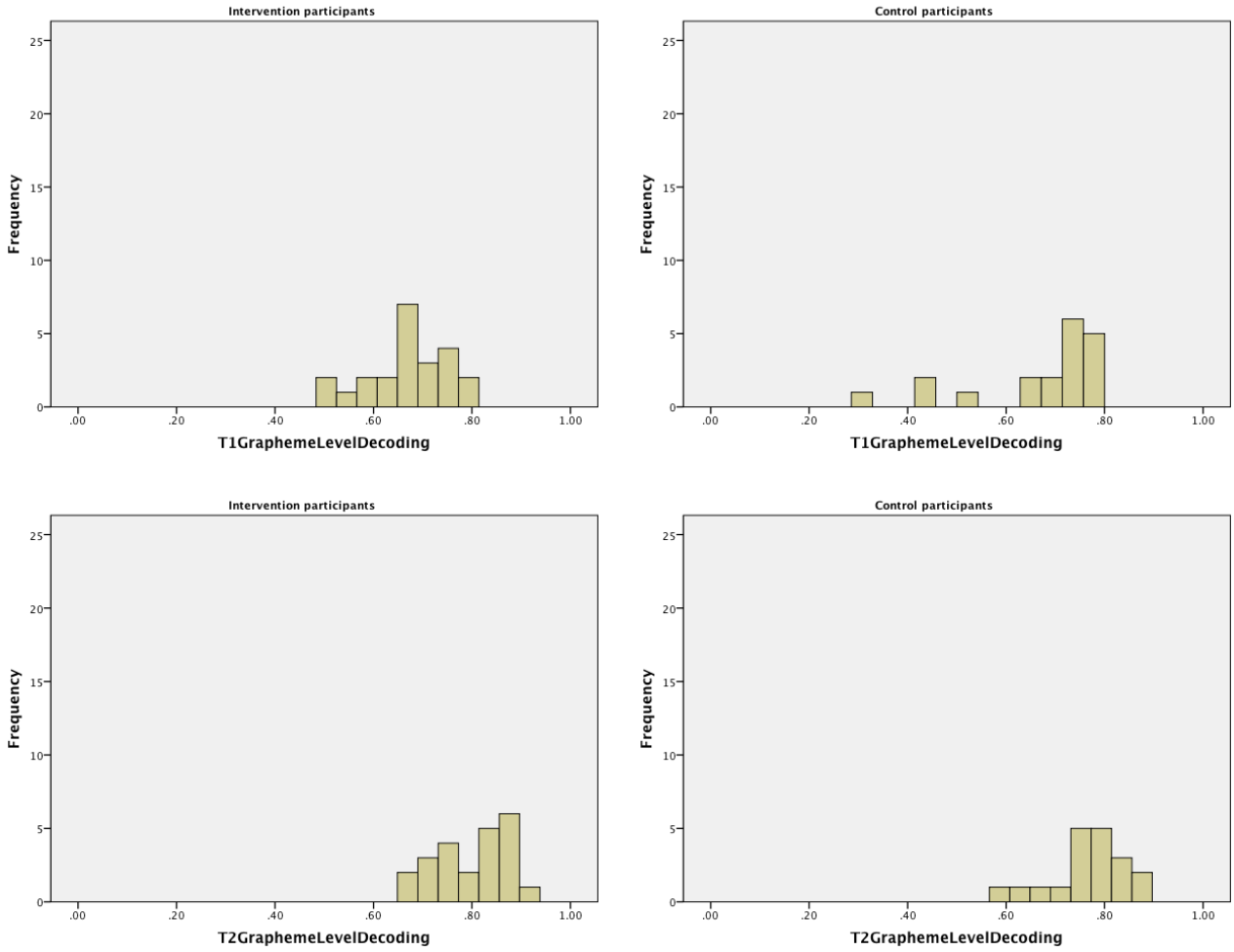



The grapheme-level decoding scores of University C were then analysed. The descriptive statistics are presented in Table 4.9. The histograms are shown in Figure 4.10.

*Table 4.9 Percentage decoding accuracy in participants in University C, as measured at the grapheme-level*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=23)** | 67.77% | 67.27% | 8.26% | 50% | 79% | 81.82% | 79.77% | 7.2% | 67% | 91% |
| **Comparison (N=19)** | 72.73% | 66.59% | 14.3% | 30% | 79% | 77.69% | 76.64% | 7.57% | 59% | 87% |

*Figure 4.10 Histograms of percentage decoding accuracy in participants in University C, as measured at the grapheme-level*



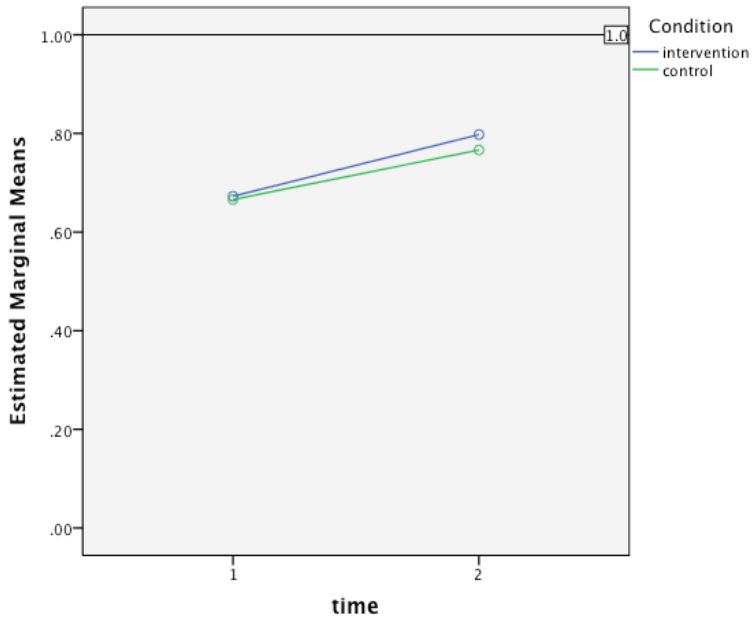

It can be seen that participants in University C achieved a higher accuracy percentage at the grapheme-level than at the word-level. Both the intervention group and the comparison group achieved roughly 35% accuracy in word-level decoding at t1, and approximately 39% at t2. In contrast, both groups correctly decoded around 70% of the total graphemes at t1, and nearly 80% at t2.

The grapheme-level decoding scores of University C were firstly analysed using

non-parametric techniques. A Wilcoxon Signed Ranks test showed that the grapheme-level decoding scores at t1 differed significantly from those at t2 for both intervention students ($Z$ = -4.20, $p$ < .001) and comparison students ($Z$ = -3.55, $p$ < .001). However, when comparing the scores of the two groups at each time point, a Mann-Whitney test found no significant difference at either t1 ($U$ = 186.50, $Z$ = -.81, $p$ = .42) or t2 ($U$ = 173.50, $Z$ = -1.14, $p$ = .25). This shows that the intervention group did not significantly differ from the comparison group at either time point, indicating that the instruction programme did not lead to significantly greater progress in phonological decoding for University C, as measured at the individual grapheme level.

The grapheme-level decoding scores of University C were also analysed using ANOVA, as a point of comparison for the non-parametric tests. There was a significant main effect of time, $F(1, 40) = 64.20$, $p$ <. 001, and the effect size was large ($r$ = .78). The main effect of condition was found to be non-significant, $F(1, 40) = .53$, $p$= .47. The interaction between time and condition was also non-significant, $F(1, 40) = .76$, $p$ =.39. The interaction graph is shown in Figure 4.11. This, again, is in accordance with the findings of the non-parametric tests. The analysis of the grapheme-level decoding scores is in accordance with that of the word-level decoding scores, where no significant difference was observed between intervention and comparison groups in University C.

*Figure 4.11 Estimated marginal means of percentage decoding accuracy in participants in University C, as measured at the grapheme-level*

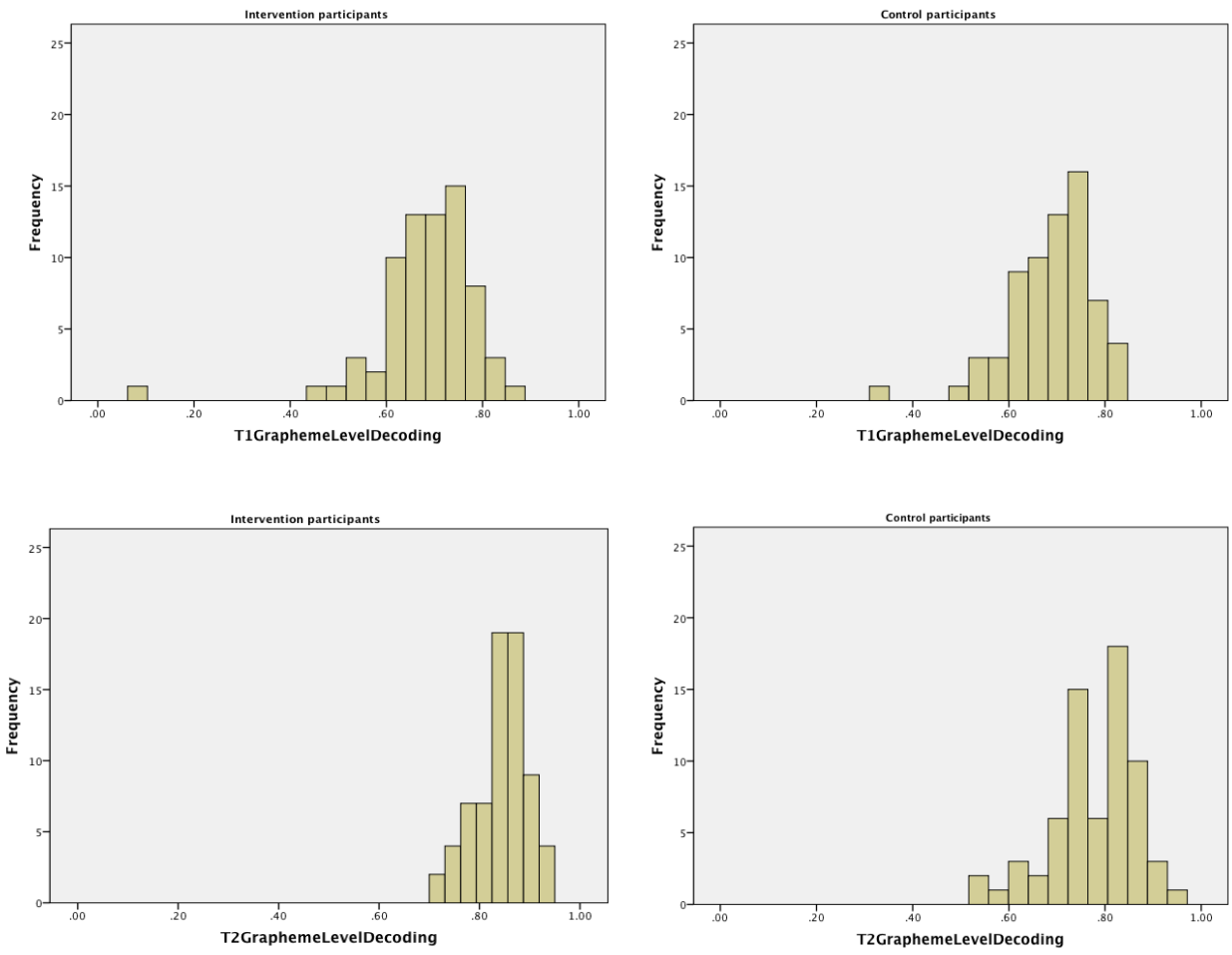



The descriptive statistics for the grapheme-level decoding scores of Universities A and B combined are shown in Table 4.10. The histograms are presented in Figure 4.12.

*Table 4.10 Percentage decoding accuracy in participants in Universities A and B, as measured at the grapheme-level*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 69.42% | 68.28% | 10.52% | 8% | 86% | 85.12% | 84.31% | 4.93% | 71% | 93% |
| **Comparison (N=67)** | 70.25% | 69.10% | 8.82% | 33% | 83% | 79.34% | 77.74% | 8.75% | 54% | 96% |

*Figure 4.12. Histograms of percentage decoding accuracy in participants in each group in Universities A and B combined, as measured at the grapheme-level*




It can be seen that the intervention participants correctly decoded 16% more of the graphemes at t2 than at t1, though their word-level progress was only 12% (see section 5.2). The comparison group's grapheme-level progress (from t1 to t2) was roughly 8%, while their word-level progress was 6%. In other words, both intervention and comparison groups appear to have made greater progress at the grapheme-level than the word-level. Moreover, both groups only correctly decoded approximately 10 words at t1, which was 35% of the total number. In contrast, their grapheme-level accuracy percentage had already reached nearly 70% at t1. Similarly, the intervention group achieved a mean score of 15 words at t2, slightly above 50% of

the total number of words, which was in marked contrast to their 85% accuracy percentage at the grapheme-level.

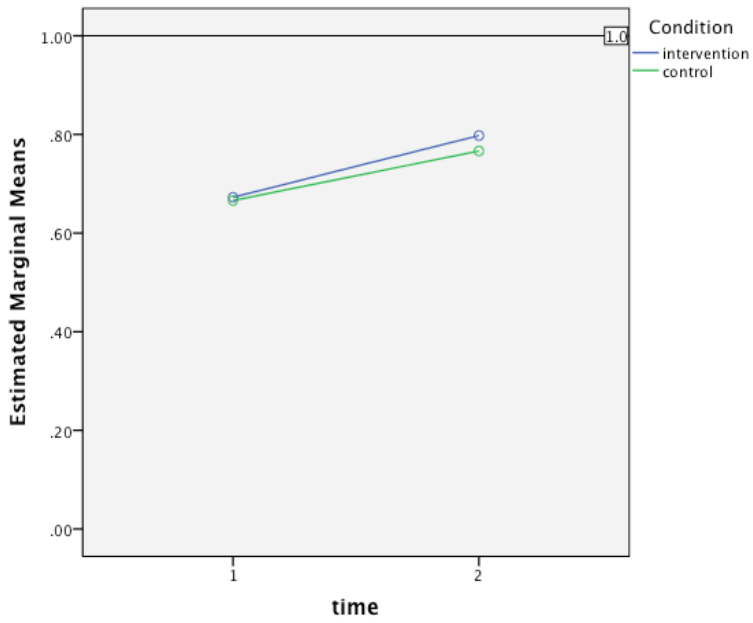
Non-parametric tests were firstly conducted to analyse the aggregated grapheme-level decoding scores of each group (intervention and comparison) in Universities A and B. Firstly, the differences between each group's scores at t1 and t2 were compared using a Wilcoxon Signed Rank test. For both groups, scores at t1 differed significantly from those at t2 (intervention group: $Z = -7.33$, $p < .001$, $r = -.62$; comparison group: $Z = -6.30$, $p < .001$, $r = -.54$). As the mean grapheme-level decoding scores were higher at t2 than t1 for both groups, this demonstrates that both groups made significant progress in English decoding when measured at the grapheme level. Then, a Mann-Whitney test was conducted to examine the difference between the scores of the two groups at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U = 2298.5$, $Z = -.23$, $p = .73$), but at t2 they differed significantly, with a medium effect size ($U = 1204.5$, $Z = -5.01$, $p < .001$, $r = -.30$).

An ANOVA was also conducted as a point of comparison. There was a significant main effect of time with a large effect size, $F(1, 136) = 281.53$, $p < .001$, $r = .82$, as well as a significant main effect of condition with a small effect size, $F(1, 136) = 5.32$, $p < .05$, $r = .19$. The interaction between time and condition was also significant, $F(1, 136) = 25.30$, $p < .001$, with a medium effect size ($r = .40$). The interaction graph is presented in Figure 4.13. These results support the findings of the non-parametric

tests.

Overall, the findings suggest that the phonics instruction led to significantly greater progress in terms of grapheme-level decoding for the intervention group than the comparison group in universities A and B.

*Figure 4.13 Estimated marginal means of percentage decoding accuracy in participants in Universities A and B, as measured at the grapheme-level*




In summary, the analysis of the grapheme-level decoding scores yields similar findings to the analysis of word-level decoding results. Intervention participants in Universities A and B made significantly greater progress between t1 and t2 than comparison groups in their English phonological decoding, both when measured at the word-level and when measured at the grapheme-level. Conversely, the phonics instruction did not lead to significantly more progress for the intervention group in

University C in either word-level or grapheme-level decoding scores. The comparison between word-level and grapheme-level decoding scores also reveals a considerable gap between the two: though intervention participants in Universities A and B only decoded slightly over 50% of the words correctly after the phonics instruction, their grapheme-level accuracy reached 85%. This will be further discussed in section 8.2.2.

### 4.4 Overall time of decoding

As described in section 3.5.2.3, the time between the onset of the first stimulus to the end of the final production of the last stimulus was recorded as the overall time spent on the decoding test. The assumptions of a two-way mixed factorial ANOVA with one between-subjects variable (condition) and one within-subjects variable (time) were firstly checked. Firstly, Mauchly's test of sphericity was not considered as the repeated measures had only two levels. Secondly, the assumption of Normality was retained at t1 but rejected at t2[23]. Thirdly, the assumption of homogeneity of variance was met at t1 but rejected at t2[24]. As a result, the data was primarily analysed using non-parametric tests. However, ANOVA tests were also conducted as a point of comparison.

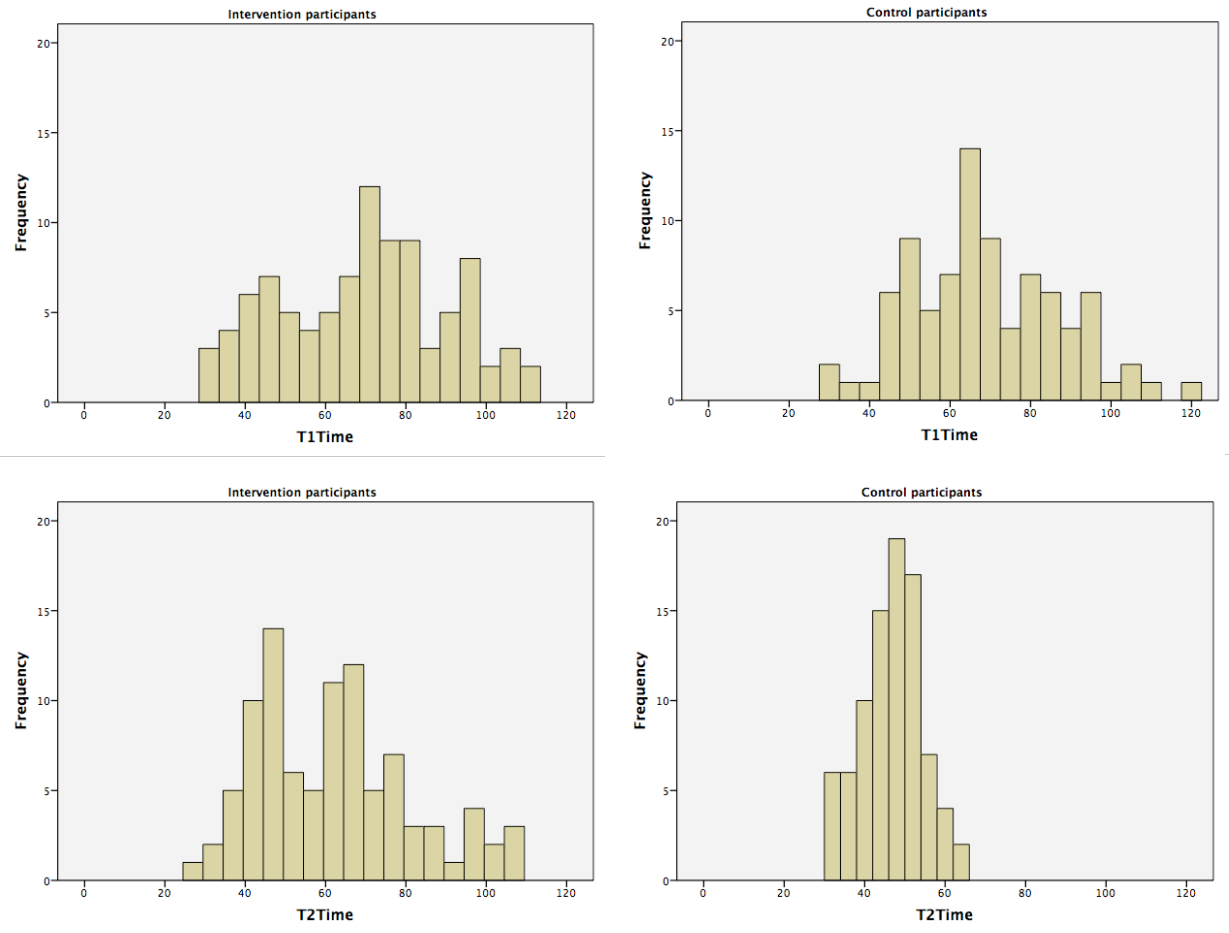
The descriptive data for all intervention and comparison groups is shown in Table 4.11. The histograms are shown in Figure 4.14.

[23] t1 *z*-score of skewness = .48, *z*-score of kurtosis = -1.72; t2 *z*-score of skewness= 6.62, *z*-score of kurtosis= 3.32
[24] t1 $F(1, 178) = .1.45$, $p = .23$; t2, $F(1, 178) = 51.96$, $p < .001$

*Table 4.11 Overall time of decoding for all participants (measured in seconds)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 72 | 70 | 21 | 31 | 111 | 61 | 62 | 19 | 27 | 107 |
| **Comparison (N=86)** | 67 | 69 | 19 | 30 | 119 | 47 | 46 | 8 | 30 | 62 |

*Figure 4.14. Overall time of decoding for all participants (measured in seconds)*




Decoding proficiency may be operationalized in terms of both the accuracy and the speed with which written words are named (*e.g.* Hamada & Koda, 2008). However, it can be seen that the intervention group, whose decoding scores significantly increased

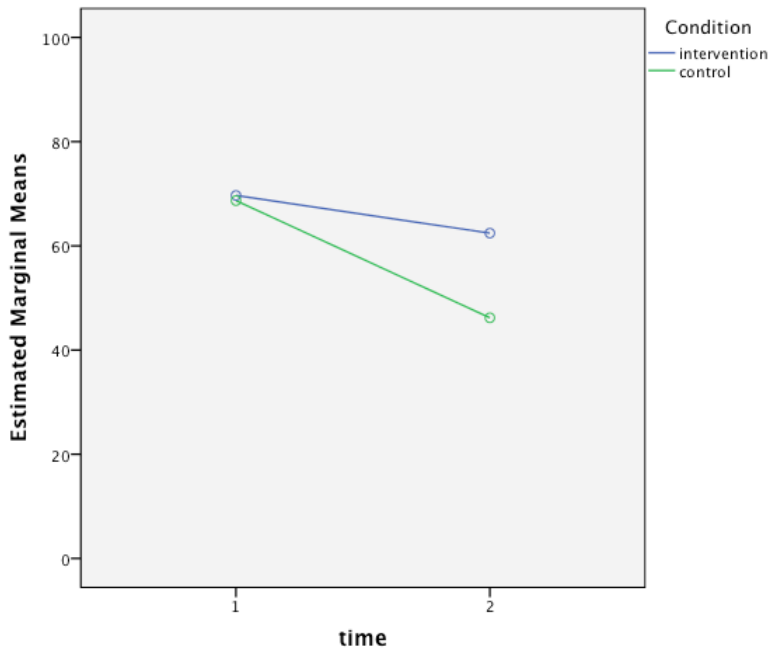
at t2, showed only a slight decrease in overall time of decoding: i.e. they became only slightly quicker at decoding the items. Interestingly, however, the comparison group, though not showing a significant increase in decoding *accuracy* between time t1 and t2, appeared to spend much less time decoding the stimuli. In other words, an absence of explicit phonics instruction seems to have been associated with faster decoding at time 2. The variation in terms of overall time of decoding also appears to be much lower among comparison groups at t2, as the standard deviation was much smaller than at t1. By contrast, this pattern was not observed for the intervention group. This will be discussed in section 8.2.3.

Non-parametric tests were firstly conducted to analyse the overall time of decoding for all participants. Firstly, the differences between each group's overall time of decoding at t1 and t2 were compared using a Wilcoxon Signed Rank test. For both the intervention and comparison group, the overall time of decoding at t1 differed significantly from that at t2 (intervention group: $Z = -4.03$, $p < .001$, $r = -.29$; comparison group: $Z = -7.73$, $p < .001$, $r = -.56$). As the overall time of decoding was shorter at t2 than at t1 for both groups, this shows that both the intervention and the comparison group responded significantly faster at t2. Then, a Mann-Whitney test was conducted to examine the differences between the overall time of decoding of the two groups at each time point. The results show that the overall time of decoding of the two groups was not significantly different at t1 ($U = 3834.5$, $Z = -.59$, $p = .55$) but was significantly different at t2, with a medium effect size ($U = 1910.5$ $Z = -6.11$, $p$

< .001, $r$ = -.32), indicating that the participants who did not follow the programme of phonics instruction showed a significantly greater decrease in overall time of decoding compared to those who participated in the instruction programme.

A two-way mixed factorial ANOVA was also conducted as a point of comparison. The results showed a significant main effect of time, $F(1, 178) = 136.77$, $p < .001$, with a large effect size ($r = .66$), as well as a significant main effect of condition with a small effect size, $F(1, 178) = 14.55$, $p < .001$, $r = .27$. The interaction between time and condition was also significant, $F(1, 178) = 36.03$, $p < .001$, with a medium effect size ($r = .41$). The interaction graph is presented in Figure 4.15. The results are in accordance with the findings of the non-parametric tests as shown above, suggesting that the absence of explicit phonics instruction was associated with significant decreases in the overall time of decoding.

*Figure 4.15 Estimated marginal means of overall time of decoding for the intervention and comparison groups*




Participants' overall time of decoding was also analysed individually for each university. The descriptive statistics for the three universities are shown in Table 4.12. Assumptions of the two-way factorial ANOVA test were checked. Firstly, the assumption of Normality was violated for all three universities[25]. Secondly, Levene's test confirmed the homogeneity of variance for University C, but the null hypothesis that the variances were equal was rejected for University A and University B[26]. As a result, both non-parametric and parametric tests were conducted as a point of comparison.

[25] University A: t1 *z*-score of skewness = 2.08, *z*-score of kurtosis = .46; t2 *z*-score of skewness = 4.18, *z*-score of kurtosis = 2.74; University B: t1 *z*-score of skewness = -1.12, *z*-score of kurtosis = .02; t2 *z*-score of skewness = 3.37, *z*-score of kurtosis = .62; University C: t1 *z*-score of skewness = 2.83, *z*-score of kurtosis = 1.90; t2 *z*-score of skewness = 0.59, *z*-score of kurtosis = .01

[26] University A: t1 $F(1, 43) = 4.49$, $p < .05$; t2 $F(1, 43) = 15.36$, $p < .001$; University B: t1 $F(1, 91) = .89$, $p = .35$; t2 $F(1, 91) = 18.39$, $p < .001$; University C: t1 $F(1, 40) = 1.45$, $p = .24$; t2 $F(1, 40) = .04$, $p = .84$

*Table 4.12 Overall time of decoding grouped by university (measured by second)*

| | | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Uni A** | **Intervention (N=24)** | 55 | 61 | 21 | 32 | 111 | 57 | 63 | 20 | 40 | 106 |
| | **Comparison (N=21)** | 59 | 59 | 13 | 40 | 86 | 47 | 47 | 8 | 33 | 62 |
| **Uni B** | **Intervention (N=47)** | 80 | 82 | 14 | 40 | 109 | 67 | 72 | 15 | 46 | 107 |
| | **Comparison (N=46)** | 75 | 75 | 16 | 32 | 106 | 48 | 47 | 7 | 30 | 62 |
| **Uni C** | **Intervention (N=23)** | 48 | 53 | 15 | 31 | 80 | 44 | 43 | 8 | 27 | 60 |
| | **Comparison (N=19)** | 63 | 65 | 23 | 30 | 119 | 43 | 44 | 7 | 31 | 61 |

The overall time of decoding of participants in each university were firstly analysed using non-parametric tests.

For University A, a Wilcoxon Signed Ranks test found that the overall time of decoding at t1 did not differ significantly from that at t2 for the intervention group ($Z$ = -1.26, $p$ = .21), but it did for the comparison group ($Z$ = -3.46, $p$ < .005). Then, a Mann-Whitney test found that the overall time of decoding of the two groups were not significantly different at t1 ($U$ = 241.50, $Z$ = -.24, $p$ = .81), but were significantly different at t2 ($U$ = 127, Z = -2.85, $p$ < .005), with a medium effect size ($r$ = .30). This shows that the lack of intervention was associated with a significantly greater decrease in overall time of decoding for the comparison group in University A.

For University B, a Wilcoxon Singed Ranks test found that the overall time of

decoding at t1 differed significantly from that at t2 for both the intervention group ($Z$ = -4.14, $p$ < .001) and the comparison group ($Z$ = -5.91, $p$ < .001), suggesting both groups took less time to complete the decoding test at t2 than t1. Then, a Mann-Whitney test found that the overall time of decoding of the two groups were not significantly different at t1 ($U$ = 788.5, $Z$ = -2.25, $p$ = .25) but were significantly different at t2 ($U$ = 100, $Z$ = -7.54, $p$ < .001), with a large effect size ($r$ = -.55). This is in accordance with the findings for University A, suggesting that the participants who did not follow the phonics instruction programme showed a significantly greater decrease in overall time of decoding than those who received the instruction.

For University C, a Wilcoxon Signed Ranks test found that the overall time of decoding at t1 differed significantly from that at t2 for both the intervention group ($Z$ = -3.00, $p$ < .005) and the comparison group ($Z$ = -3.32, $p$ < .005), indicating that both groups completed the decoding test faster at t2 than t1. Then, a Mann-Whitney test found that the overall time of decoding of the two groups was not significantly different either at t1 ($U$ = 152.5, Z = -1.67, $p$ = .10), or at t2 ($U$ = 204.5, Z = -.36, $p$ = .72), suggesting that the instruction programme did not distinguish the intervention and the comparison group in terms of the overall time of decoding.

Three two-way factorial ANOVA tests were also conducted to compare the overall time of decoding between intervention and comparison groups in each university. The results are as follows.

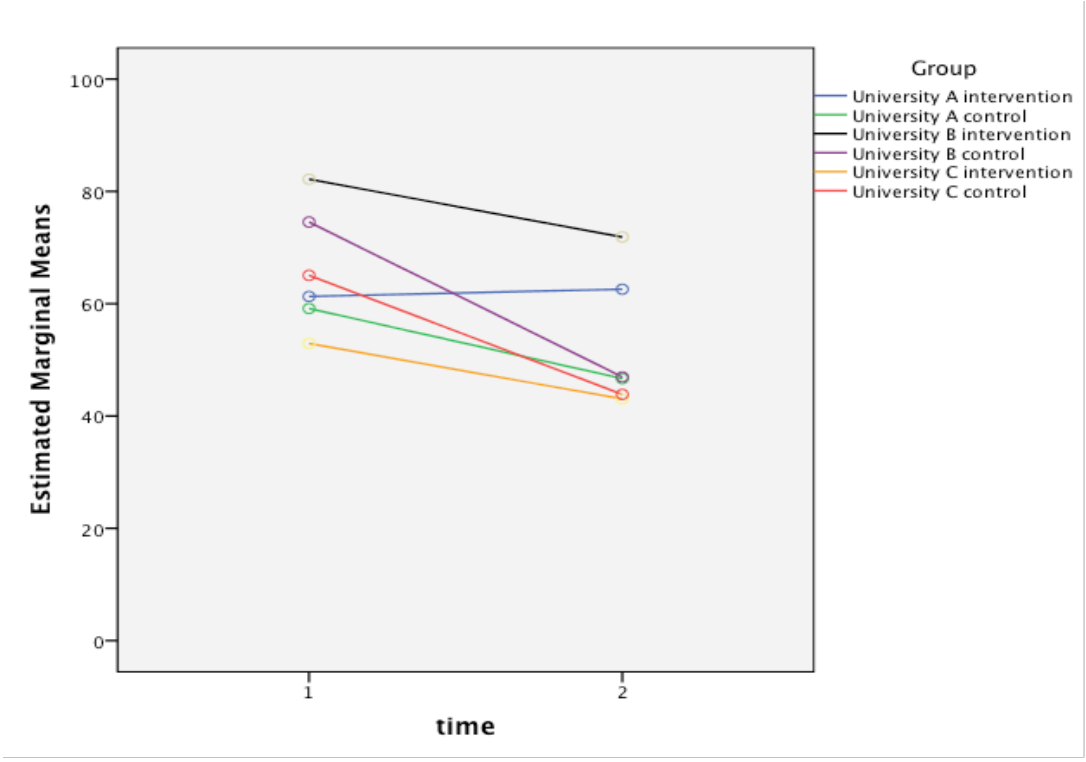
For University A, there was a significant main effect of time with a medium effect size, $F(1,43) = 4.14$, $p < .05$, $r = .30$, and also a significant main effect of condition with a medium effect size, $F(1,43) = 4.87$, $p < .05$, $r = .32$. The interaction between time and condition was also significant, $F(1, 43) = 6.28$, $p <. 05$, with a medium effect size ($r = .36$). This is in accordance with the findings of the non-parametric tests.

For University B, a significant main effect of time with a large effect size was observed, $F(1, 91) = 159.66$, $p <. 001$, $r = .80$. There was also a significant main effect of condition with a large effect size, $F(1, 91) = 44.31$, $p <. 001$, $r = .57$. The interaction between time and condition was also significant, $F(1, 91) = 33.25$, $p <. 001$, with a large effect size ($r = .52$). This is also in accordance with the findings of the non-parametric tests.

For University C, a significant main effect of time was observed, $F(1, 40) = 32.41$, $p <. 001$. A significant main effect of condition was not found, $F(1, 40) = 3.20$, $p = .08$. The interaction between time and condition was also non-significant, $F(1, 40) = 4.32$, $p = .05$. This again is in accordance with the findings of the non-parametric tests.

The interaction graph of the six groups in the three universities is presented in Figure 4.16. It can be seen that the intervention group in University A was the only group which showed an increase in the overall time of decoding at t2.

*Figure 4.16 Estimated marginal means of overall time of decoding of all six groups of participants in the three universities*




As University C was previously found to perform differently in terms of both word-level and grapheme-level decoding scores (see section 4.3), it is of interest to conduct another analysis that includes only University A and University B, in order to provide an evaluation of the phonics instruction programme on participants' overall time of decoding without the possibly confounding effect of concurrently learning another foreign language (as was the case for participants in University C).

The descriptive statistics of the overall time of decoding of participants in Universities A and B are presented in Table 4.13. The histograms are shown in Figure 4.17. The assumptions of the two-way mixed factorial ANOVA were firstly checked. The assumption of Normality[27] and homogeneity of variance[28] were both retained at

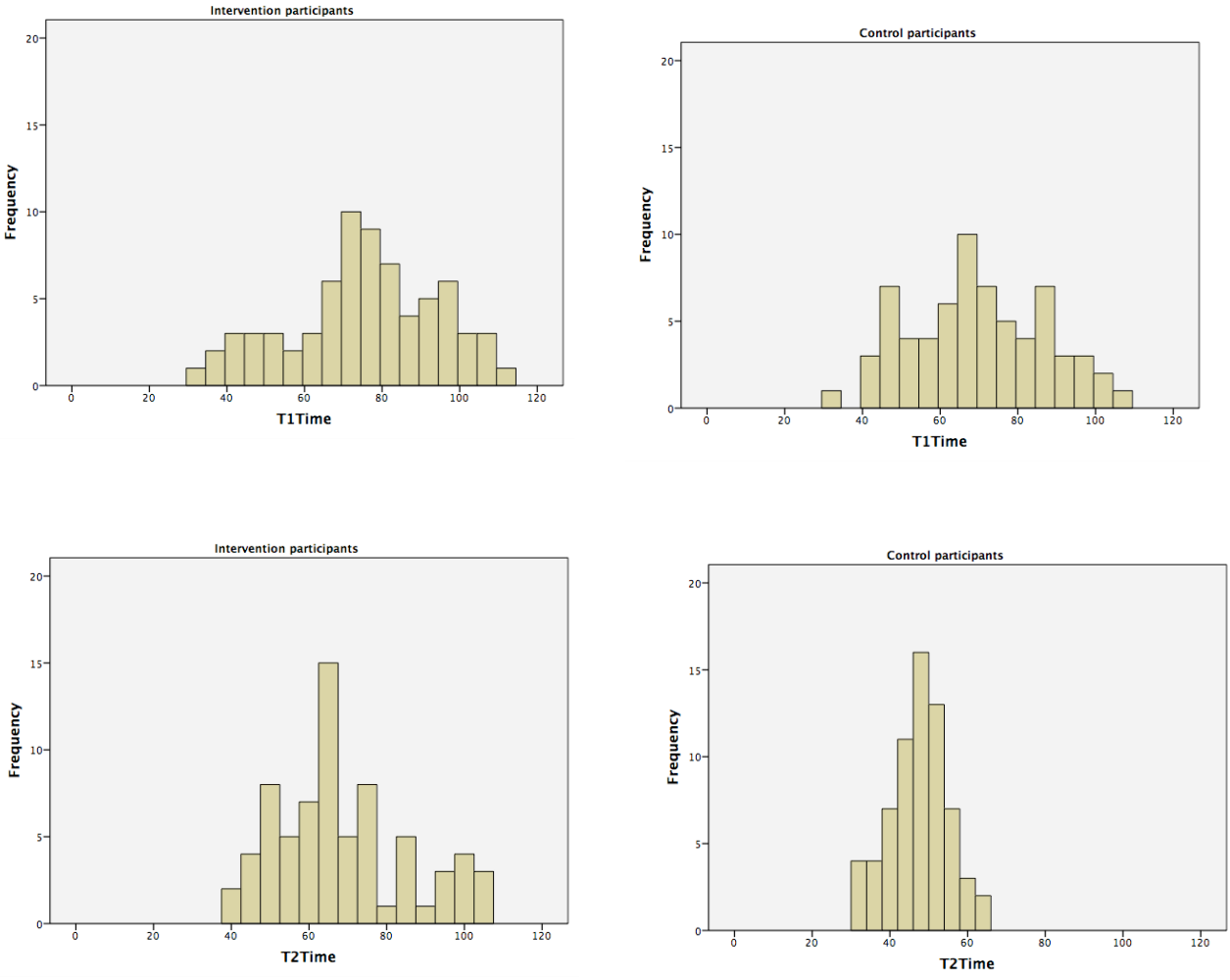
[27] t1 *z*-score of skewness= -.48, *z*-score of kurtosis = -1.51; t2 *z*-score of skewness= 4.91, *z*-score of kurtosis= 1.40

t1 but rejected at t2. Therefore, the data was primarily analysed using non-parametric techniques while the ANOVA was also conducted as a point of comparison.

*Table 4.13. Overall time of decoding of participants in Universities A and B (measured by second)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 75 | 75 | 19 | 32 | 111 | 66 | 69 | 17 | 40 | 107 |
| **Comparison (N=67)** | 69 | 70 | 17 | 32 | 106 | 47 | 47 | 7 | 30 | 62 |

*Figure 4.17 Histograms of overall time of decoding of participants in Universities A and B*



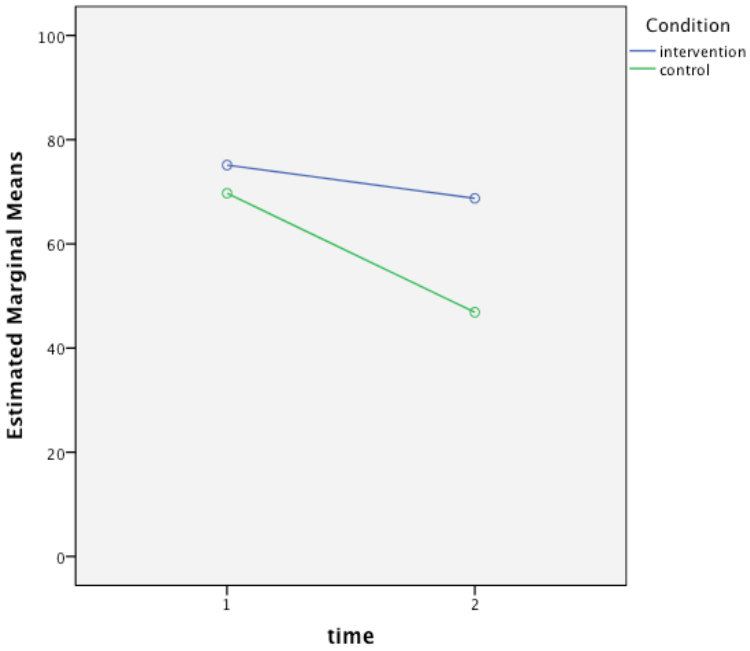

[28] At t1 $F(1, 136) = .48$, $p = .49$; at t2 $F(1, 136) = 32.45$, $p < .001$.

Non-parametric tests were firstly conducted to analyse the overall time of decoding of participants in University A and B taken together. Firstly, the differences between the intervention group's and the comparison group's overall time of decoding at t1 and t2 were compared using a Wilcoxon Signed Rank test. For both the intervention and comparison group, the overall time of decoding at t1 differed significantly from that at t2 (intervention group: $Z = -3.02$, $p < .01$, $r = -.26$; comparison group: $Z = -6.95$, $p < .001$, $r = -.58$). As the mean overall time of decoding was longer at t1 than t2 for both groups, this indicates that both the intervention and comparison group responded significantly faster at t2. Then, a Mann-Whitney test was conducted to examine the differences between the overall time of decoding of the two groups at each time. The results show that the overall time of decoding of the two groups was not significantly different at t1 ($U = 1912$, $Z = -1.95$, $p = .06$) but at t2 it was significantly different, with a medium effect size ($U = 569.5$, $Z = -7.71$, $p < .001$, $r = -.46$). These results indicate that the participants in Universities A and B who did not follow the programme of phonics instruction showed a significantly greater decrease in overall time of decoding than those who did participate in the instruction programme.

A two-way mixed factorial ANOVA was also conducted. There was a significant main effect of time with a large effect size, $F(1, 136) = 102.77$, $p < .001$, $r = .65$, and also a significant main effect of condition with a medium effect size, $F(1, 136) = 34.55$, $p < .001$, $r = .45$. The interaction between time and condition was also significant, $F(1, 136) = 32.63$, $p < .001$, with a medium effect size ($r = .44$).

Compared with the previous ANOVA, it was found that the effect size of the between-subjects variable (condition) was only small when University C was included ($r$ = .27) but grew to medium when University C was excluded ($r$ = .45), and the effect size of the interaction effect also showed a small growth (from $r$ = .41 to $r$ = .44). These results are in accordance with the findings of the non-parametric tests, suggesting that the lack of explicit phonics instruction was associated with a significantly greater decrease in overall time of decoding for the comparison groups in Universities A and B. The interaction graph is presented in Figure 4.18.

*Figure 4.18 Estimated marginal means of overall time of decoding of Universities A and B participants*




In summary, the results show that the participants who did not receive the phonics instruction showed a significantly greater decrease in overall time of decoding than those who participated in the phonics instruction programme.

## 4.5 Summary

The various strands of quantitative analysis conducted in this chapter suggest that the participants who followed the phonics instruction programme made significantly more progress in both word-level and grapheme-level decoding accuracy compared to those in the comparison group, even though the two groups were matched in their English proficiency and vocabulary knowledge, as measured by the NCEEE and the BPVS. On average, participants in the intervention group correctly decoded 4 more words out of the 28 test items after the intervention, though they still had plenty of room for progress as they still decoded only around half of all the test items correctly at t2. The intervention participants correctly decoded less than 70% of the total graphemes at t1, but more than 85% of the total graphemes at t2, which also clearly demonstrates their progress in decoding at grapheme level. These results provide convincing evidence of the effectiveness of the phonics instruction programme in terms of promoting English decoding accuracy. In addition, it was found that the comparison groups showed a significantly greater decrease in the overall time of decoding than the intervention participants, which seems to be counter-intuitive, as greater accuracy of decoding is often associated with faster decoding speed (or shorter overall time of decoding as operationalized in this study). This will be further discussed in Chapter 8.

The analysis by university reveals that, unlike in the other two universities, the

intervention group in University C did not demonstrate significantly greater progress in accuracy of decoding, either at the word level or the grapheme level. In addition, the two groups in University C did not differ significantly in terms of their overall time of decoding at either t1 or t2. A possible reason for this difference between University C and the other two universities is that the participants in University C were concurrently learning another foreign language in addition to English during the intervention period. For the intervention participants in University C, this additional language (French) also uses the same Roman alphabet as English, but has different GPCs.

Having established that University C behaved differently, some additional analyses were also conducted excluding University C (i.e. the combined scores of Universities A and B were analysed). The results of these analyses confirmed those of the original analyses (i.e. with all three universities) but the differences between the intervention and comparison groups were larger.

## Chapter 5. Findings II. Phonological Decoding Test – Accuracy of Individual GPCs

The previous chapter provided a quantitative analysis of the phonological decoding test results, by comparing intervention and comparison groups in terms of both the accuracy and speed of their decoding before and after the instruction programme. However, what this level of analysis did not investigate was whether the phonics instruction was more effective for some graphemes than others, which could possibly help inform the design of future instruction programmes. Therefore, this chapter provides a more nuanced line of analysis by examining participants' accuracy percentages for individual GPCs before and after instruction. More specifically, three sets of results are presented. Firstly, the accuracy percentages for individual graphemes at t1 are examined, in order to provide a baseline for the decoding test and to identify participants' strengths and weaknesses in English decoding before the instruction programme. Secondly, the accuracy percentages for individual graphemes at t2 are also examined, in order to demonstrate participants' command of individual graphemes after the instruction programme. Thirdly, the progress rates for individual graphemes are also calculated in order to examine whether some graphemes were more amenable to the instruction programme than others. Therefore, this chapter addresses Research Question 2:

*RQ2: Is the programme of phonics instruction more effective for some GPCs than*

*others?*

As the previous chapter has shown, the intervention participants in University C did not demonstrate significantly greater progress in English decoding than the comparison groups after the instruction programme. Given the purpose of this chapter is to identify whether the phonics instruction programme is more effective for some graphemes than others, there is no point in looking for variations where there is a lack of overall change, which is probably due to the confound of learning another foreign language.

Given that University C behaved differently probably because of learning French at the same time (which is supported by the errors they made, such as pronouncing <oi> as /wa/, which is how the grapheme is pronounced in French. Interested readers may refer to Appendix 5 for more details), it was therefore decided to conduct this analysis of individual graphemes only for Universities A and B, which had a 'purer' form of intervention (i.e. the participants were learning to decode only in English and not another Roman alphabetic writing system concurrently).

Graphemes that correspond to regular English spellings of vowels (henceforth 'vowel GPCs') and those that correspond to regular English spelling of consonants (henceforth 'consonant GPCs') were analysed separately. This is not only because English phonemes are conventionally divided into these two articulatory categories

(Chomsky, 1968), but also because Chinese EFL learners seem to encounter more difficulty in the decoding of vowels than in the decoding of consonants, based on the researcher's observation in the pilot study. Where a grapheme has both consonant and vowel phonemic realisations (as in the case of <y> = /j/ or /aɪ/ or /ɪ/), these cases were analysed separately. Likewise, when a single grapheme has two vowel or two consonant phonemic realisations (as in the case of <c> = /s/ or /k/), the two grapheme-phonemes correspondences were analysed separately.

This chapter is divided into three sections. Section 5.1 presents the accuracy percentages for the consonant GPCs in the decoding test. Section 5.2 presents the accuracy percentages for the vowel GPCs. Section 5.3 summarises the results of this chapter.

## 5.1 Consonant GPCs

The 35 consonant GPCs and the number of occurrences in the test for each correspondence are presented in Table 5.1. As explained above, some graphemes correspond to more than one phoneme in the test. In such cases, these are shown separately in the table and analysed separately. It is worth noting that the number of occurrences for some GPCs was smaller than others. In order to examine whether the frequency of occurrence of the GPCs in the test mirrors their frequency of occurrence in the language generally, a Spearman's correlation test was conducted, as both

variables (the number of occurrences of the GPCs in the test and their frequency of occurrence in the language) were ordinal. The frequencies of occurrence of individual GPCs in the language are taken from Gontijo et al. (2003), where a computational analysis of the frequency of occurrence of GPCs in British English was conducted using a corpus of more than 160,000 words. A significant and strong correlation was found ($\rho = .78$, $p < .001$) between frequency of occurrence in the decoding test and frequency of occurrence in the language. This suggests that though the numbers of occurrences were different for the different consonant GPCs in the test, this generally mirrors their frequency of occurrence in the language. This lends support to the robustness of the grapheme-level analysis of the decoding test from a different perspective, given that the more frequently encountered GPCs were tested more often in the decoding test. That being said, it is still worth noting that those GPCs that had more occurrences in the decoding test might afford more reliable analyses, given that they were tested in various orthographic contexts.

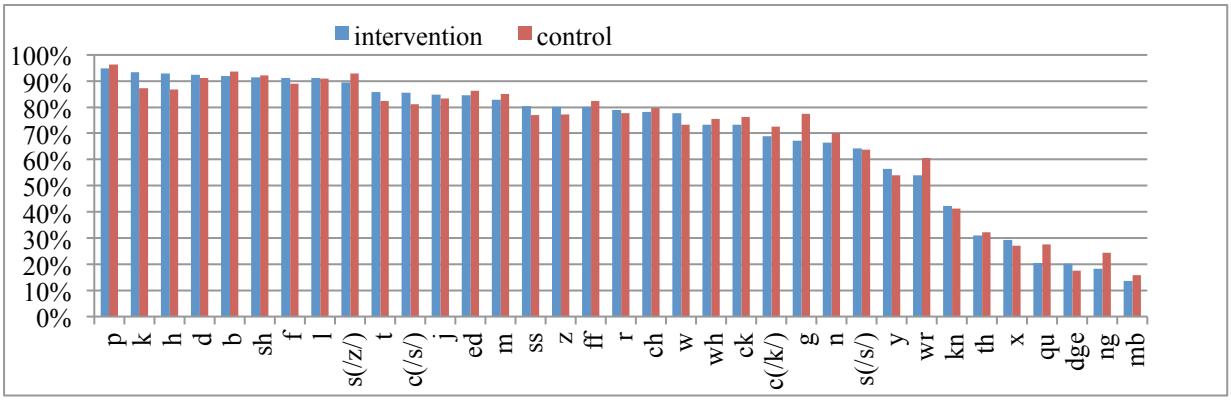
*Table 5.1 Number of occurrences in the decoding test of each consonant GPCs*

| Number of occurrences | GPCs |
|---|---|
| 1 | <sh> = /ʃ/; <x> = /ks/; <z> = /z/; <dge> = /dʒ/; <mb> = /m/; <ff> = /f/; <ss> = /s/; <th> = /θ/; <c> = /k/ |
| 2 | <y> = /j/; <wh> = /w/; <kn> = /n/; <ck> = /k/; <ch> = /tʃ/; <ed> = /d/; <qu> = /kw/; <h> = /h/; <j> = /dʒ/; <ng> = /ŋ/; <wr> = /r/ |
| 3 | <k>= /k/; <w>= /w/; <s>= /z/ |
| 5 | <g>= /g/ |
| 6 | <f> = /f/; <c> = /s/ |
| 7 | <s> = /s/ |
| 9 | <p> = /p/ |
| 10 | <m> = /m/ |
| 11 | <d> = /d/; <r> = /r/ |
| 12 | <b> = /b/; <n> = /n/; <l> = /l/ |
| 15 | <t> = /t/ |

### 5.1.1. Mean accuracy percentage at t1

The mean accuracy percentages for all the consonant GPCs of Universities A and B at t1 are presented in Figure 6.1 (the corresponding phonemes are not presented in the figure to save space, but they can be found in Table 5.1). The graphemes are ranked from left to right in descending order of the accuracy percentage that the intervention participants achieved.

*Figure 5.1. Universities A and B's mean accuracy of consonant GPCs at t1*



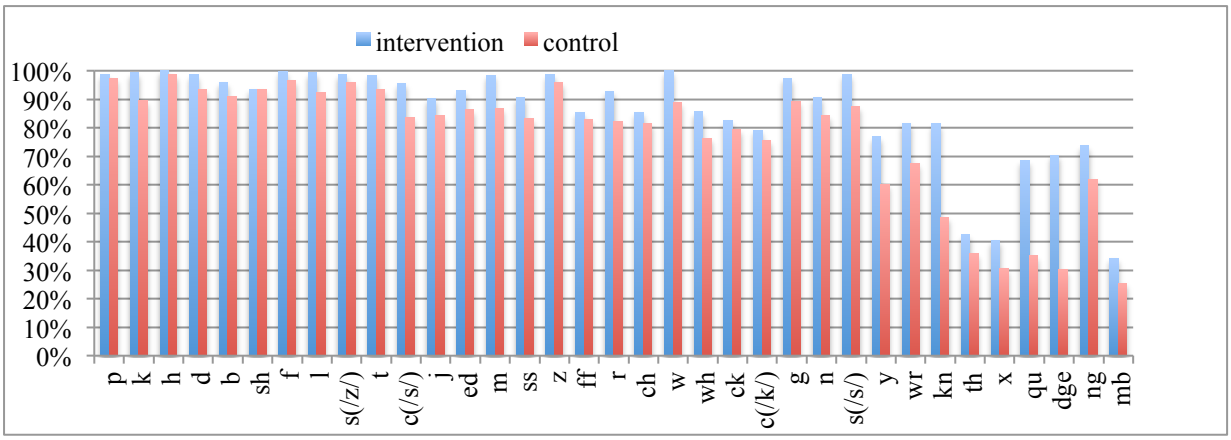
Visual inspection of Figure 6.1 suggests that the intervention and comparison groups achieved fairly similar accuracy percentages across the consonant GPCs, suggesting that the two groups were roughly matched in their decoding of consonant GPCs at t1. Statistical analysis was also conducted to compare the scores achieved in all the consonant GPCs by the two groups of participants (where each correctly decoded grapheme was scored as 1 and otherwise 0, following the grapheme-level scoring system as discussed in Chapter 4). A paired samples t-test found no significant difference between the two groups, $t(34) = -.39$, $p = .71$. This suggests that the intervention and comparison groups in Universities A and B were similar in the accuracy with which they decoded consonant graphemes at t1.

It can also be seen from Figure 6.1 that both the intervention and comparison groups already appeared to have a good command of most of the consonant GPCs at t1, though a few difficult graphemes still posed a challenge (such as <kn>, <th>, <x>, <qu>, <dge>, <ng> and <mb>).

### 5.1.2 Mean accuracy percentage at t2

The mean accuracy percentages for all the consonant GPCs at t2 are presented in Figure 5.2. The graphemes are presented in the same order as in Figure 5.1 for convenience of comparison.

*Figure 5.2 Universities A and B's mean accuracy of consonant GPCs at t2*




Visual inspection of Figure 5.2 shows that the intervention participants in Universities A and B, at a descriptive level, outperformed their comparison group counterparts in the accuracy with which they decoded all consonant graphemes at t2. This was also tested by statistical analysis, where the overall scores achieved by each group on all consonant GPCs were compared using a paired-samples t-test. Significant differences were found between the two groups, $t(34) = 6.25$, $p < .001$. This finding, together with the results presented in Figure 5.2, suggests that the intervention participants in Universities A and B significantly outperformed the comparison groups in the decoding of the consonant graphemes at t2.

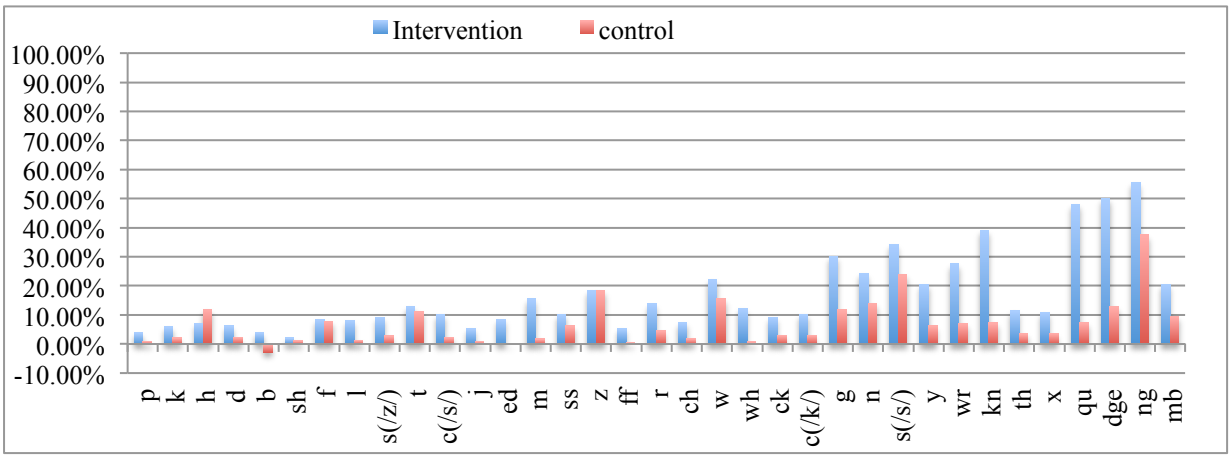
It can also be seen from Figure 5.2 that the two groups of participants achieved similar accuracy percentages in decoding the consonant graphemes towards the left-hand end of Figure 5.2, which were the graphemes that they decoded best at t1. This suggests that for the consonant GPCs that the participants already had good command of at t1, the phonics instruction programme did not lead to distinctively better performance for the intervention group, though this is expected as these GPCs

had a high baseline (and therefore relatively little room for improvement). In contrast, the intervention group's advantage was very clear for some GPCs towards the right-hand end of Figure 5.2, which were the ones with the lowest mean accuracy at t1. For instance, the intervention group appears to have outperformed the comparison group by more than 40% in decoding <qu> and <dge>, both of which only had an accuracy of roughly 20% at t1. However, it is interesting to notice that the intervention group's advantage was much smaller for some GPCs that had low accuracy at t1. An example was the decoding of grapheme <th>, where the intervention group only outperformed the comparison group by approximately 5 percentage points. This will be further discussed in section 8.3.1.

### 5.1.3 Progress

The difference between each consonant GPC's mean accuracy at t2 and t1 was calculated as a measure of progress, which is presented in Figure 5.3. The graphemes are presented in the same order as in Figures 5.1 and 5.2, that is, in descending order of accuracy of decoding at t1.

*Figure 5.3 Difference between mean accuracy of consonant GPCs at t2 and t1*




It can be seen that the intervention participants appeared to have consistently made more progress than the comparison groups in decoding almost all the consonant graphemes with only one exception (<h>), though both groups achieved nearly 100% accuracy in decoding the grapheme <h> at t2, as can be seen in Figure 6.2. The progress made by the two groups of participants in all consonant GPCs were also compared by statistical analysis, using a paired-samples t-test. A significant difference was found between the two groups, $t(34) = 5.73$, $p < .001$. This finding, together with the results presented in Figure 5.3, indicates that the intervention group made more progress than the comparison group in the accuracy of decoding consonant graphemes.

It can be seen from Figure 5.3 that the intervention group appeared to make limited progress in decoding most of the consonant graphemes towards the left-hand end, which is presumably because these graphemes were already correctly decoded in most cases at t1; hence the room for progress was small. In contrast, the graphemes at the right-hand end, which were the ones with the lowest mean accuracy at t1, generally

saw considerably more progress at t2. Indeed, three graphemes had an increase in accuracy of approximately 50 percentage points (<qu>, <dge>, <ng>). In contrast, some graphemes which had low accuracy at t1 seemed to resist the effects of instruction at t2, with graphemes <th> and <x> seeing an increase in accuracy of only 10 percentage points. Another thing worth noticing is that the comparison group also made good progress in decoding some consonant graphemes even without the phonics instruction. For instance, the grapheme <s>= /s/ saw more than 20 percentage points of progress, and the grapheme <ng> more than 30.

Even though the baseline varies across different consonant GPCs, it is still worth noticing that some GPCs, despite having similar accuracy percentages at t1, nonetheless varied in the degree of improvement with which they were associated for the intervention group. For instance, the decoding of the graphemes <ch> <w> and <wh> all had approximately 75% accuracy at t1; at t2, the decoding of <ch> and <wh> saw an increase in accuracy of approximately 1 percentage point, while <w> witnessed a large increase in accuracy of nearly 25 percentage points. Similarly, the decoding of <qu> <dge> <ng> and <mb> all presented great difficulty at t1, with average accuracy scores below 20%; at t2, the decoding of the graphemes <qu> <dge> and <ng> saw considerable progress of more than 40 percentage points, while the decoding of <mb> only saw moderate progress of 20 percentage points. This will be further discussed in chapter 8.3.1.

### 5.1.4 Summary

In summary, the analysis of the accuracy percentages of the consonant GPCs shows that both the intervention and comparison groups at Universities A and B already had a good command of many consonant graphemes at t1, achieving over 80% accuracy in over a third of the total consonant graphemes. After the phonics instruction programme, more than two thirds of the consonant GPCs had high accuracy (over 80%) for the intervention participants, but some GPCs that were poorly decoded still presented problems. In addition, the intervention participants made more progress in almost all GPCs, and in some cases considerably more compared to the comparison groups. It was also found that for consonant GPCs with a similar baseline, some witnessed more progress than others after the phonics instruction, providing an affirmative answer to Research Question 2. The reasons that might contribute to this finding will be discussed in section 8.3.1.

## 5.2 Vowel GPCs

The 28 vowel GPCs and their number of occurrences in the test are presented in Table 5.2. Some vowel graphemes correspond to more than one phoneme in the test (for instance, <a> is decoded as /æ/ in some test items but as /ə/ in others); these are shown separately in the table and analysed separately. Similarly to the consonant GPCs, the number of occurrences for some vowel GPCs was smaller than for others. In order to examine whether the frequency of occurrence for the vowel GPCs in the

test mirrors their frequency of occurrence in the language generally, a Spearman's correlation test was again conducted with the corresponding figures from Gontijo et al. (2003). A significant, strong correlation was found ($\rho = .70$, $p < .001$), indicating that the number of occurrences of different vowel GPCs mirrors their frequency of occurrence in the language.

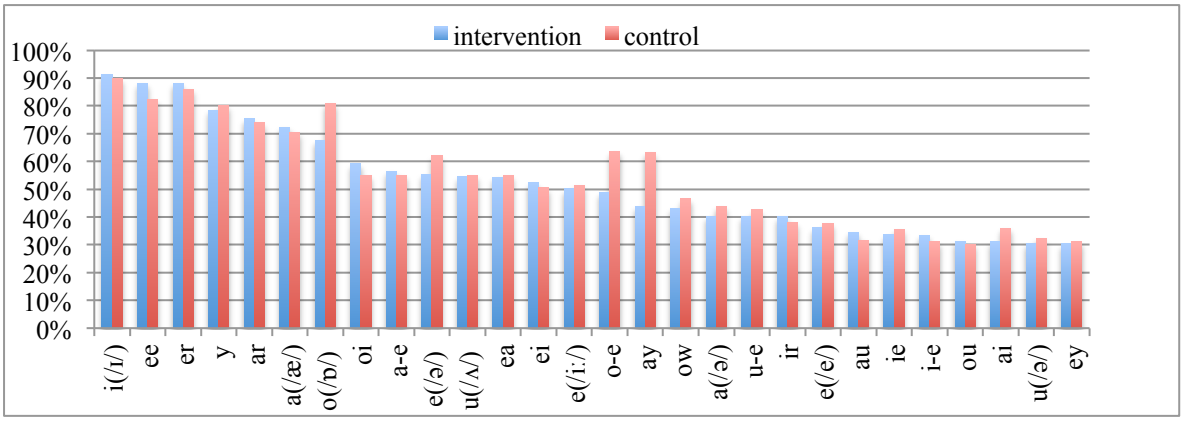
*Table 5.2 Number of occurrences in the decoding test for each vowel GPCs*

| Number of occurrences | GPCs |
|---|---|
| 1 | <ie> = /aɪ/; <ai> = /eɪ/; <ir>= /ɜː/; <a-e> = /eɪ/; <ey> = /eɪ/; <ei> = /iː/; <u-e> = /uː/; <au> = /ɔː/; <ou> = /aʊ/; <ar> = /ɑː/; <ow> = /əʊ/ |
| 2 | <ay> = /eɪ/; <oi> = /ɒɪ/; <e> = /ə/; <er> =/ə/; <ea> = /iː/; <e> =/iː/ |
| 3 | <ee> = /iː/; <y> = /iː/; <o-e> = /əʊ/; <i-e> = /aɪ/ ; <a> = /ə/; <u>= /ə/ |
| 4 | <o> = /ɒ/ |
| 7 | <i> = /ɪ/ |
| 8 | <u> = /ʌ/ |
| 9 | <e> = /e/ |
| 10 | <a> =/æ/ |

### 5.2.1 Mean accuracy percentage at t1

The mean accuracy percentages for all the vowel GPCs from Universities A and B at t1 are presented in Figure 5.4 (the corresponding phonemes are not presented in the figure to save space, but they can be found in Table 5.2). The graphemes from left to right are ranked in descending order of the mean accuracy percentage that the intervention participants achieved.

*Figure 5.4 Universities A and B's mean accuracy of vowel GPCs at t1*




Visual inspection of Figure 5.4 suggests that, at a descriptive level, the intervention and comparison groups in Universities A and B achieved broadly similar results across the decoding of most vowel graphemes in the decoding test (with the exception of graphemes <o>= /ɒ/, <o-e> and <ay>, where the comparison group seems to have had an advantage), indicating that the two groups appear to have achieved similar accuracy across the decoding of all the vowel graphemes. Statistical analysis was also conducted to compare the scores achieved in all the vowel GPCs by the two groups of participants (where each correctly decoded grapheme was scored as 1 and otherwise 0, following the grapheme-level scoring system as discussed in Chapter 4). A paired-samples t-test found no significant difference between the intervention and comparison group, $t(27) = -1.49$, $p = .15$, suggesting that the two groups in Universities A and B did not have any significant overall difference in the accuracy of decoding vowel graphemes at t1.

In contrast to the consonant GPCs, in most of which both groups of participants already achieved around 80% accuracy at t1, the vowel GPCs saw much lower

accuracy percentages. Participants achieved over 80% accuracy in only three out of the 28 vowel GPCs (<i> = /ɪ/, <ee> and <er>), while the accuracy for 9 vowel GPCs fell below 40%: <ir>, <e>= /e/, <au>, <ie>, <i-e>, <ou>, <ai>, <u> = /ə/ and <ey>.

One thing worth noting is that participants' decoding accuracy varied widely for some vowel graphemes with multiple phonemic realisations. For instance, both intervention and comparison groups achieved approximately 70% accuracy where grapheme <a> is decoded as /æ/, but only around 40% accuracy where <a> is decoded as /ə/. Similarly, both intervention and comparison groups achieved approximately 50% accuracy where the grapheme <u> is decoded as /ʌ/, but only 30% accuracy where it is decoded as /ə/. This indicates that many participants might not be aware that these graphemes have multiple phonemic realisations; hence they may have tended to decode these graphemes consistently, without considering the orthographic context. This will be further elaborated upon in section 6.2.2.

Another interesting observation is that, for vowel graphemes with single phonemic realisations, participants' accuracy showed wide variation between the decoding of different graphemes. For instance, both intervention and comparison groups achieved more than 80% accuracy in decoding <ee> and <er>, but only around 30% accuracy in decoding <ou> and <ai>. This will be further discussed section 8.3.2.

### 5.2.2. Mean accuracy percentage at t2

The mean accuracy for each of the vowel GPCs for Universities A and B combined at t2 is shown in Figure 5.5. The graphemes are presented in the same order as in Figure 5.4 to facilitate comparison.

*Figure 5.5 Universities A and B's mean accuracy of vowel GPCs at t2*

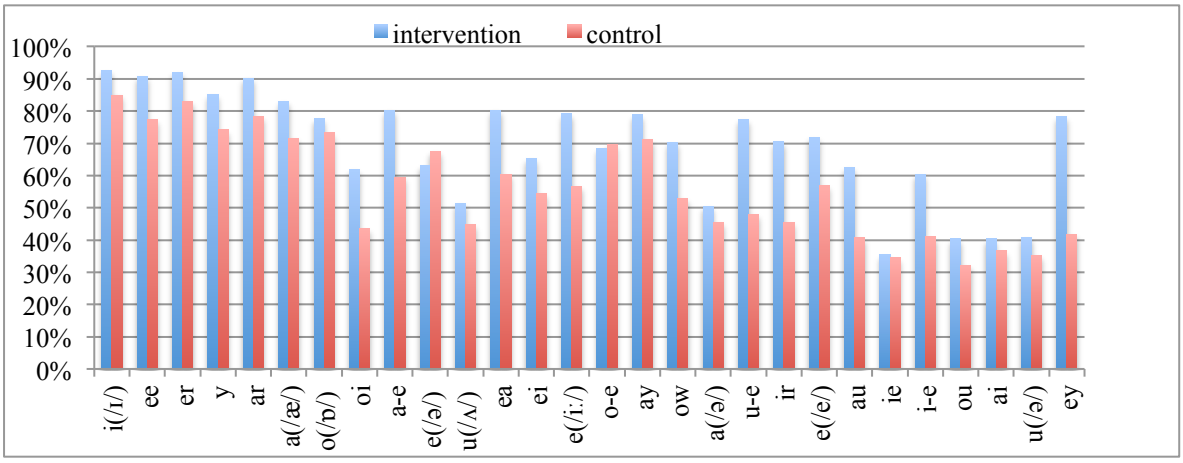



Figure 5.5 shows that, at a descriptive level, the intervention participants in Universities A and B outperformed the comparison groups in the decoding accuracy of almost all vowel graphemes at t2, with only two exceptions, <e> = /ə/ and <o-e> = /əʊ/. However, it should be noted that for these two GPCs, the baseline score was lower for the intervention participants than for the comparison groups, as can be seen in Figure 5.4. The two groups achieved similar accuracy in these two GPCs at t2, suggesting that the phonics instruction helped the intervention participants catch up with their comparison counterparts in these two GPCs. Taking this into consideration, there appears to have been a consistent effect of the intervention across all vowel

GPCs.

Statistical analysis was conducted to compare the scores achieved in all the vowel GPCs of the two groups of participants at t2. Significant differences were found between the two groups, as revealed by a paired-samples t-test, $t(27) = 5.14$, $p < .001$. This finding, together with the results presented in Figure 5.5, shows that the intervention group in Universities A and B significantly outperformed the comparison group in the decoding of the vowel graphemes at t2.

It can be seen that most of the vowel GPCs at the left-hand end of the table were still decoded more accurately than those on the right, indicating that the GPCs with low accuracy at t1 generally remained more difficult to decode even after the phonics instruction programme.

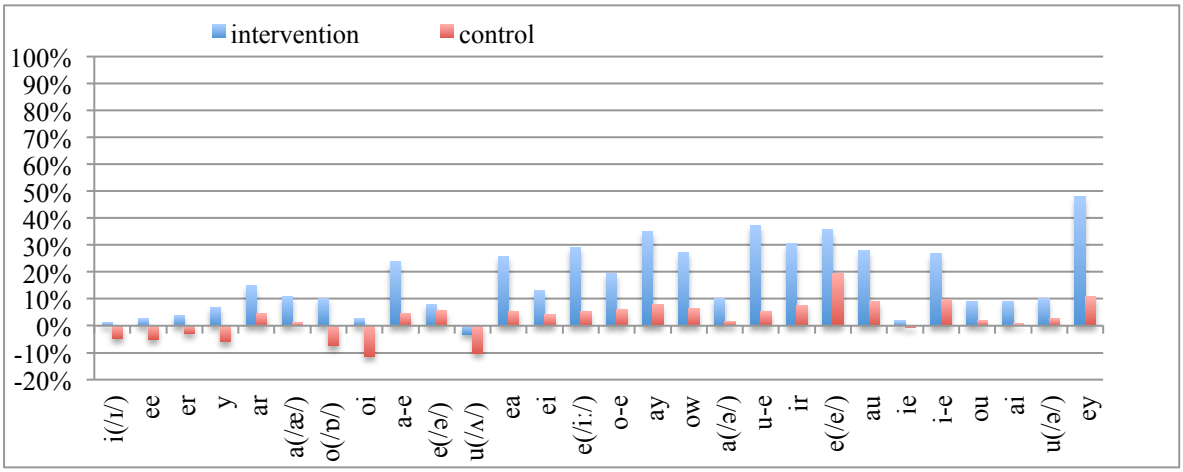
For the vowel graphemes with multiple phonemic realisations, some variations in terms of accuracy percentages can still be observed in various cases at t2, though the mean accuracy increased across all the different realisations. For instance, at t2, the intervention participants achieved approximately 80% accuracy when grapheme <a> is decoded as /æ/ and around 50% accuracy when it is decoded as /ə/, compared to 70% and 40% respectively at t1. However, the gap between the two phonemic realisations of grapheme <a> was still large, indicating that some GPCs remained difficult to master even after the phonics instruction programme. This will be further elaborated

upon in section 6.2.2.

### 5.2.3 Progress

The difference between the mean accuracy for each vowel GPC at t2 and t1 was calculated as a measure of progress. These progress scores are presented in Figure 5.6. The graphemes are presented in the same order as in Figures 5.4 and 5.5 for ease of comparison, that is, in descending order of accuracy of decoding at t1.

*Figure 5.6. Difference between mean accuracy of vowel GPCs at t2 and t1*




The progress made by the two groups of participants in accuracy of decoding vowel graphemes was also compared by statistical analysis, using a paired-samples t-test. Significant differences were found between the two groups, $t(27) = 6.07$, $p < .001$. This finding, together with the results presented in Figure 6.6, suggests that the intervention group made more progress than the comparison group in the accuracy of decoding vowel graphemes.

It can be seen that intervention participants in Universities A and B made progress in all the vowel GPCs except one (<u> = /ʌ/) (whose accuracy scores remained around 50%; this GPC was tested 8 times in the test, so this cannot be an artefact of a single test item). The progress varied widely across different vowel GPCs. For instance, the decoding of <ey> saw an increase in accuracy above 40 percentage points, while the decoding of graphemes such as <ai> and <ou> saw an increase in accuracy of approximately 10 percentage points. The comparison groups also made some small progress in most vowel GPCs, but the progress appears to have been much less substantial than that of the intervention participants. Moreover, some vowel GPCs were associated with a decrease in accuracy in several cases for the comparison groups, but only in one case for the intervention participants.

Similarly to the findings for the consonant GPCs, it can be observed here that the intervention participants made more progress in some vowel GPCs than others, despite having a similar baseline score. For instance, the decoding of the graphemes <a-e>, <e> = /ə/ and <u> = /ʌ/ all had approximately 55% accuracy at t1. After the instruction programme, intervention participants had made considerable progress in decoding <a-e> but little progress in decoding <e> = /ə/, while the accuracy of decoding <u> = /ʌ/ even showed a decrease. Likewise, though the graphemes <ie>, <i-e>, <ou>, <ai>, <u> = /ə/ and <ey> were all among the most poorly decoded graphemes at t1, with accuracy lower than 40%, the decoding of <ey> saw clear progress at t2 and the decoding of <i-e> witnessed moderate progress, while the

decoding of <ou>, <ai>, <u> = /ə/ only saw small progress and the decoding of <ie> remained roughly the same. This again provides an affirmative answer to Research Question 2. Possible reasons contributing to these findings will be further elaborated upon in section 8.3.2.

### 5.2.4 Summary

In summary, the analysis of the accuracy for vowel GPCs demonstrates that participants in Universities A and B had a good command of only a few vowel GPCs at t1. Only a quarter of the vowel GPCs achieved accuracy percentages above 60%, while many vowel GPCs appeared to pose great challenges. After the phonics instruction programme, the intervention participants had made clear progress with most vowel GPCs, with two thirds of the total vowel GPCs now achieving accuracy higher than 60%. Similarly to the consonant GPCs, some vowel GPCs showed more progress than others, despite having a similar baseline, again providing an affirmative answer to RQ2.

## 5.3 Summary

This chapter has examined the phonological decoding test results from a different perspective than Chapter 4, by analysing the accuracy percentages for individual GPCs before and after the instruction programme.

The results show that the both the intervention and comparison groups in Universities A and B already demonstrated good command of most consonant GPCs at t1. In contrast, most vowel GPCs had low accuracy at t1, suggesting that the vowel GPCs posed more challenge to the participants than the consonant GPCs.

After the phonics instruction programme, the intervention participants in Universities A and B demonstrated a clear advantage over the comparison groups in almost every GPC, again confirming the effectiveness of the decoding instruction programme. The intervention participants correctly decoded most of the consonant graphemes and many vowel graphemes at t2. There were also GPCs that had poor accuracy even after the phonics instruction, and this will be further analysed in the next chapter.

When comparing the difference in accuracy between the two time points, it was found that the intervention participants clearly made more progress than the comparison groups for most of the GPCs. It was also found that the phonics instruction programme was indeed more effective for some GPCs than others, even in cases where they were realised with similar accuracy at t1. This provides an affirmative answer to RQ2. Possible causes of this finding will be discussed in Chapter 8.

## Chapter 6. Findings III. Participants' Realisation of English Graphemes

The previous chapter discussed participants' strengths and weaknesses in English decoding by analysing the accuracy percentages of individual grapheme. It is also of interest to examine how participants actually pronounced the words in the decoding test, which complements the previous quantitative analysis and provides a more comprehensive picture of their English decoding. Therefore, this chapter pursues a qualitative line of analysis and addresses Research Question 3:

*RQ3: What are the features and problems of Chinese university EFL learners' English decoding at each time point?*

As the previous chapter has demonstrated, participants encountered more problems in the decoding of vowel graphemes, while most of the consonant graphemes already witnessed a high accuracy percentage at t1. As a result, the consonant graphemes would be treated more briefly in this chapter, covering only the main kinds of problems that occurred. As the vowel graphemes were particularly challenging for the participants, error analysis was conducted for each individual vowel grapheme. In addition, other errors beyond individual graphemes, namely whole-word errors and misordered graphemes, were also examined.

In line with the previous chapter, only the results of Universities A and B are

presented in this chapter. Nonetheless, the decoding output of participants in University C was also subjected to qualitative analysis and interested readers can find this analysis in Appendix 5. This analysis sheds interesting light on the influence of learning a third language (in this case, French) on L2 decoding and further justifies the removal of University C from the main quantitative analyses presented earlier.

This chapter is divided into four sections. Section 6.1 examines the common problems in the decoding of consonant graphemes. Section 6.2 presents participants' realisations of individual vowel graphemes. Section 6.3 looks at the other errors beyond individual graphemes (e.g. whole word errors; graphemes decoded in the wrong order). Section 6.4 summarises the findings and concludes this chapter.

## 6.1 Consonant graphemes

When examining participants' decoding at t1, three types of common errors were observed, namely (a) epenthesis, (b) omission and (c) approximation, which are presented in sections 6.1.1 to 6.1.3 respectively.

### 6.1.1 Epenthesis

One common error observed in the analysis of participants' decoding at t1 was epenthesis, defined as adding a phoneme or multiple phonemes that do not exist into

the pronunciation of a target word. This can be further divided into two categories.

The first type of epenthesis was found in an individual consonant digraph/ trigraph, namely breaking up an individual consonant grapheme into two graphemes by adding a phoneme. These consonant graphemes include <kn>, <wr>, <qu>, <wh> and <dge>. For instance, the grapheme <kn> in the word *knoink* and *knaf* was commonly pronounced as /kən/; the grapheme <wr> in the word *wrault* was pronounced as /wər/; the grapheme <qu> in the word *squow* and *quiles* was pronounced as /kju:/ or /kəʊ/; the grapheme <wh> in the word *whumb* was pronounced as /wəh/; and the grapheme <dge> in the word *cedge* was pronounced as /di:dʒ/. All the errors made by University A and B participants in this category and the error rate[29] before and after the phonics instruction are presented in Table 6.1.

[29] Error rate is defined as the number of errors divided by the total number of occurrences of the target word in the test.

*Table 6.1 Epenthesis in individual consonant graphemes by Universities A and B*

| Mispronounced Grapheme | Target word | Error | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <kn> | *knaf* | /kən/ | 39% | 4% | 44% | 30% |
| | | /ken/ | 0% | 0% | 3% | 0% |
| | *knoink* | /kən/ | 42% | 7% | 40% | 35% |
| | | /kin/ | 7% | 0% | 7% | 10% |
| <wr> | *wrault* | /wər/ | 11% | 0% | 15% | 8% |
| | *wrey* | /wer/ | 9% | 0% | 5% | 3% |
| | | /wər/ | 4% | 0% | 5% | 2% |
| <qu> | *squow* | /kju:/ | 47% | 11% | 39% | 30% |
| | | /kɔ:r/ | 11% | 4% | 15% | 19% |
| | *quiles* | /kju:/ | 44% | 14% | 34% | 18% |
| | | /kwɔ:/ | 13% | 4% | 20% | 10% |
| <wh> | *whumb* | /wəh/ | 7% | 0% | 5% | 0% |
| | | /wi:h/ | 4% | 0% | 0% | 0% |
| <dge> | *cedge* | /di:dʒ/ | 40% | 9% | 28% | 25% |
| | | /dedʒ/ | 33% | 9% | 27% | 33% |

It can be seen that this type of error was commonly observed in participants' decoding at t1, which is in line with the findings in section 5.1.1 as these graphemes (except for <wh>) were among the ones that saw the lowest mean accuracy before the instruction programmes. The results show that the most commonly inserted phoneme was schwa. Most inserted phonemes were vowels; only in the word *squow* was a consonant /r/ added by some participants. It should also be noted that this type of epenthesis was not found in all the consonant diagraphs/ trigraph in the decoding test: for instance, no epenthesis was found in the graphemes <sh> <ch> and <ng>. A possible reason for this finding is that these graphemes also exist in Pinyin and share similar pronunciations in English and Chinese, and this will be further discussed in section 8.4.1.

At t2, this type of epenthesis appeared to have clearly decreased in frequency for the intervention participants, suggesting the phonics instruction helped them curb the tendency to pronounce an individual consonant digraph/ trigraph as several separate graphemes. This was especially true for the graphemes <wr> and <wh>, as none of the intervention participants made epenthesis errors for these two graphemes at t2. However, this was not the case for the comparison groups. Though the comparison groups made fewer errors in most cases at t2, the differences between the two time points were not very large. When comparing the error rate of intervention and comparison groups with respect to epenthesis in an individual consonant digraph/ trigraph, it can be seen that the two groups made similar numbers of errors at t1, but the intervention participants appeared (on a descriptive level) to have made many fewer errors than the comparison groups at t2. This again suggests the effectiveness of the phonics instruction in combating this type of error.

The second type of epenthesis was found in consonant strings, namely a consecutive string of two or more consonant graphemes. For instance, the grapheme <tw> in the word *twem* was pronounced as /tɪw/ or /təw/; and <str> in the word *straced* was pronounced as /stər/. All the errors made by Universities A and B's participants in this category and the error rates at t1 and t2 are presented in Table 6.2.

*Table 6.2 Epenthesis in consonant strings by Universities A and B*

| Consonant string | Target word | Error | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <ft> | *bufty* | /fət/ | 9% | 3% | 8% | 5% |
| | | /fi:t/ | 3% | 0% | 5% | 0% |
| | | /fet/ | 0% | 0% | 2% | 0% |
| <tw> | *twem* | /təw/ | 8% | 4% | 9% | 4% |
| | | /tɪw/ | 4% | 0% | 0% | 0% |
| <str> | *straced* | /stər/ | 11% | 4% | 13% | 5% |
| <nk> | *knoink* | /nɪk/ | 16% | 5% | 12% | 5% |

The results show that compared to the epenthesis in an individual consonant grapheme, participants made fewer epenthesis errors in consonant strings at t1. All the added phonemes were vowels. Schwa was a commonly inserted phoneme, as well as the phoneme /ɪ/.

At t2, both intervention and comparison groups appeared (on a descriptive level) to have made fewer errors in consonant strings. It is also interesting to note that when comparing the error rates of the two groups with respect to epenthesis in consonant strings, the intervention and comparison groups made similar numbers of errors at both time points, suggesting that both the phonics instruction and the phonology instruction had a positive impact in terms of combating epenthesis in consonant strings.

### 6.1.2 Omission

Another commonly observed error type was omission of consonant graphemes,

defined as omitting some of the consonant graphemes when pronouncing a target word. This type of error was mostly observed in the decoding of long words with multiple syllables such as *monglustamer* and *untroikest*. For instance, the graphemes <g> and <s> in the word *monglustamer* were often omitted, as was the grapheme <s> in the word *untroikest*. All the omission errors made by participants in Universities A and B and the error rates at the two time points are summarised in Table 6.3. The omitted consonants are underlined in the table.

*Table 6.3 Omission of consonant graphemes by Universities A and B*

| Omitted Consonant | Target word | Illustrative example of errors | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <b> | *bafmot<u>b</u>em* | /bæfmɒtem/, /bæfmɒtəm/ /bæfməutem/, /bæfməutəm/ | 6% | 4% | 6% | 8% |
| <t> | *bafmo<u>t</u>bem* | /bæfmɒbem/, /bæfmɒbəm/ /bæfməubem/, /bæfməubəm/ | 3% | 1% | 15% | 5% |
| <f> | *ba<u>f</u>motbem* | /bæmɒtbem/, /bæmɒtbəm/ /bæməutbem/, /bæməutbəm/ | 8% | 0% | 11% | 3% |
| <g> | *mon<u>g</u>lustamer* | /mɒnlʌstəmə/, /mɒnlju:stəmə/ | 42% | 2% | 43% | 15% |
| | *ci<u>g</u>bet* | /sɪbet/, /sɪbɪt/ | 18% | 0% | 15% | 2% |
| <s> (/s/) | *monglu<u>s</u>tamer* | /mɒnglʌtəmə/, /mɒnglju:təmə/ | 31% | 0% | 27% | 8% |
| | *untroike<u>s</u>t* | /ʌntrɒɪkət/, /ʌntrɒɪkɪt/ | 28% | 2% | 35% | 15% |
| <s> (/z/) | *cei<u>s</u>minadolt* | /si:minədəʊlt/, /saɪminədəʊlt/ | 9% | 1% | 7% | 4% |
| <ng> | *manci<u>ng</u>ful* | /mansɪfəl/ | 44% | 11% | 39% | 20% |
| <l> | *pelnid<u>l</u>um* | /pelnɪdəm/ | 7% | 0% | 8% | 6% |

It can be seen that omission of consonant graphemes was commonly observed for both the intervention and comparison groups at t1. More specifically, the grapheme <g> in the word *monglustamer* and the grapheme <ng> in the word *mancingful* were the ones that were most frequently omitted, with nearly half of the participants not pronouncing them at t1. It is interesting to note that all the omitted consonant graphemes were in consonant strings, which again points to the difficulty in decoding consonant strings, as observed in the previous section. Moreover, when examining the position of the omitted consonant graphemes, it is clear that all the omitted consonants are at the start of a consonant string; in other words, all the omitted consonants are after a vowel, while other consonant graphemes before vowels were always pronounced. For instance, in the word *mancingful*, the grapheme <ng> in the consonant string <ngf> was frequently unpronounced, while the grapheme <f> was never omitted. It is also interesting to note that in the word *mancingful*, <nc> is also a consonant string but neither of the consonant graphemes was omitted by the participants. A possible explanation is that the graphemes <a> and <n> were processed together as <an>, which is a commonly seen letter combination in both Pinyin and English. This will be further discussed in section 8.4.1.2.

At t2, both the intervention and comparison groups made fewer errors of this type (on a descriptive level). However, the advantage of the intervention group appeared still to be clear: the error rate dropped to nearly zero except for the grapheme <ng> in the word *mancingful*. This suggests that the phonics instruction was effective in terms of

combating the omission of consonant graphemes in long words with multiple syllables.

### 6.1.3 Approximation

Another type of error was approximation, defined as wrongly pronouncing a consonant grapheme as another grapheme which resembles the target grapheme either in form or in pronunciation. For instance, the grapheme <v> in the word *vunhip* was frequently wrongly pronounced as /w/, which is an approximation of its pronunciation /v/ and also resembles the letter graphically.

All the errors of this type made by participants in Universities A and B were summarised in Table 6.4.

*Table 6.4 Approximation of consonant graphemes by Universities A and B*

| Mispronounced grapheme | Target word | Error | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <v> | *vunhip* | /wʌnhɪp/ | 15% | 2% | 13% | 3% |
| <th> | *throbe* | /srəʊb/ | 66% | 56% | 65% | 62% |
| <x> | *adjex* | /adjes/ | 45% | 34% | 50% | 40% |
| | *adjex* | /adjek/ | 24% | 23% | 23% | 29% |
| <wr> | *wrault* | /wɔːlt/ | 11% | 2% | 8% | 2% |

When comparing the error rate with respect to 'approximation' at t1 and t2, it can be seen that intervention participants made fewer errors in decoding graphemes <v> and <wr> after the instruction, but the graphemes <th> <x> and <mb> were still poorly

decoded. The comparison groups seemed to produce similar results, with the error rates for the graphemes <v> and <wr> dropping to almost zero at t2, while the error rates for <th> and <x> remained approximately the same as at t1. When comparing the results of the intervention and comparison groups with respect to 'approximation', it can be seen that both groups made similar numbers of errors at both t1 and t2, indicating that the phonics instruction was not more effective in promoting the accurate decoding of these graphemes than the phonology instruction received by comparison groups. Possible reasons for this finding will be discussed in section 8.4.1.3.

## 6.2 Vowel graphemes

Compared to consonant graphemes, both intervention and comparison groups in Universities A and B made more errors in the decoding of vowel graphemes. As different types of vowel graphemes showed different errors, this section divides the vowel graphemes into four categories, namely (a) split digraphs, (b) vowel graphemes with multiple possible realisations, (c) vowel digraphs that exist in Pinyin and (d) vowel digraphs that do not exist in Pinyin.

### 6.2.1 Split digraphs

At t1, the split digraphs, namely graphemes <u-e>, <i-e>, <o-e> and <a-e> (as

exemplified in the test items *rejune*, *depine*, *throbe* and *straced*) were frequently decoded wrongly. The errors made in decoding split digraphs by participants in Universities A and B, and the error rate at the two time points, are summarised in Table 6.5.

*Table 6.5 Errors in split digraphs by Universities A and B*

| Mispronounced grapheme | Target word | Illustrative example of errors | Intervention | | Comprison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <u-e> | *rejune* | /rɪdʒuːnə/, /redʒuːnə/ /redʒyːnə/ | 24% | 2% | 27% | 24% |
| | | /redʒuniː/ | 19% | 0% | 12% | 2% |
| | | /rɪdʒyːn/, /redʒyːn/ | 15% | 20% | 18% | 26% |
| <i-e> | *depine* | /dɪpaɪnə/, /dɪpɪnə/, /depɪnə/ | 32% | 5% | 30% | 21% |
| | | /dɪpɪniː/, /depɪniː/ | 10% | 0% | 9% | 2% |
| | | /dɪpɪn/, /dəpɪn/ | 24% | 34% | 29% | 36% |
| | *quiles* | /kwɒləs/, /kwɒləz/, /kwaɪləs/, /kɪləs/, /kwɪləs/, /kjuːləs/ | 36% | 4% | 22% | 18% |
| | | /kwɒliːs/, /kwɒliːz/, /kɪliːs/, /kwɪliːs/, /kjuliːs/ | 20% | 5% | 28% | 23% |
| | | /kjuls/, /kwɒls/, /kwiːlz/ | 10% | 30% | 18% | 19% |
| <o-e> | *throbe* | /θrəʊbə/, /srəʊbə/ | 5% | 0% | 3% | 0% |
| | | /θrəʊbiː/, /srəʊbiː/, /θrɒbiː/ /srɒbiː/ | 26% | 1% | 20% | 15% |
| | | /θrɔːb/, /srɔːb/ | 19% | 40% | 12% | 25% |
| <a-e> | *straced* | /stræsəd/, /stresəd/, /strʌsəd/ | 11% | 0% | 8% | 3% |
| | | /stræsiːd/ /stresiːd/ /strʌsiːd/ | 13% | 8% | 15% | 18% |
| | | /strʌst/, /stræst/ | 20% | 10% | 20% | 19% |

The results show that the errors in decoding split digraphs can be further divided into three types. The first type was breaking the split digraph into two graphemes and sounding out the silent <e> as /ə/. The second type was also breaking the split digraph into two graphemes but sounding out the silent <e> as /iː/. The third type consisted of 'wild forms' (Alison Porter, verbal presentation, 2015), namely those pronunciations that, whilst they correctly treated the split digraph as one single grapheme, were still

incorrect. It can be seen that at t1, the first two types of errors accounted for the majority of mistakes, suggesting that the participants had great difficulty in recognising split digraphs as one single grapheme. This was not surprising, as the grapheme <e> in final-word position is always pronounced in Pinyin. After the phonics instruction, the intervention participants made fewer errors in decoding split digraphs, with the error of sounding out the silent <e> almost being eliminated. However, the intervention participants produced more wild forms for three out of the four graphemes compared to t1. This suggests that even though the intervention participants did not have much problem in seeing split digraphs as one single grapheme, they were still trying to figure out the correct pronunciation. This will be further discussed in section 8.4.2.

In comparison, the comparison groups in Universities A and B still made similar numbers of errors in terms of sounding out the silent <e> in split digraphs at t2, though the error type was less varied compared to t1. For instance, 27% of the comparison groups pronounced the silent <e> in the word '*rejune*' as /ə/ at t1, and 12% sounded out the silent <e> as /i:/. At t2, 24% of them pronounced the silent <e> in this word as /ə/, but only 2% pronounced it as /i:/. The comparison groups also produced slightly more wild forms at t2, but not as many as the intervention participants did, suggesting that the phonology instruction programme may not be as useful as the phonics instruction programme in terms of promoting the decoding of spilt digraphs.

### 6.2.2 Vowel graphemes with multiple realisations

For the vowel graphemes with multiple possible realisations, namely <a> <e> <i> <o> and <u>, participants had great difficulty in determining their correct pronunciations in specific orthographic contexts. The errors made by participants in Universities A and B in decoding these graphemes at the two time points are presented in Tables 6.6. The mispronounced grapheme in the target word is underlined.

*Table 6.6 Errors in vowel graphemes with multiple realisations by Universities A and B*

| Mispronounced grapheme | Target word | Error | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <a> (/æ/) | *bab, knaf, bafmotbem, raff, dat, glack, darlanker, mancingful, adjex,* | /a:/ | 20% | 7% | 19% | 15% |
| | | /e/ | 2% | 5% | 3% | 6% |
| | | /eɪ/ | 7% | 5% | 7% | 8% |
| <a> (/ə/) | *byrcal, monglustamer, ceisminadolt* | /a:/ | 29% | 9% | 30% | 25% |
| | | /æ/ | 15% | 12% | 12% | 15% |
| | | /eɪ/ | 15% | 27% | 14% | 15% |
| <e> (/e/) | *twem, adjex, yeng, cedge, hend, sess, brecked, pelnidlum* | /i:/ | 28% | 9% | 31% | 24% |
| | | /æ/ | 25% | 21% | 31% | 16% |
| <e> (/i:/) | *rejune, depine* | /e/ | 40% | 20% | 40% | 40% |
| <e> (/ə/) | *untroikest* | /e/ | 25% | 19% | 25% | 13% |
| | | /i:/ | 20% | 16% | 20% | 18% |
| <u> (/ʌ/) | *shum, dud's, bufty, vunhip, whumb, ful's, untroikest* | /u/ | 30% | 21% | 35% | 31% |
| | | /y/ | 18% | 12% | 10% | 12% |
| | | /æ/ | 18% | 17% | 10% | 12% |
| <u> (/ə/) | *mancingful, pelnidlum* | /ʌ/ | 15% | 27% | 6% | 21% |
| | | /u/ | 55% | 33% | 60% | 44% |
| <o> (/ɒ/) | *op, monglustamer* | /æ/ | 30% | 22% | 20% | 28% |
| <o> (/əu/) | *ceisminadolt* | /ɒ/ | 40% | 14% | 45% | 36% |
| <i> (/ɪ/) | *bim, plip, vunhip, cigbet, pelnidlum* | /aɪ/ | 10% | 6% | 9% | 15% |

The results show that participants in Universities A and B had much difficulty in determining the correct pronunciation for the vowel graphemes with multiple possible realisations. For instance, the grapheme <a> in words *bab*, *knaf*, *bafmotbem*, *raff*, *dat*,

*glack*, *darlanker* and *mancingful* was wrongly decoded as either /a:/, /e/ or /eɪ/, all of which are possible realisations of this grapheme in other orthographic contexts (e.g. *after*, *many*, *rate*). However, when comparing the error rate of these three realisations, it can be seen that /a/ was the predominant error at t1 for both the intervention and comparison groups when the grapheme <a> should be decoded as /æ/. When <a> should be decoded as /ə/, the predominant error type was also /a:/. In other words, participants frequently decoded the grapheme <a> as /a:/ regardless of the orthographic context. Similarly, both intervention and comparison groups predominantly decoded the grapheme <e> as /e/, the grapheme <u> as /u/, and the grapheme <o> as /ɒ/ (i.e. according to their canonical pronunciations in monosyllabic words such as 'bed', 'nut' and 'hot'). Though /ɪ/ was only represented by <i> in the decoding test, many participant wrongly decoded the split digraph <i-e> in a way that included the sound /ɪ/ as well (see section 6.2.1). This indicates that the participants might wrongly perceive English GPCs as being as consistent, as is the case in Pinyin: i.e., they might think that a given grapheme always represents a particular sound, rather than its pronunciation varying according to its orthographic context, as in English. At t2, though the intervention participants in Universities A and B still had some difficulty in decoding these graphemes, they did not predominantly commit one particular error any more. Still taking the grapheme <a> (/æ/) as an example, the error rate for /a/ was 7% at t2, that for /e/ was 5% and /eɪ/ was 5%. This indicates their awareness of the multiple possible realisations for this grapheme, even though they still have not reached the correct pronunciation in all cases. In contrast, the error rate

for /a/ was 15%, /e/ was 6% and /eɪ/ was 8% for the comparison groups at t2, suggesting that they still tended to overgeneralise the realisation for this grapheme.

### 6.2.3 Vowel digraphs that exist in Pinyin

For vowel digraphs that also exist in Pinyin, participants predominantly decoded them in the same way as in Pinyin at t1. For instance, the grapheme <ei> was commonly decoded as /eɪ/, the grapheme <ie> as /ɪe/, the grapheme <ou> as /ou/ and the grapheme <ai> as /aɪ/. The errors in decoding these graphemes by participants in Universities A and B, and the error rates at each time point, are presented in Table 6.7.

*Table 6.7 Errors in vowel digraphs that exist in Pinyin by Universities A and B*

| Mispronounced grapheme | Target word | Illustrative example of errors | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <ei> | *ceisminadolt* | /ceɪzminədəʊlt/ | 42% | 29% | 45% | 40% |
| <ie> | *whie* | /wɪe/ | 36% | 36% | 34% | 30% |
| <ou> | *gouch* | /goutʃ/ | 70% | 60% | 70% | 67% |
| <ai> | *laip* | /laip/ | 68% | 59% | 63% | 63% |

It can be seen that the participants in Universities A and B consistently decoded these vowel digraphs in the same way as in Pinyin at t1. At t2, the rate of these Pinyin-resembling errors only saw a slight decrease (on a descriptive level) for the intervention participants, suggesting that the phonics instruction was not very effective in promoting the decoding of these vowel digraphs. Comparison groups made similar numbers of Pinyin-resembling errors at the two time points.

### 6.2.4 Vowel digraphs that do not exist in Pinyin

For vowel digraphs that do not exist in Pinyin, the error patterns were less clear. All the errors in this type and the error rate at each time point for participants in Universities A and B are shown in Table 6.8.

*Table 6.8 Errors in vowel digraphs that do not exist in Pinyin by Universities A and B*

| Mispronounced grapheme | Target word | Error | Intervention | | Comparison | |
|---|---|---|---|---|---|---|
| | | | t1 | t2 | t1 | t2 |
| <ee> | *dee*, *yee*, *ree* | /e/ | 10% | 7% | 17% | 22% |
| <er> | *monglustamer, darlanker* | /i:/ | 11% | 7% | 14% | 16% |
| <y> | *bufty*, *byrcal* | /aɪ/ | 21% | 14% | 20% | 25% |
| <ar> | *darlanker* | /eə/ | 24% | 9% | 26% | 21% |
| <oi> | *knoink, untroikest* | /ɒ/ | 40% | 38% | 45% | 56% |
| <ea> | *weaf*, *weat* | /e/ | 55% | 19% | 55% | 39% |
| <ay> | *tay*, *tayed* | /ai/ | 56% | 20% | 35% | 28% |
| <ow> | *squow* | /au/ | 20% | 15% | 25% | 17% |
| | | /ɔ:/ | 36% | 15% | 29% | 30% |
| <ir> | *zirdn't* | /iə/ | 40% | 20% | 40% | 30% |
| | | /i:/ | 19% | 8% | 22% | 24% |
| <au> | *wrault* | /au/ | 35% | 21% | 44% | 35% |
| | | /əu/ | 30% | 16% | 24% | 24% |
| <ey> | *wrey* | /i:/ | 50% | 21% | 48% | 40% |
| | | /e/ | 19% | 0% | 20% | 18% |

Comparing the error rate at each time point, it can be seen that the error rate for intervention participants in Universities A and B saw a clear decrease (on a descriptive level) for all these vowel graphemes at t2, suggesting that the phonics instruction was effective in promoting the knowledge of these graphemes. In contrast, the error rate for the comparison groups saw a slight increase for some graphemes, while remaining roughly the same for the others.

Though the error types were less clear for these vowel digraphs, some interesting observations can be made. For instance, the graphemes <ee> <ea> and <ey> were all wrongly decoded as /e/, which is the most common decoding of the grapheme <e>. Similarly, <oi> was frequently decoded as /ɒ/ and <ow> was decoded as /ɔ:/, both of which are possible realisations of the grapheme <o>. As a result, there is a chance that the participants pronounced one letter of the vowel digraph when they did not know its exact pronunciation.

## 6.3 Other errors

The previous two sections have demonstrated participants' realisations of different graphemes at the two time points. However, some errors beyond individual graphemes were also noticed, namely (a) whole-word errors and (b) misordered graphemes.

### 6.3.1 Whole-word errors

As all the target words in the decoding test were pseudo words, 'whole-word error' refers to the situation in which participants wrongly pronounced a pseudo word as a real word with similar orthography. It should be noted that some errors in this category also fit into other error categories as discussed in this chapter. For instance, pronouncing *dat* as *date* can also be considered as a mispronunciation of the vowel

grapheme <a>. However, when the product of decoding was a legitimate and high-frequency word (of which participants were highly likely to have prior knowledge), such as *date* and *cry*, these were considered as whole-word errors other than errors in the other categories. All the whole-word errors made by participants in Universities A and B are summarised Table 6.9.

*Table 6.9 Whole-word errors by Universities A and B*

| Target word | Illustrative examples of errors | Intervention | | Comparison | |
|---|---|---|---|---|---|
| | | t1 | t2 | t1 | t2 |
| *bab* | bad | 10% | 1% | 12% | 7% |
| *whie* | well, wine, while, where, wheel | 35% | 7% | 29% | 22% |
| *twem* | twin, team | 14% | 3% | 17% | 19% |
| *yeng* | year, young | 21% | 0% | 18% | 9% |
| *wrey* | worry, when, well, wine, white, wall, way, very | 42% | 7% | 49% | 37% |
| *knoink* | knock, kick | 41% | 21% | 37% | 24% |
| *wrault* | worried, result, wallet, wreck | 17% | 4% | 21% | 15% |
| *dat* | date | 21% | 8% | 27% | 15% |
| *cyr* | cry | 28% | 0% | 24% | 18% |
| *quiles* | quiz | 0% | 0% | 4% | 4% |
| *weat* | water, wake | 0% | 0% | 7% | 0% |

The results show that whole-word errors range from changing only one letter of the target word, such as pronouncing *dat* as *date*, to totally different spellings, such as pronouncing *wrault* as *result*. It can be seen that the intervention participants appeared to have made clearly fewer whole-word errors at t2, while the comparison groups appeared to have made roughly the same amount of whole-word errors.

### 6.3.2 Misordered graphemes

Another type of error was misordered graphemes, meaning the order of the graphemes in a target word was wrong in participants' decoding. These errors made by participants in Universities A and B in this category are summarised in Table 6.10.

*Table 6.10 Misordered graphemes by Universities A and B*

| Target word | Illustrative examples of errors[30] | Intervention | | Comparison | |
|---|---|---|---|---|---|
| | | t1 | t2 | t1 | t2 |
| *knoick* | knocki (/nɒki:/, /kənɒki:/) | 14% | 0% | 12% | 6% |
| *bufty* | bfuty (/bfjuti:/, /bəfuti:/) | 11% | 3% | 15% | 12% |
| *mancingful* | mangcinful (/mangsınfəl/) | 7% | 0% | 4% | 4% |
| *byrcal* | brycal (/bri:kəl/, /braıkəl/) | 14% | 0% | 11% | 6% |
| *pelnidlum* | pelindlum (/pelındləm/) | 7% | 0% | 4% | 4% |

The results show that many of these errors are associated with consonant strings. For instance, the word *pelnidlum* was pronounced as /pelındləm/, perhaps because of the problem with pronouncing the two consonants /l/ and /n/ together; similarly, the word *bufty* was decoded as /bəfjuti:/, perhaps because of the difficulty in the consonant string <ft>. At t2, this type of error appeared to have been almost eradicated for the intervention participants, while the comparison groups appeared to have made similar numbers of errors as t1.

## 6.4 Summary

This chapter examines the features and problems of participants' decoding before and

[30] The spellings in this column are the 'reconstruction' of the written forms of participants' decoding

after the instruction programmes. More specifically, participants' decoding was examined from three perspectives, namely the realisation of consonant graphemes, the realisation of vowel graphemes and other errors. It should be noted that all the patterns identified in this chapter are purely descriptive and have not been subjected to statistical analysis.

The examination of participants' decoding of consonant graphemes at t1 revealed three types of errors, namely epenthesis, omission and approximation. Epenthesis was found in both individual consonant grapheme and consonant strings, where a vowel, mostly schwa, was added in between the two consonants. At t2, both the intervention and comparison groups in Universities A and B made fewer errors in decoding consonant strings, but the intervention group appeared to have made many fewer errors than the comparison group in terms of epenthesis in individual consonant graphemes. Omission was observed in long words with multiple syllables at t1, where the first letter of consonant strings was frequently omitted. At t2, both the intervention and comparison groups in Universities A and B made fewer errors of this type, but the advantage of the intervention group still appeared to be clear. Some consonant graphemes were mispronounced as similar-sounding or a similar-looking graphemes at t1, and were still poorly decoded by the intervention group even after the phonics instruction.

The picture for the vowel graphemes was more complicated. The examination of

participants' decoding of vowel graphemes revealed four types of errors. The first type of error was found in split digraphs, where the silent <e> was frequently sounded out; the second type of error was found in graphemes with multiple possible realisations, where the decoding of the graphemes was frequently overgeneralised regardless of the orthographic contexts. The third type of error was Pinyin-resembling pronunciations for the vowel digraphs that also exist in Pinyin. The last type of error was wild forms for vowel digraphs that do not exist in Pinyin. The results show that the intervention participants in Universities A and B appeared to have decoded all the vowel graphemes more accurately at t2, except for the vowel digraphs that also exist in Pinyin.

In addition, some whole-word errors and misordered graphemes were observed at t1. At t2, both these types of errors were almost eliminated in the intervention participants in Universities A and B, suggesting that the phonics instruction was effective in terms of promoting intraword analysis.

## Chapter 7. Findings IV: Vocabulary Memorisation Test Results

The previous three chapters examined the results of the phonological decoding test. This chapter presents the results of the vocabulary memorisation task. Intervention and comparison groups' scores in the four vocabulary recall and recognition tests were compared in order to examine whether the phonics instruction programme led to better performance in vocabulary memorisation. As a result, this chapter addresses Research Question 4:

> *RQ4: Do participants who have followed the programme of phonics instruction also show gains in their ability to recall (productive knowledge) and recognise (receptive knowledge) new English words?*

As demonstrated in Chapter 4, the intervention participants in University C did not demonstrate significantly greater gains in the accuracy of decoding at t2. However, in experimental trials in education (as in other fields of inquiry), it is recommended that the primary analysis always be based on 'Intention to Treat' (Torgerson & Torgerson, 2013). In other words, all participants who were originally allocated to the different experimental groups (i.e. who were 'intended to be treated') should be included in the analysis, rather than excluding those who (for whatever reason) did not receive the programme in its intended form. Therefore, this chapter will firstly examine the aggregated recall and recognition test results of all three universities together (the full

original sample). However, subsequently, there will also be a separate analysis of (a) the aggregated results of Universities A and B and (b) the results of University C, in recognition of the fact that the phonics intervention may not have functioned as intended in University C due to the students' concurrent learning of the French writing system.

This chapter is divided into five sections. Section 7.1 presents the results of the oral recall test, where the participants saw the Chinese translations and were asked to pronounce the English words. Section 7.2 presents the results of the written recall test, where the participants saw the Chinese translations and were asked to spell the English words. An analysis of participants' errors in the written recall test is also presented. Section 7.3 presents the results of the aural recognition test, where the participants heard the English pronunciations and were asked to write the Chinese translations. Section 7.4 presents the results of the written recognition test, where the participants saw the English words and were asked to write the Chinese translations. In this way, both receptive and productive knowledge of the words in the vocabulary memorisation task were assessed. Section 7.5 summarises the findings of this chapter. Table 7.1 presents a brief summary of the four recall and recognition tests.

*Table 7.1 Summary of the four recall and recognition tests*

| Order | Test form | Task | Requirement | Type of knowledge assessed |
|---|---|---|---|---|
| 1 | Oral recall | See Chinese | Say English | Productive |
| 2 | Written recall | See Chinese | Write English | Productive |
| 3 | Aural recognition | Hear English | Write Chinese | Receptive |
| 4 | Written recognition | See English | Write Chinese | Receptive |

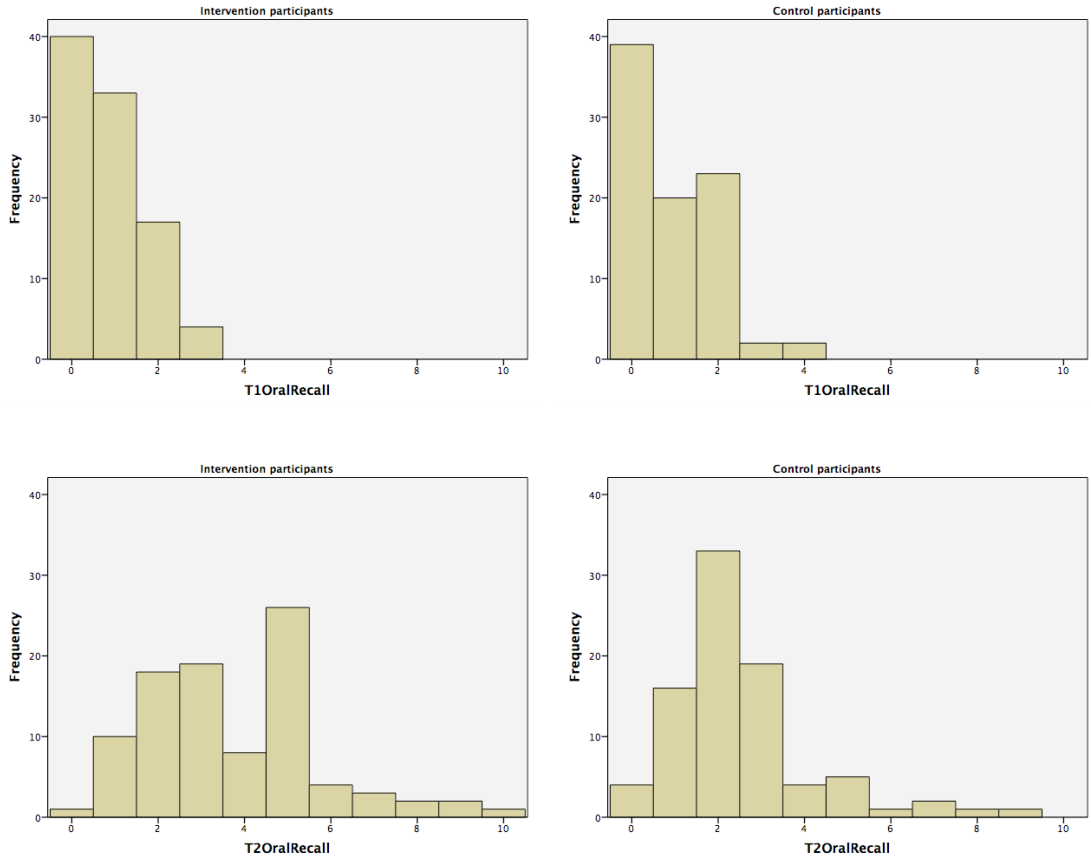
## 7.1 Oral recall test results (See Chinese, Say English)

As discussed in Chapter 4, the number of correctly recognised words was counted as the score on the oral recall test. The descriptive statistics for the oral recall test results of all three universities taken together are summarised in Table 7.2, and the histograms are presented in Figure 7.2.

*Table 7.2 Oral recall scores of all three Universities (out of 10)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 1 | 0.8 | 0.7 | 0 | 3 | 3 | 3.8 | 2.0 | 0 | 10 |
| **Comparison (N=86)** | 1 | 0.9 | 0.7 | 0 | 4 | 2 | 2.5 | 1.7 | 0 | 9 |

*Figure 7.1 Histograms of oral recall scores (out of 10) of all three universities*



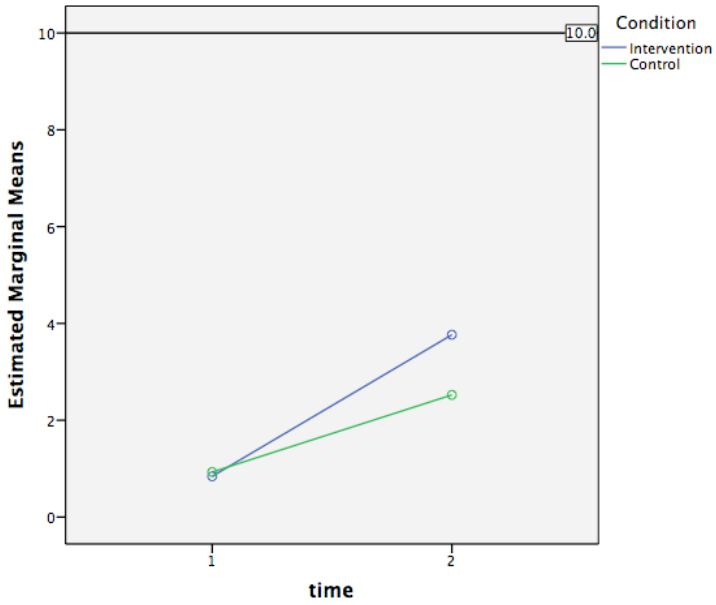

The assumptions of a two-way factorial ANOVA with one within-subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated, as the z-scores skewness and kurtosis were both larger than 1.96 (Field, 2005: 139)[31]. Secondly, the issue of sphericity was not considered here as the repeated measures had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was violated at both t1 and t2[32]. As a result, the data was primarily analysed using non-parametric tests. However, parametric tests were also conducted as a means of comparison.

[31]t1 *z*-score of skewness = 4.77, *z*-score of kurtosis = 1.02; t2 *z*-score of skewness = 5.51, *z*-score of kurtosis = 2.61
[32]t1 $F(1, 178) = 9.09$, $p < .01$, t2 $F(1, 178) = 12.20$, $p < .01$

Firstly, a Wilcoxon Signed Rank test was conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for both the intervention participants ($Z$= -7.98, $p < .001$, $r = .59$) and the comparison groups ($Z$= -6.41, $p < .001$, $r = .48$). As the mean oral recall scores were higher at t2 compared to t1 for both groups, this suggests that both the intervention and comparison groups in all three universities made significant progress in the oral recall test. Then a Mann-Whitney test was conducted to examine the differences between the two groups' oral recall scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U = 3927.5$, $Z = -.35$, $p = .73$) but were significantly different at t2, with a medium effect size ($U = 2483.5$, $Z = -4.56$, $p < .001$, $r = .34$).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 178) = 215.54$, $p < .001$, $r = .74$, as well as a significant main effect of condition with a small effect size, $F(1, 178) = 13.35$, $p < .001$, $r = .26$. The interaction between time and condition was also significant, $F(1, 178) = 39.82$, $p < .001$, with a medium effect size, $r = .31$. The interaction graph is presented in Figure 7.2. These results support the findings of the non-parametric tests, indicating that the phonics instruction led to significantly greater progress in terms of oral recall test results compared to the phonology instruction.

*Figure 7.2 Estimated marginal means of oral recall test scores (out of 10) of all three universities*

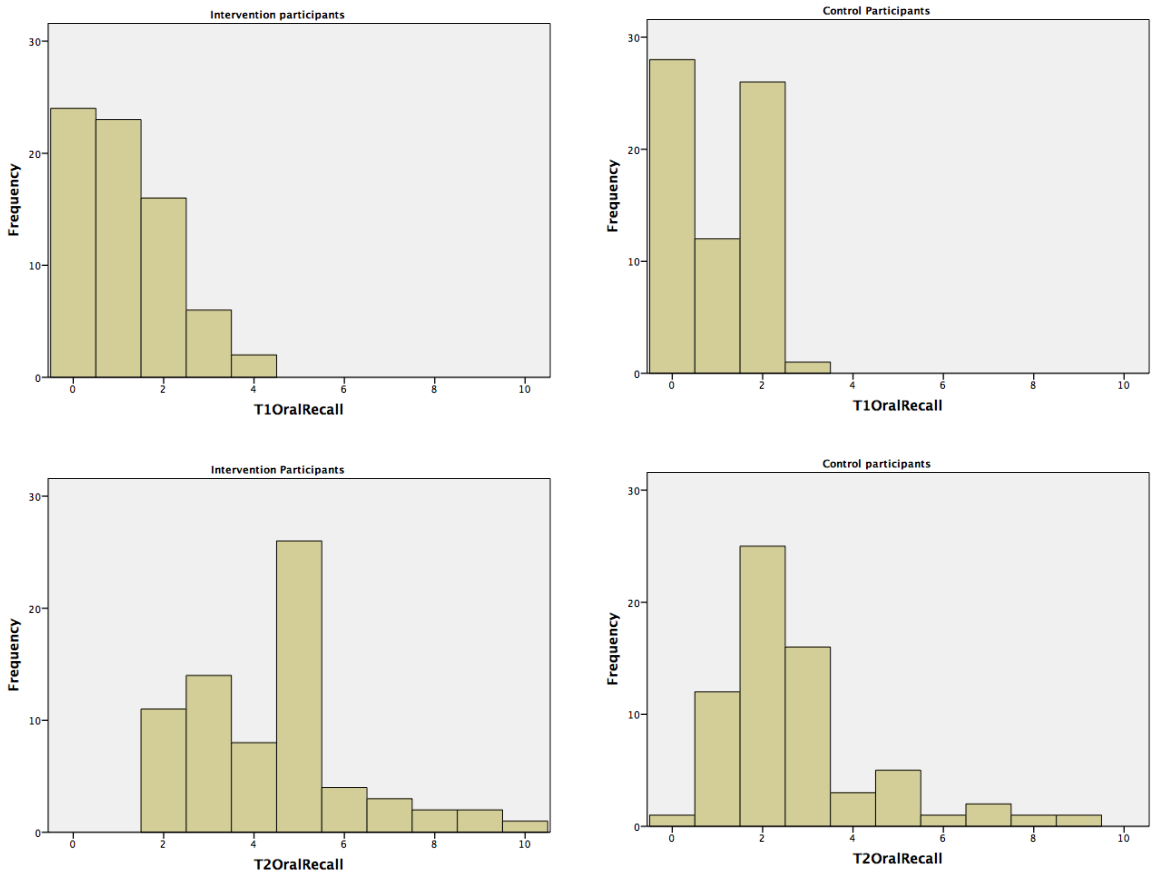



The aggregated oral recall test results of Universities A and B were then analysed separately. The descriptive statistics of the oral recall test results of Universities A and B are summarised in Table 7.3, and the histograms are presented in Figure 7.3.

*Table 7.3 Oral recall scores of Universities A and B (out of 10)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 1 | 1.0 | 0.9 | 0 | 4 | 5 | 4.2 | 1.5 | 2 | 10 |
| **Comparison (N=67)** | 1 | 1.0 | 0.9 | 0 | 3 | 2 | 2.5 | 1.4 | 0 | 9 |

*Figure 7.3 Histograms of oral recall scores (out of 10) of Universities A and B*



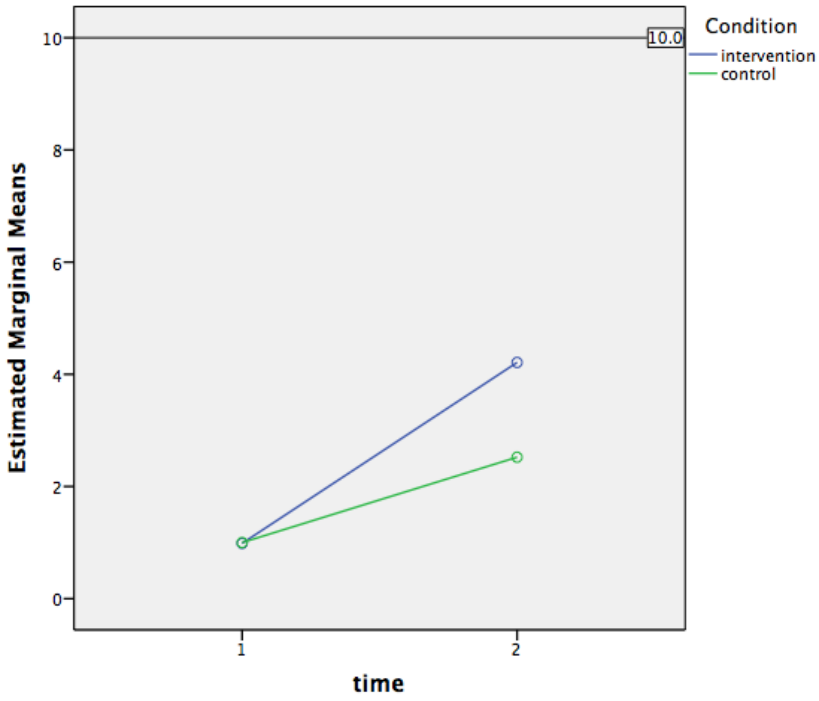

The assumptions of a two-way factorial ANOVA with one within-subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[33]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at both t1 and t2[34]. As with the analysis of the oral recall test results of all three universities, both non-parametric and parametric tests were conducted as a point of comparison. However, in view of the assumption violations, the results of the parametric tests should be treated with caution.

[33] t1 *z*-score of skewness = 2.52, *z*-score of kurtosis = -.47; t2 *z*-score of skewness = .83, *z*-score of kurtosis= -2.21
[34] t1 $F(1,136) = 3.48$, $p = .06$ ; t2 $F(1,136) = 1.80$, $p = .18$

A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -6.59, $p$ < .001, $r$ = -.55) but not the comparison group ($Z$ = -1.70, $p$ = .09). As the mean oral recall scores were higher at t2 compared to t1 for both groups, this suggests that only the intervention participants in Universities A and B made significant gains after the phonics instruction. Then, a Mann-Whitney test was also conducted to examine the differences between the two groups' oral recall scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U$ = 2323, $Z$ = -.25, $p$ = .80) but were significantly different at t2, with a medium effect size ($U$ = 880.5, $Z$ = - 6.52, $p$ < .001, $r$ = .39).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F$(1, 136) = 268.82, $p$ < .001, $r$ = .81, as well as a significant main effect of condition with a medium effect size, $F$(1, 136) = 48.35, $p$ < .001, $r$ = .44. The interaction between time and condition was also significant, $F$(1, 136) = 34.59, $p$ < .001, with a medium effect size, $r$ = .45. The interaction graph is presented in Figure 7.4. These results support the findings of the non-parametric tests, indicating that the phonics instruction led to significantly greater progress in terms of oral recall test results compared to the phonology instruction in Universities A and B.

*Figure 7.4. Estimated marginal means of oral recall test scores (out of 10) of Universities A and B*

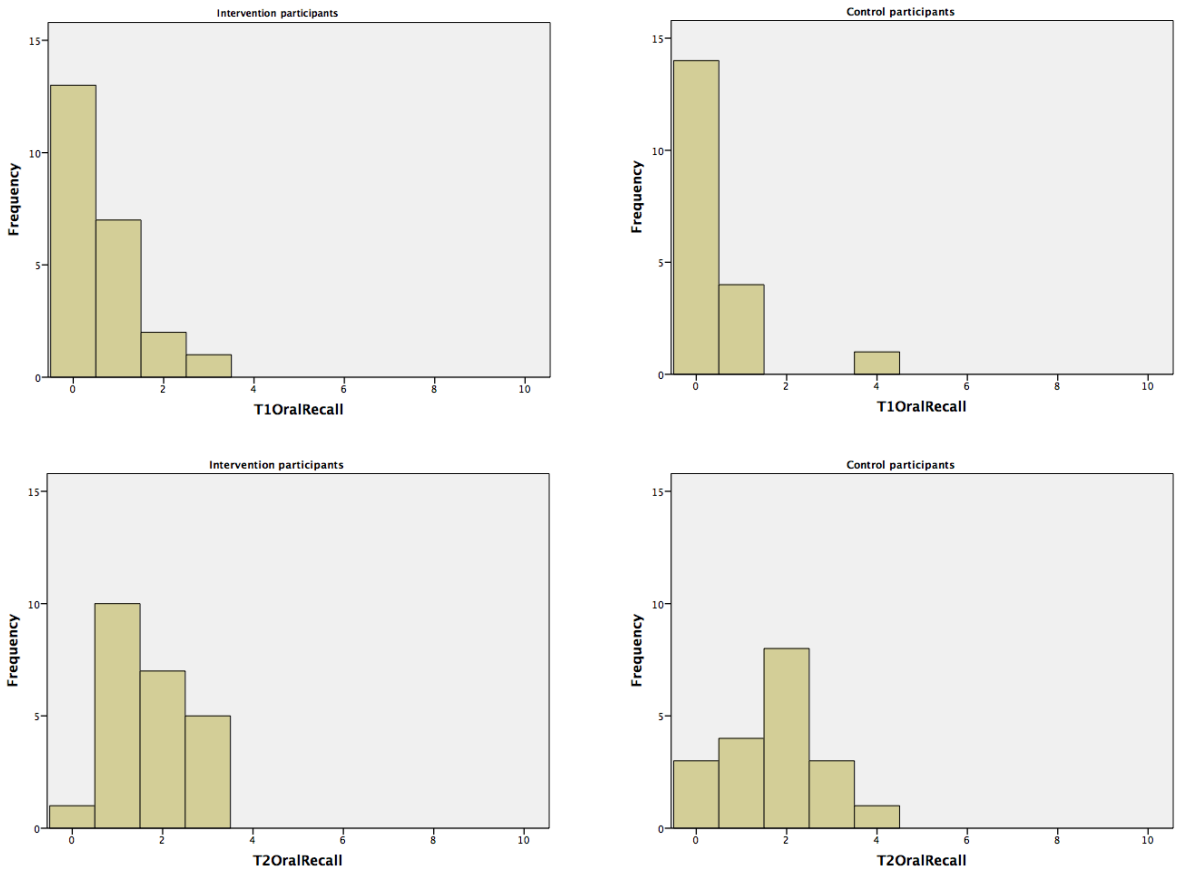



Finally, the oral recall test results of participants in University C alone were analysed. The descriptive statistics of the oral recall test results of Universities C are summarised in Table 7.4, and the histograms are presented in Figure 7.5.

*Table 7.4. Oral recall scores of University C (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=23)** | 0 | 0.6 | 0.8 | 0 | 3 | 2 | 1.7 | 0.9 | 0 | 3 |
| **Comparison (N=19)** | 0 | 0.4 | 1.0 | 0 | 4 | 2 | 1.7 | 1.1 | 0 | 4 |

*Figure 7.5 Histograms of oral recall scores (out of 10) of Universities C*



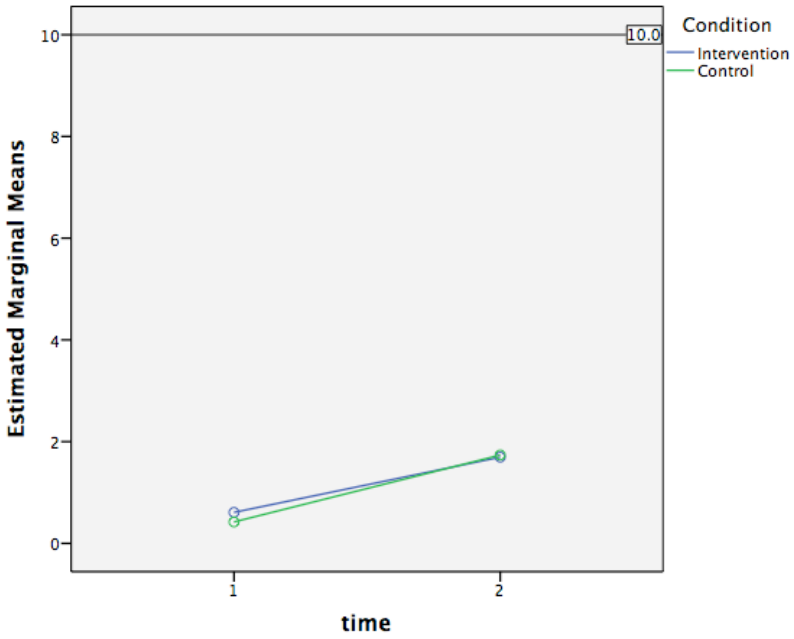

The assumptions of a two-way factorial ANOVA with one within subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[35]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at both t1 and t2[36]. Following the previous analyses, both non-parametric and parametric tests were conducted as a point of comparison.

[35] t1 *z*-score of skewness= 2.90, *z*-score of kurtosis= 1.74; t2 *z*-score of skewness= .47, *z*-score of kurtosis= -2.31
[36] t1 $F(1,40)= .14$, $p= .71$ ; t2 $F(1,136)= .41$, $p= .53$

A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -3.20, $p$ < .01, $r$ = -.49) as well as the comparison group ($Z$ = -2.92, $p$ < .01, $r$ = -.45). As the mean oral recall scores were higher at t2 compared to t1 for both groups, this suggests that both the intervention participants and the comparison groups in University C made significant gains after the instruction programmes. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' oral recall scores at each time point. It was found that the scores of the two groups were neither significantly different at t1 ($U$ = 180, $Z$ = -1.15, $p$ = .25) nor significantly different at t2 ($U$ = 211, $Z$ = -.20, $p$ = .84).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 40) = 30.03$, $p < .001$, $r = .68$. However, there was no significant main effect of condition, $F(1, 40) = .13$, $p = .73$. The interaction between time and condition was also non-significant, $F(1, 40) = .31$, $p = .58$. The interaction graph is presented in Figure 7.6. These results support the findings of the non-parametric tests, suggesting that, in University C, the phonics instruction did not lead to significantly greater gains in terms of oral recall test results compared to the phonology instruction.

*Figure 7.6. Estimated marginal means of oral recall test scores (out of 10) of University C*




In summary, analysis of the oral recall test results shows that the intervention participants as a whole (i.e. in all three universities taken together) made significantly greater gains compared to the comparison groups in the oral recall test, corresponding to the hypothesis that the phonics instruction programme was effective in facilitating the learning recall of phonological forms in vocabulary learning. The analysis of the aggregated results of participants in Universities A and B only also revealed similar results. However, the analysis of results of participants in University C suggests that the intervention participants in University C did not demonstrate significantly greater gains compared to the comparison groups in the same university. Recall that the intervention participants in University C did not make significantly greater gains in phonological decoding than their comparison counterparts either (presumably because of the disruptive effects of learning the French writing system concurrently).

Therefore, this finding in University C is still consistent with the hypothesis concerning the effects of improved decoding proficiency on vocabulary learning: in University C, the phonics instruction was less effective, there were no significant improvements in participants' decoding proficiency, and as a result there was also no improvement in their vocabulary learning (as measured by the oral recall test).

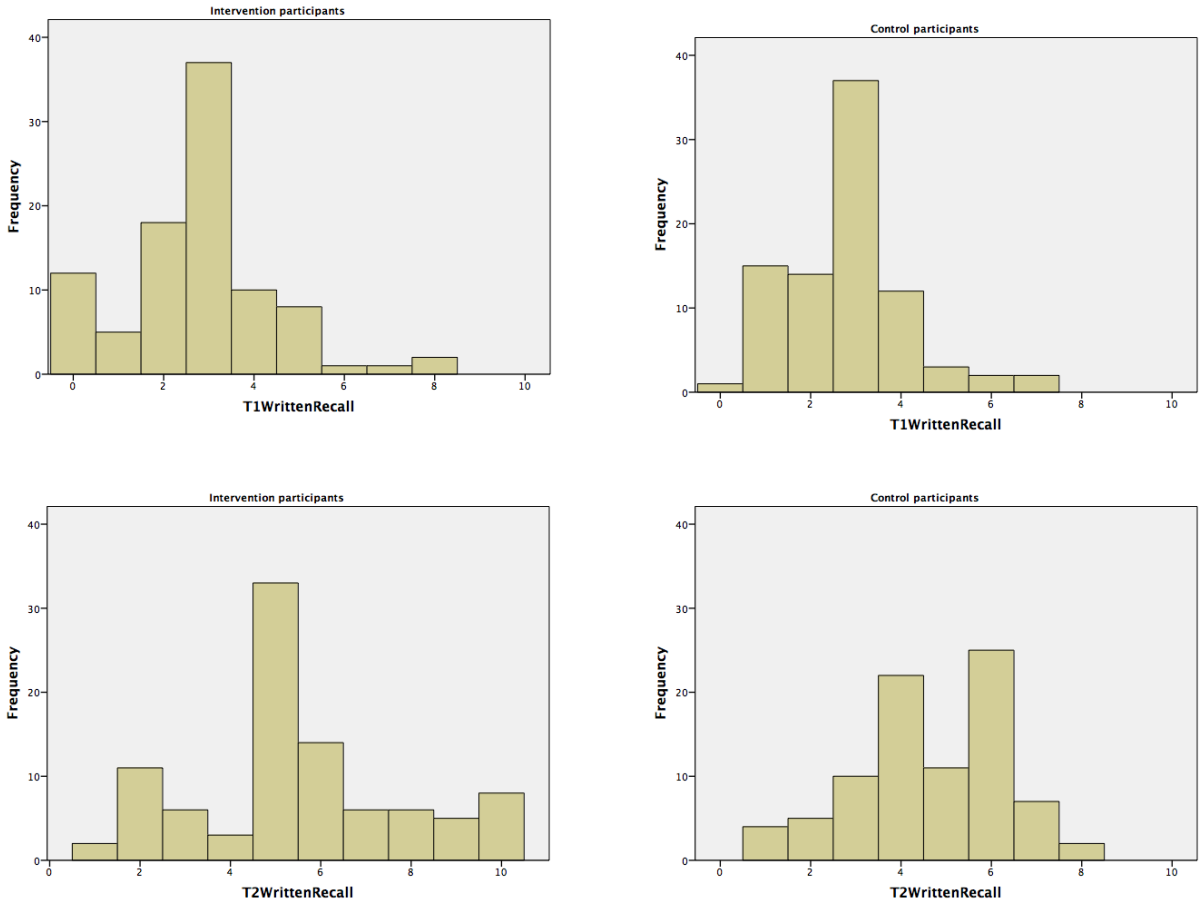
## 7.2 Written recall test results (See Chinese, Write English)

The number of correctly recalled words was counted as the score of the written recall test. The descriptive statistics of the aggregated written recall test results for all three universities are summarised in Table 7.5, and the histograms are presented in Figure 7.7.

*Table 7.5 Written recall scores of all three Universities (out of 10)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 3 | 2.8 | 1.7 | 0 | 8 | 5 | 5.5 | 2.3 | 1 | 10 |
| **Comparison (N=86)** | 3 | 2.8 | 1.4 | 0 | 7 | 5 | 4.7 | 1.7 | 1 | 8 |

*Figure 7.7 Histograms of oral recall scores (out of 10) of all three universities*



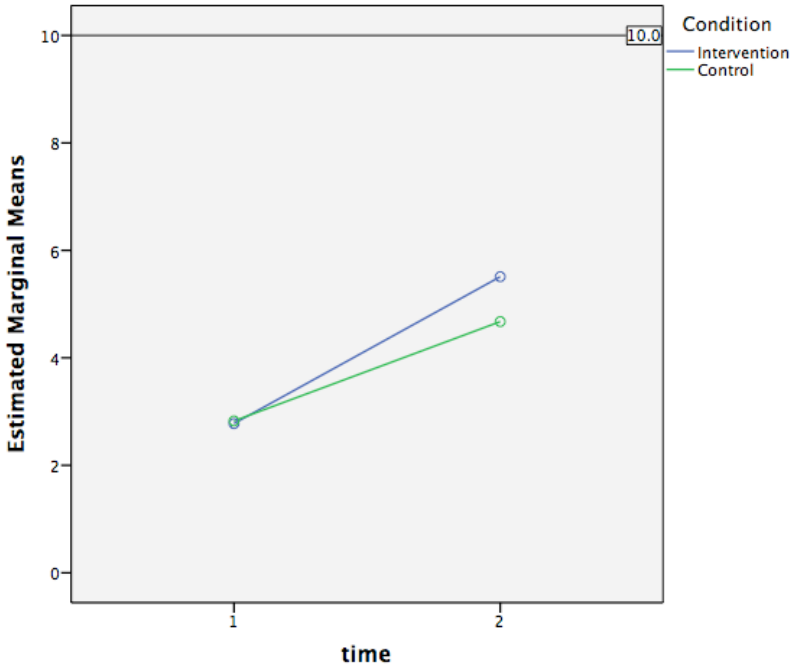

The assumptions of a two-way factorial ANOVA with one within-subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[37]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at t1 but violated at t2[38]. As a result, the data was primarily analysed using non-parametric tests. However, parametric tests were also conducted as a means of comparison.

[37] t1 *z*-score of skewness = 3.02, *z*-score of kurtosis = 3.69; t2 *z*-score of skewness = 1.59, *z*-score of kurtosis = .23
[38] t1 $F(1,178) = 2.03$, $p = .16$ ; t2 $F(1,178) = 4.75$, $p < .05$

Firstly, a Wilcoxon Signed Rank test was conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for both the intervention participants ($Z$ = -7.74, $p$ < .001, $r$ = -.58) and the comparison groups ($Z$ = -6.27, $p$ < .001, $r$ = -.47). As the mean oral recall scores were higher at t2 compared to t1 for both groups, this indicates that both the intervention and comparison groups in all three universities made significant progress in the written recall test. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' written recall scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U$ = 3971, $Z$ = -.21, $p$ = .83) but were significantly different at t2, with a small effect size ($U$ = 3286, $Z$ = -2.19, $p$ < .05, $r$ = .16).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 178) = 195.37$, $p$ < .001, $r$ = .72. The effect of condition was non-significant, $F(1, 178) = 3.43$, $p$ = .07. The interaction between time and condition was significant, $F(1, 178) = 7.29$, $p$ < .005, with a small effect size, $r$ = .20. The interaction graph is presented in Figure 7.8. These results support the findings of the non-parametric tests, indicating that the phonics instruction led to significantly greater progress in terms of the written recall test results than the phonology instruction. This will be further discussed in section 8.5.1.

*Figure 7.8. Estimated marginal means of written recall test scores (out of 10) of all three universities*

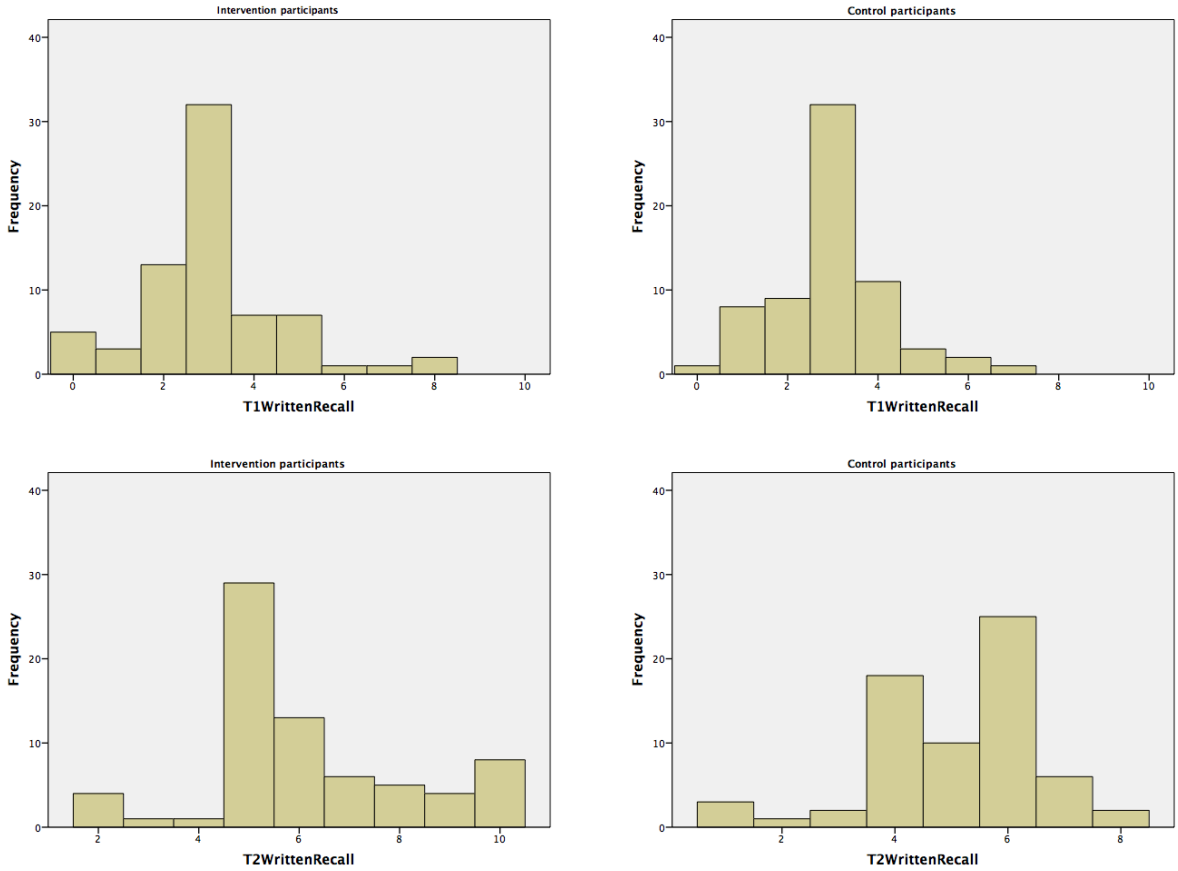



The aggregated written recall test results of Universities A and B were then analysed separately. The descriptive statistics of the written recall test results of Universities A and B are summarised in Table 7.6, and the histograms are presented in Figure 7.9.

*Table 7.6 Written recall scores of Universities A and B (out of 10)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 3 | 3.1 | 1.6 | 0 | 8 | 6 | 6.1 | 2.1 | 2 | 10 |
| **Comparison (N=67)** | 3 | 3.0 | 1.6 | 0 | 7 | 5 | 5.1 | 2.3 | 1 | 8 |

*Figure 7.9 Histograms of written recall scores (out of 10) of Universities A and B*




The assumptions of a two-way factorial ANOVA with one within-subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[39]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at both t1 but rejected at t2[40]. Following the previous analysis, both non-parametric and parametric tests were conducted as a point of comparison.

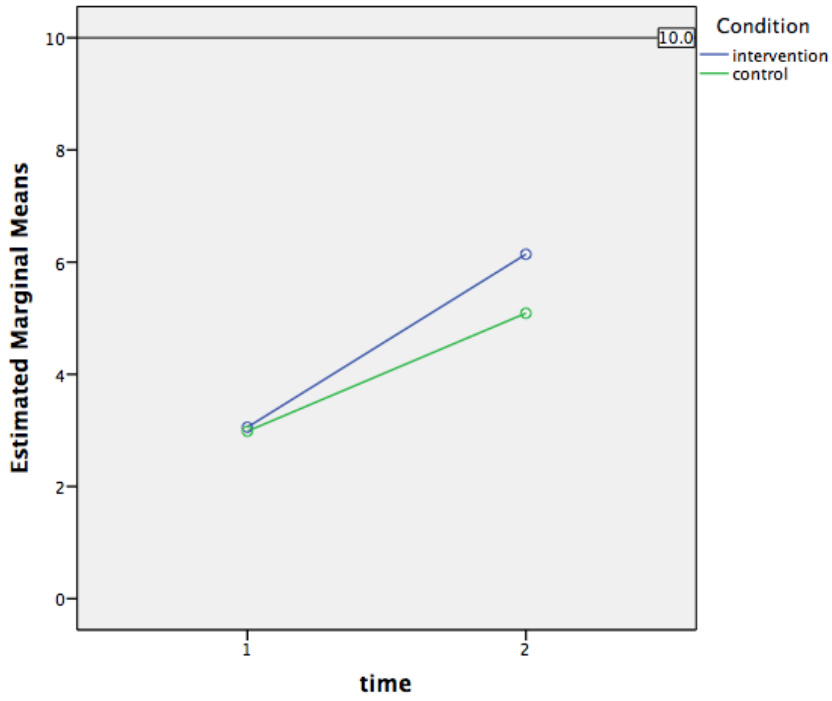
[39] t1 *z*-score of skewness = 3.16, *z*-score of kurtosis = 4.58; t2 *z*-score of skewness = 1.43, *z*-score of kurtosis = 1.87

[40] t1 $F(1,136) = 1.31$, $p = .25$ ; t2 $F(1,136) = 4.58$, $p < .05$

A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -6.90, $p$ < .001, $r$ = -.22) and the comparison group ($Z$ = -5.72, $p$ < .001, $r$= -.21). As the mean written recall scores were higher at t2 compared to t1 for both intervention and comparison groups, this suggests that both groups in Universities A and B made significant gains after the instruction programmes at t2. Then, a Mann-Whitney test was also conducted to examine the differences between the two groups' written recall scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U$ = 2372.5, $Z$ = -.03, $p$ = .98) but were significantly different at t2, with a small effect size ($U$ =1771.5, $Z$ = -2.65, $p$ < .01, $r$ = -.22).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F$(1, 136) = 172.03, $p$ < .001, $r$ = .75, as well as a significant main effect of condition with a small effect size, $F$(1, 136) = 8.12, $p$ < .01, $r$ = .24. The interaction between time and condition was also significant, $F$(1, 136) = 16.55, $p$ < .05, with a small effect size, $r$ = .21. The interaction graph is presented in Figure 7.10. These results support the findings of the non-parametric tests, indicating that the phonics instruction led to significantly greater progress in terms of written recall test results than the phonology instruction, for participants in Universities A and B.

*Figure 7.10 Estimated marginal means of written recall test scores (out of 10) of Universities A and B*

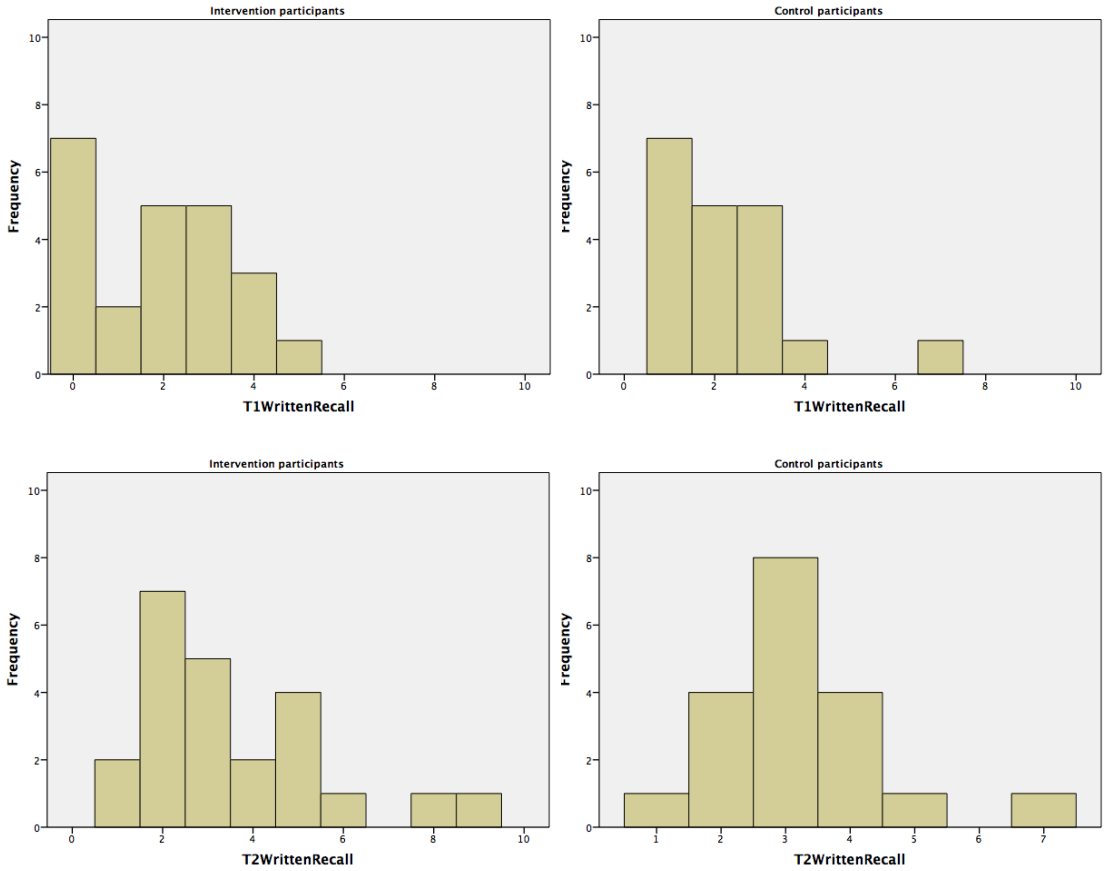



Finally, the written recall test results of participants in University C were analysed. The descriptive statistics of the written recall test results of Universities C are summarised in Table 7.7, and the histograms are presented in Figure 7.11.

*Table 7.7 Written recall scores of University C (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=23)** | 2 | 1.9 | 1.6 | 0 | 5 | 3 | 3.6 | 2.1 | 1 | 9 |
| **Comparison (N=19)** | 2 | 2.3 | 1.5 | 1 | 7 | 3 | 3.2 | 1.3 | 1 | 7 |

*Figure 7.11 Histograms of written recall scores (out of 10) of University C*




The assumptions of a two-way factorial ANOVA with one within subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[41]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at t1 but rejected at t2[42]. As in the previous analyses, both non-parametric and parametric tests were conducted as a point of comparison.

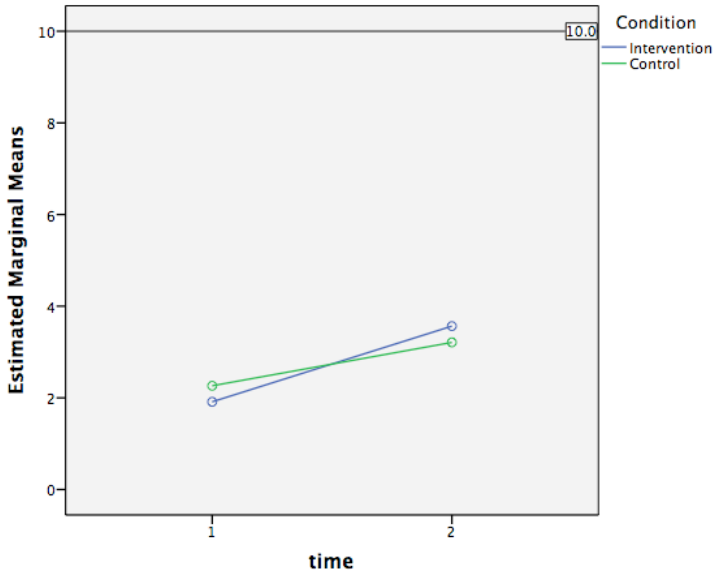
[41] t1 *z*-score of skewness = 2.09, *z*-score of kurtosis = 1.62; t2 *z*-score of skewness = 3.62, *z*-score of kurtosiss= 2.81

[42] t1 $F(1,40) = .80$, $p = .38$ ; t2 $F(1,136) = 4.51$, $p < .05$

A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -3.65, $p$ < .001, $r$ = -.56) as well as the comparison group ($Z$ = -2.97, $p$ < .01, $r$ = -.46). As the mean written recall scores were higher at t2 compared to t1 for both groups, this indicates that both the intervention participants and the comparison groups in University C made significant gains after the instruction programmes. Then, a Mann-Whitney test was also conducted to examine the differences between the two groups' written recall scores at each time point. It was found that the scores of the two groups were neither significantly different at t1 ($U$ = 197.5, $Z$ = -.54, $p$ = .59) nor significantly different at t2 ($U$ = 214, $Z$ = -.12, $p$ = .91).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 40) = 40.08$, $p$ < .001, $r$ = .71. However, there was no significant main effect of condition, $F(1, 40) = .01$, $p$ = .99. The interaction between time and condition was also non-significant, $F(1, 40) = 2.58$, $p$ = .09. The interaction graph is presented in Figure 7.12. These results support the findings of the non-parametric tests, suggesting that, in University C, the phonics instruction did not lead to significantly greater gains in terms of the written recall test results compared to the phonology instruction. This also echoes the analysis of the oral recall test results, where the intervention participants in University C likewise did not show significantly greater gains compared to the comparison groups at t2.

*Figure 7.12 Estimated marginal means of written recall test scores (out of 10) of University C*




In summary, the analysis of the written recall test results yielded very similar findings to the analysis of the oral recall test results. The intervention participants in all three universities made significantly greater gains compared to the comparison groups in the written recall test, suggesting that phonics instruction programme was useful in facilitating the recall of written forms in vocabulary learning. The analysis of the aggregated results of participants in Universities A and B alone also revealed similar results. Similar to the findings of the analysis of the oral recall test, intervention participants in University C did not demonstrate significantly greater gains compared to their comparison counterparts in the written recall test either.

In order to examine whether the phonics instruction would lead to any changes in the errors in the written recall test, an analysis of the errors made by participants in Universities A and B were also conducted. Here, errors are defined as incorrect or

incomplete spellings; therefore, if a participant did not produce an answer, this was not counted as an error. University C was excluded from the analysis because the phonics instruction programme did not lead to significantly greater gains in phonological decoding for the intervention participants in University C compared to their comparison counterparts.

The examination of the spellings of participants in Universities A and B revealed 206 errors (intervention participants: 100, comparison groups: 106) at t1 and 276 errors (intervention participants: 151, comparison groups: 125) at t2. Note that the increase in number of errors between t1 and t22 reflects the fact that participants attempted more spellings of words at t2: i.e. there were more blank responses at t1, and these were not counted as errors for the purposes of this analysis.

Participants' errors can be further divided into three main types, namely: (a) phonologically plausible errors, which can be pronounced in the same way as the target word but are spelt wrongly (e.g. *birlap* for *burlap*; *calles* for *callus*); (b) positional errors, which have the same letters as the stimuli or part of the stimuli but are spelt in the wrong order (e.g. *zepyhr* for *zephyr*, *pliodes* for *ploidy*) and (c) other errors, which do not fall into the previous two categories (e.g. '*doodish*' for doodah, *augle* for *augean*). The other examples of the three types of errors can be found in Table 7.8.

*Table 7.8 Errors in the written recall test by Universities A and B*

| Error type | Target word | Error example |
|---|---|---|
| Phonologically plausible errors | burlap | birlap, berlap |
| | callus | calles |
| | zeugma | zugma |
| | sulcus | selcus |
| | mayhap | maihap |
| | cantor | canter |
| | augean | augin |
| | tisane | tisane, teesane |
| | rheumy | rhoomy |
| | zephyr | zepher |
| | maenad | meenad |
| Positional errors | zephyr | zepyhr |
| | ploidy | pliodies, pliody |
| | cantor | catron |
| | zeugma | zuegma |
| | rheumy | reuhmy |
| | tisane | tisaen |
| | maenad | meanad |
| Other errors | doodah | doodish, doodem |
| | rheumy | ruphy |
| | ploidy | ploipher, poidy |
| | augean | augent, argean, augle |
| | zeugma | zigma, zeegma, zeogma |
| | tisane | teesan, tasle |

The number of each type of errors committed at the two time points was also counted.

The results are presented in Table 7.9.

*Table 7.9 Number of different types of errors at each time point (by Universities A and B)*

| | Phonologically plausible errors | | Positional errors | | Other errors | |
|---|---|---|---|---|---|---|
| | t1 | t2 | t1 | t2 | t1 | t2 |
| **Intervention (N=71)** | 24 | 103 | 50 | 19 | 26 | 29 |
| **Comparison (N=67)** | 19 | 24 | 57 | 52 | 30 | 49 |

It can be seen that at t1, both the intervention and the comparison group in Universities A and B made most positional errors among the three types of errors. For both groups, phonologically plausible errors were the least frequent among the three error types at t1. However, the intervention group made many more phonologically plausible errors at t2, while the number of positional errors dropped sharply. In contrast, the comparison group made roughly the same number of phonologically plausible errors and positional errors at t1 and t2, with positional errors remaining the most commonly made errors at t2.

In addition, a question naturally arises as to whether there is a relationship between participants' decoding proficiency and the number of errors they made in each category. As a result, a Pearson's correlation test was conducted for each group at each time point with the following variables: word-level decoding scores, grapheme-level decoding scores, number of phonologically plausible errors, number of positional errors and number of other errors. All the significant correlations found are reported below.

The results show that at t1, there was a significant correlation between word-level decoding scores and the number of phonologically plausible errors (intervention group: $r = .55$, $p < .01$; comparison group: $r = .42$, $p < .01$). The correlation between grapheme-level decoding scores and the number of phonologically plausible errors was also found to be significant for both intervention and comparison groups

(intervention group: $r = .30$, $p < .05$; comparison group: $r = .35$, $p < .01$). This suggests that the participants with higher decoding proficiency were more likely to make phonologically plausible errors. The correlations between decoding scores, both word-level and grapheme-level, and the number of errors in the other two categories (positional errors and other errors) were not significant. At t2, the correlation between word-level decoding scores and the number of phonologically plausible errors remained significant for both the intervention and the comparison group, but was considerably higher for the intervention group (intervention group: $r = .57$, $p < .001$; comparison group: $r = .13$, $p < .01$). Similarly, the correlation between grapheme-level decoding scores and the number of phonologically plausible errors was found to be significant for both the intervention and the comparison group, and higher for the intervention group (intervention group: $r = .51$, $p < .001$; comparison group: $r = .12$, $p < .01$). This suggests that regardless of the instruction programme they received, participants with higher decoding proficiency were more likely to make phonologically plausible errors.

The correlation between intervention participants' word-level decoding score and the number of positional errors was found to be significant and negative at t2 ($r = -.25$, $p < .05$), as well as the correlation between intervention participants' grapheme-level decoding score and the number of positional errors ($r = -.24$, $p < .05$). This suggests that after the phonics instruction, intervention participants with higher proficiency were less likely to commit positional errors.

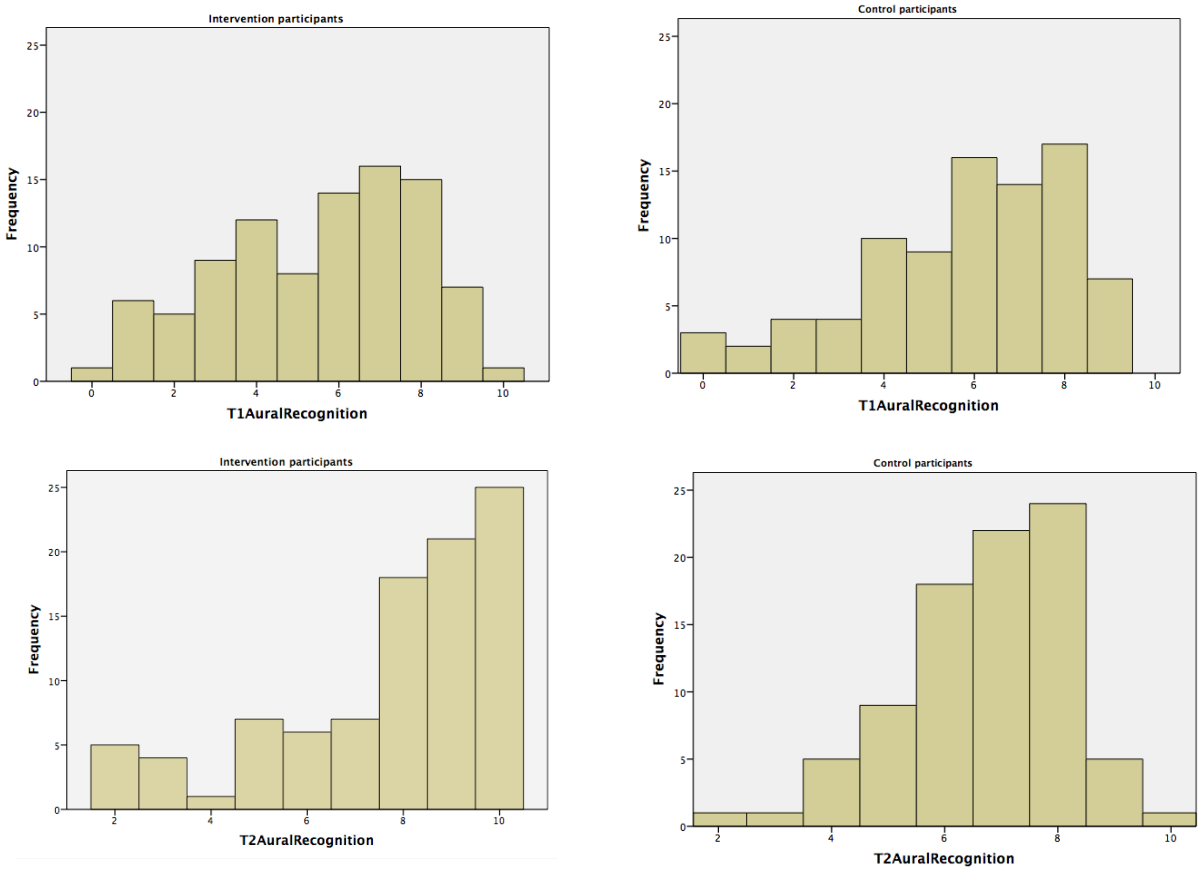
### 7.3 Aural recognition test results (Hear English, Write Chinese)

The number of correctly recognised words was counted as the score of the aural recognition test. The descriptive statistics of the aural recognition test are summarised in Table 7.10, and the histograms are presented in Figure 7.13.

*Table 7.10 Aggregated aural recognition scores for all three universities (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 6 | 5.5 | 2.4 | 0 | 10 | 8 | 7.8 | 2.3 | 2 | 10 |
| **Comparison (N=86)** | 6 | 5.8 | 2.3 | 0 | 9 | 7 | 6.7 | 2.2 | 2 | 10 |

*Figure 7.13 Histograms of aggregated aural recognition scores (out of 10) for all three universities*



The assumptions of a two-way factorial ANOVA with one within subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[43]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at t1 but rejected at t2[44]. As a result, the data was primarily analysed using non-parametric tests. However, parametric tests were also conducted as a means of comparison.

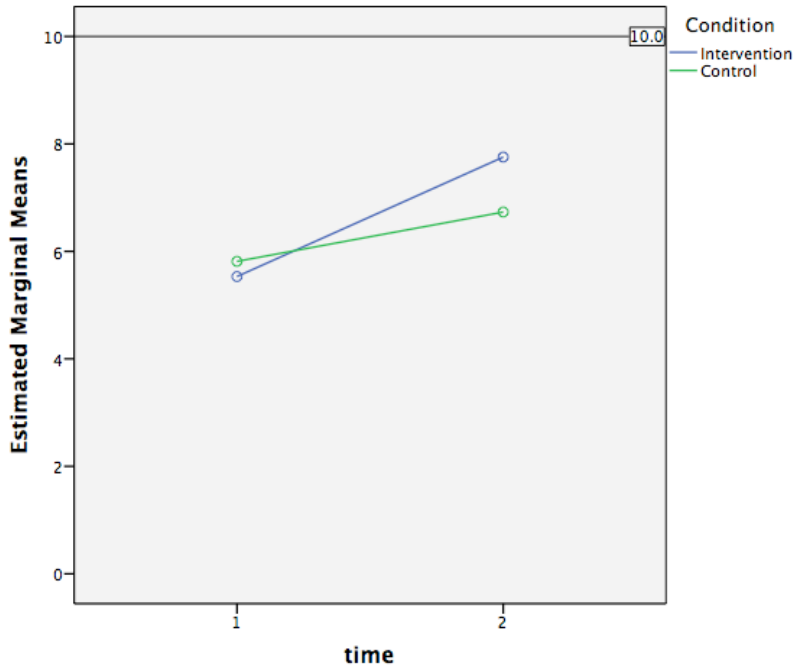
Firstly, a Wilcoxon Signed Rank test was conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention participants ($Z$= -7.93, $p < .001$, $r$ = -.59) but not the comparison groups ($Z$ = - .60, $p$ = .55). This indicates that only the intervention participants made significant gains in the aural recognition test. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' aural recognition scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U$ = 3756.5, $Z$ = -.83, $p$ = .41) but were significantly different at t2, with a small effect size ($U$ = 2431.5, $Z$ = - 4.67, $p < .001$, $r$ = .35).

[43] t1 $z$-score of skewness = -3.02, $z$-score of kurtosis = -3.82; t2 $z$-score of skewness = -11.13, $z$-score of kurtosis = 1

[44] t1 $F(1,178) = 1.05$, $p = .31$; t2 $F(1,178) = 13.81$, $p < .001$

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 178) = 83.89$, $p < .001$, $r = .57$. The effect of condition was non-significant, $F(1, 178) = 1.82$, $p = .18$. The interaction between time and condition was significant, $F(1, 178) = 14.47$, $p < .001$, with a small effect size, $r = .27$. The interaction graph is presented in Figure 7.14. These results support the findings of the non-parametric tests, indicating that the phonics instruction led to significantly greater progress in terms of the aural recognition test results compared to the phonology instruction. The results are also in line with the analyses of the oral recall and the written recall tests.

*Figure 7.14 Estimated marginal means of aural recognition test scores (out of 10) of all three universities*



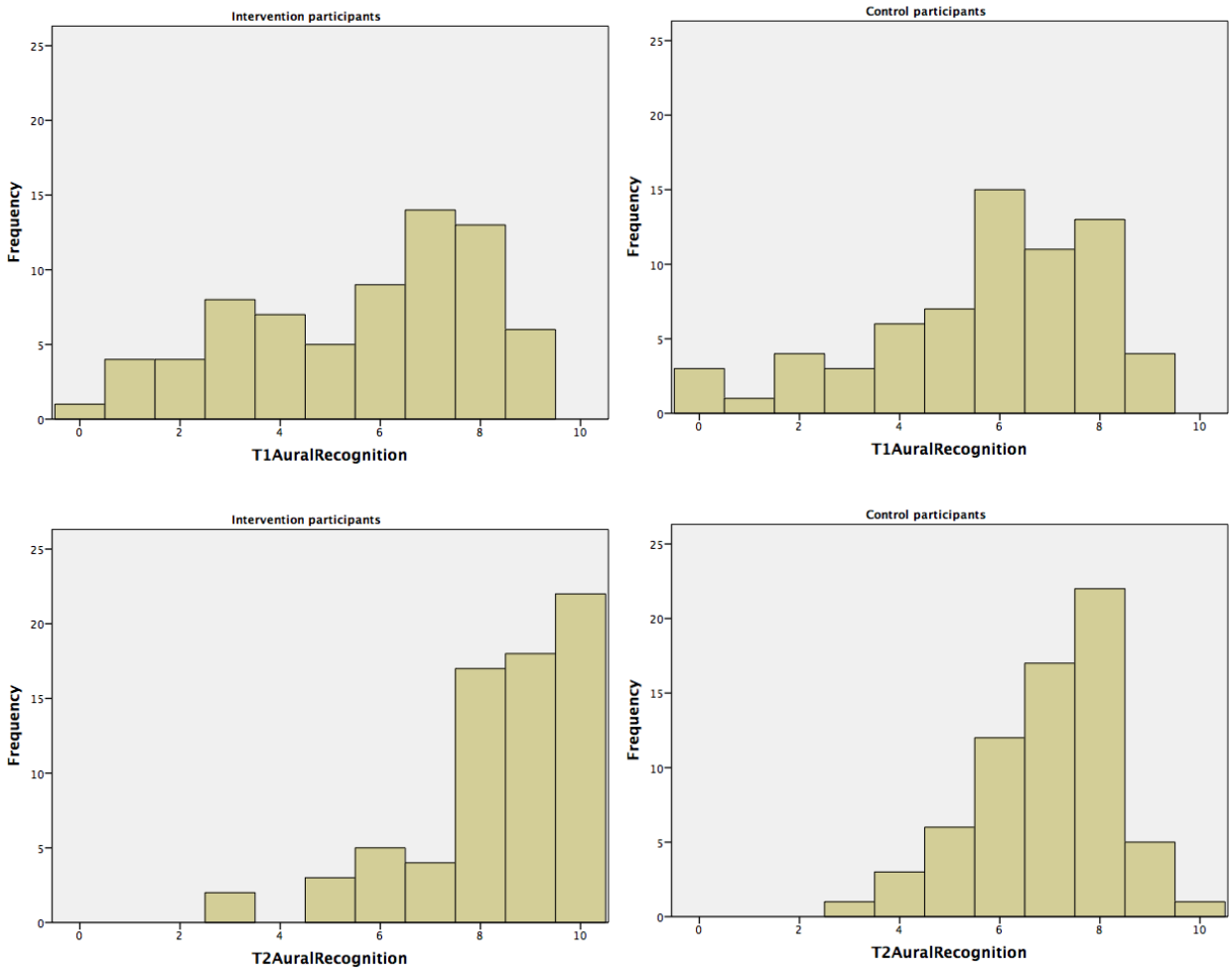

The aggregated aural recognition test results of Universities A and B were then analysed separately. The descriptive statistics of the aural recognition test results of

Universities A and B are summarised in Table 7.11, and the histograms are presented in Figure 7.15.

*Table 7.11 Aural recognition scores of Universities A and B (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 6.0 | 5.6 | 2.4 | 0 | 9 | 9.0 | 8.4 | 2.8 | 3 | 10 |
| **Comparison (N=67)** | 6.0 | 5.7 | 2.3 | 0 | 9 | 7.0 | 7.0 | 2.0 | 3 | 10 |

*Figure 7.15 Histograms of aural recognition scores of Universities A and B (out of 10)*




The assumptions of a two-way mixed factorial ANOVA with one with-in subjects

variable (time) and one between-subjects variable (condition) were checked. Firstly, the assumption of Normality was violated at both t1 and t2[45]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at both t1 and t2[46]. Following the previous analyses, both non-parametric and parametric tests were conducted and the results were compared.

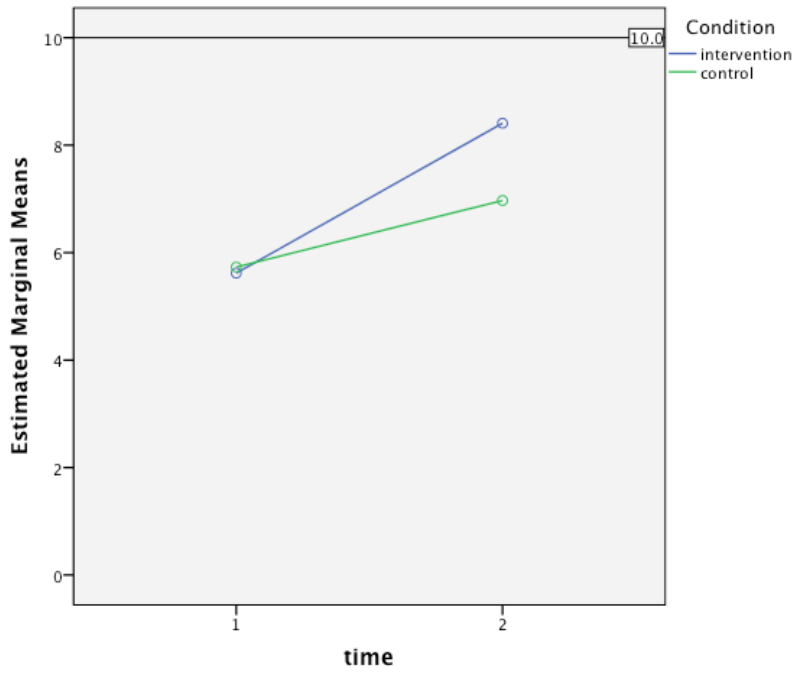
A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for both the intervention group ($Z$ = -6.58, $p$ < .001, $r$ = -.57) and the comparison group ($Z$ = -4.38, $p$ < .001, $r$ = -.37). As the mean aural recognition scores were higher at t2 compared to t1 for both intervention and comparison groups, this suggests that both groups in Universities A and B made significant gains after the instruction programmes at t2. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' aural recognition scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U$ = 2347, $Z$ = -.14, $p$ = .89) but were significantly different at t2, with a small effect size ($U$ =1077, $Z$ = -5.65, $p$ < .001, $r$ = -.34).

[45] t1 $z$-score of skewness = -3.16, $z$-score of kurtosis = -1.07; t2 $z$-score of skewness = -3.01, $z$-score of kurtosis = .01

[46] t1 $F(1, 136) = 1.51$, $p$ = .22, t2 $F(1, 136) = 1.48$, $p$ = .23

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 136) = 119.93$, $p < .001$, $r = .68$, as well as a significant main effect of condition with a small effect size, $F(1, 136) = 5.34$, $p < .05$, $r = .19$. The interaction between time and condition was also significant, $F(1, 136) = 17.76$, $p < .001$, with a medium effect size, $r = .34$. The interaction graph is presented in Figure 7.16. These results support the findings of the non-parametric tests, indicating that the phonics instruction led to significantly greater progress in terms of the aural recognition test results than the phonology instruction, for participants in Universities A and B.

*Figure 7.16 Estimated marginal means of aural recognition scores of Universities A and B*



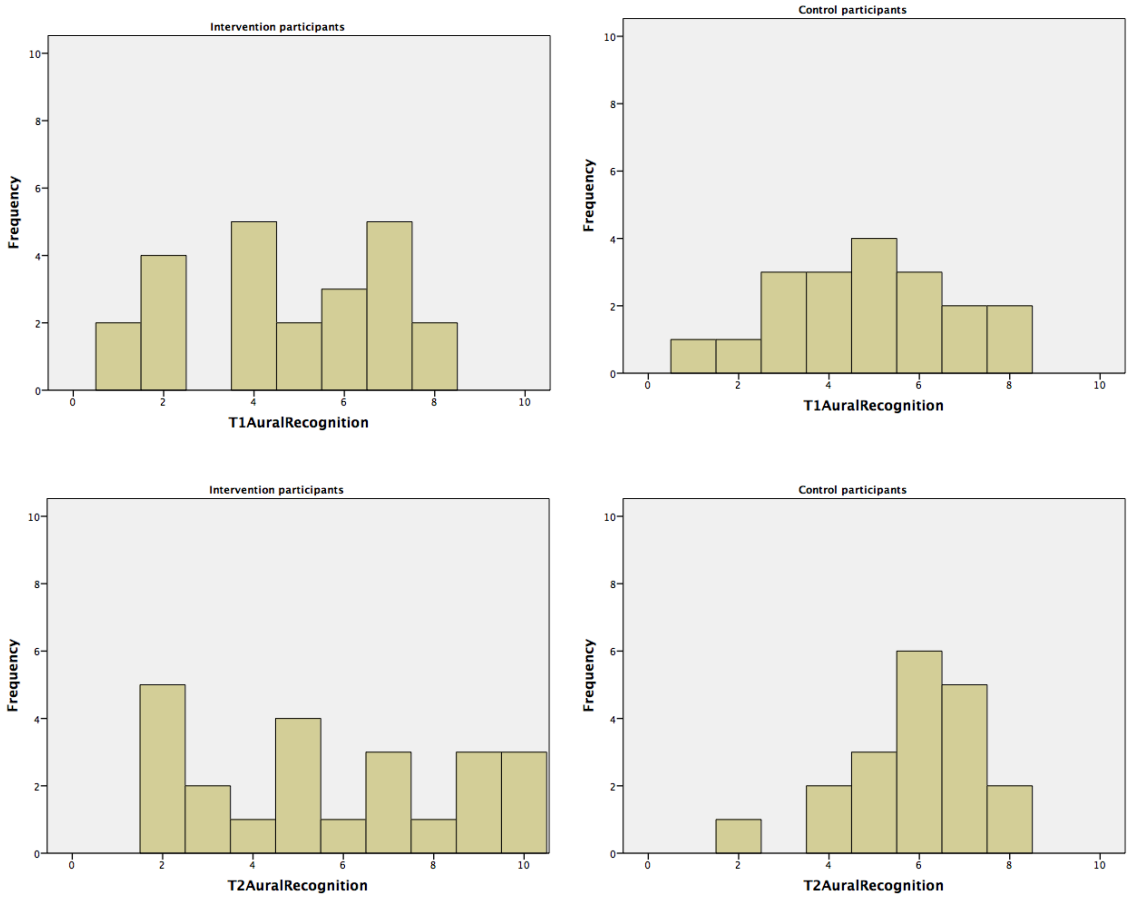

Finally, the aural recognition test results of University C were analysed. The descriptive statistics of the aural recognition test results of Universities C are

summarised in Table 7.12, and the histograms are presented in Figure 7.17.

*Table 7.12 Aural recognition scores of University C (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=23)** | 5 | 4.7 | 2.3 | 1 | 8 | 5 | 5.7 | 2.9 | 2 | 10 |
| **Comparison (N=19)** | 5 | 4.8 | 2.0 | 1 | 8 | 6 | 5.9 | 1.5 | 2 | 8 |

*Figure 7.17 Histograms of aural recognition scores of University C (out of 10)*




The assumptions of a two-way factorial ANOVA with one within subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the

assumption of Normality was violated[47]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at t1 but rejected at t2[48]. As in the previous analyses, both non-parametric and parametric tests were conducted as a point of comparison.

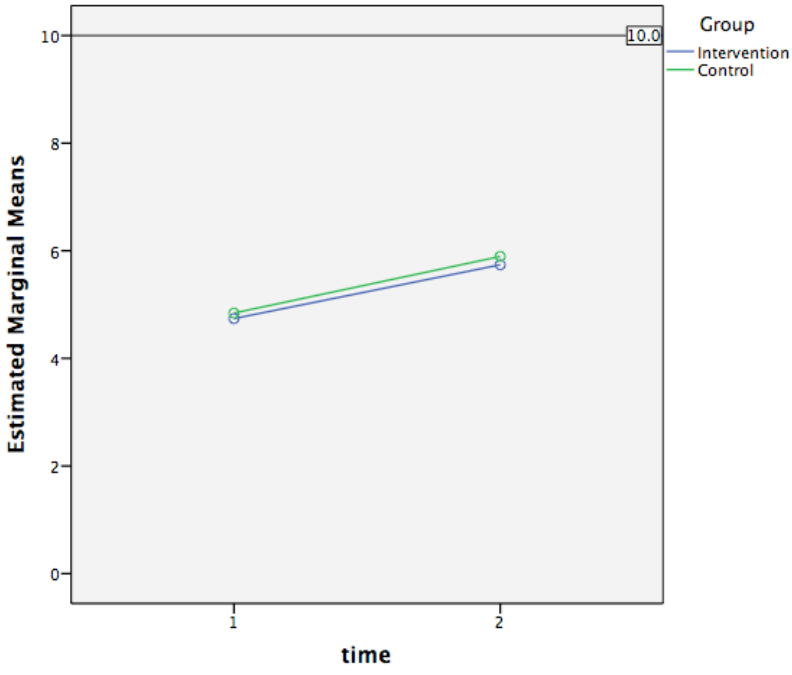
A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -2.92, $p$ < .01, $r$ = -.45) as well as the comparison group ($Z$ = -2.10, $p$ < .05, $r$ = -.32). As the mean aural recognition scores were higher at t2 compared to t1 for both groups, this indicates that both the intervention participants and the comparison groups in Universities C made significant gains after the instruction programmes. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' aural recognition scores at each time point. It was found that the scores of the two groups were neither significantly different at t1 ($U$ = 216, $Z$ = -.06, $p$ = .95) nor significantly different at t2 ($U$ = 209, $Z$ = -.24, $p$ = .81).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 40) = 16.10$, $p < .001$, $r = .54$. However, there was no significant main effect of condition, $F(1, 40) = .04$, $p = .84$. The interaction

[47] t1 $z$-score of skewness= 2.92, $z$-score of kurtosis= 1; t2 $z$-score of skewness= 1.41, $z$-score of kurtosis= 1
[48] t1 $F(1,40)= 1.16$, $p= .29$ ; t2 $F(1,136)= 14.29$, $p< .01$

between time and condition was also non-significant, $F(1, 40) = .01$, $p = .92$. The interaction graph is presented in Figure 7.18. These results support the findings of the non-parametric tests, suggesting that the phonics instruction did not lead to significantly greater gains in terms of the aural recognition test results for the intervention participants than the comparison groups in University C. This also echoes the analyses of the oral recall and written recall test results, where the intervention participants in University C also did not show significantly greater gains compared to the comparison groups at t2.

*Figure 7.18 Estimated marginal means of aural recognition scores of University C*




In summary, the analysis of the aural recognition test results yielded similar findings to the analyses of the oral recall and the written recall test results. The intervention participants in all three universities made significantly greater gains compared to the

comparison groups in the aural recognition test, indicating the effectiveness of the phonics instruction programme in promoting the recognition of phonological forms in vocabulary learning. The analysis of the aggregated results of participants in Universities A and B also revealed similar results. Similar to the findings of the analyses of the oral recall and written recall tests, participants in University C did not demonstrate significantly greater gains compared to their comparison counterparts in the aural recognition test.

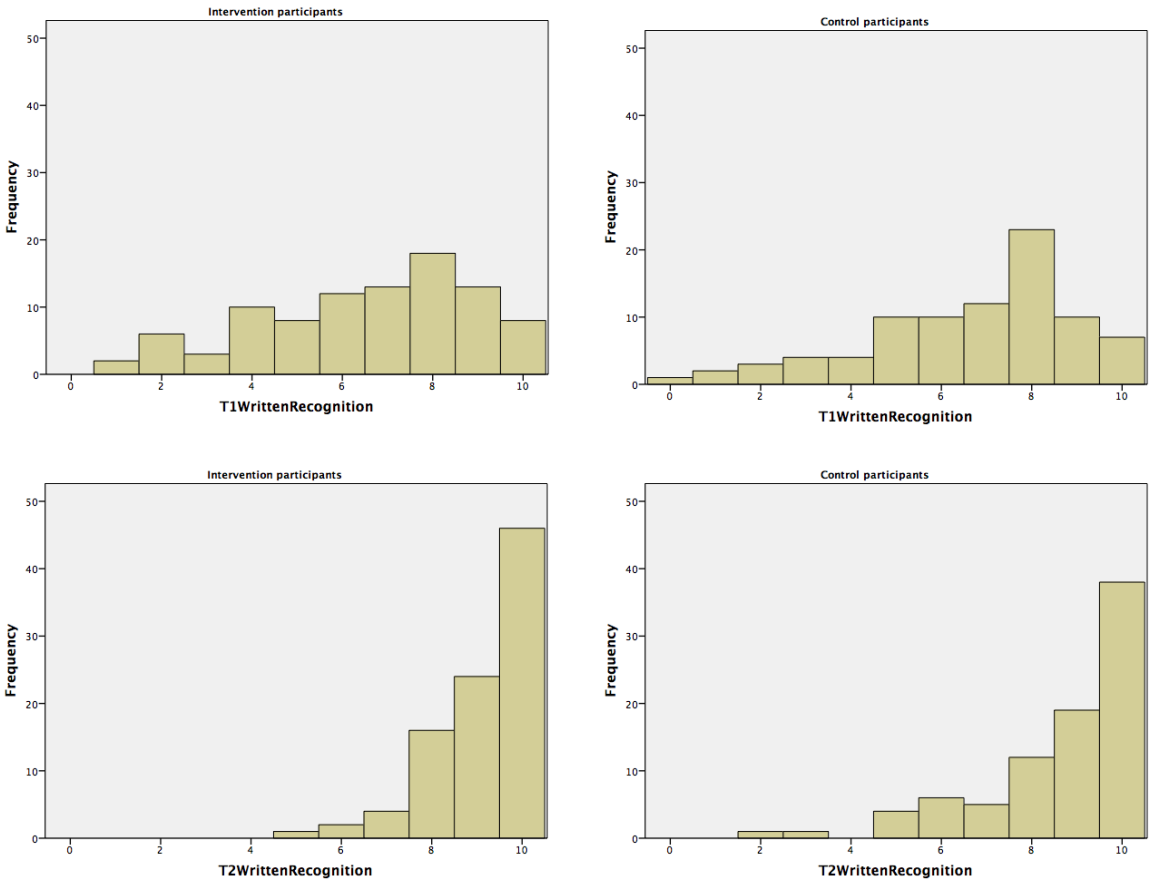
### 7.4 Written recognition test results (See English, Write Chinese)

The number of correctly recognised words was counted as the score of the written recognition test. The descriptive statistics of the written recognition test are summarised in Table 7.13, and the histograms are presented in Figure 7.19.

*Table 7.13 Written recognition scores of all three universities (out of 10)*

| | **t1** | | | | | **t2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=94)** | 7 | 6.5 | 2.4 | 1 | 10 | 9 | 8.9 | 1.1 | 5 | 10 |
| **Comparison (N=86)** | 7 | 6.7 | 2.3 | 0 | 10 | 9 | 8.6 | 1.8 | 2 | 10 |

*Figure 7.19 Histograms of written recognition scores of all three universities (out of 10)*




The assumptions of a two-way factorial ANOVA with one within subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[49]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at t1 but rejected at t2[50]. As a result, the data was primarily analysed using non-parametric tests. However, parametric tests were also conducted as a means of comparison.

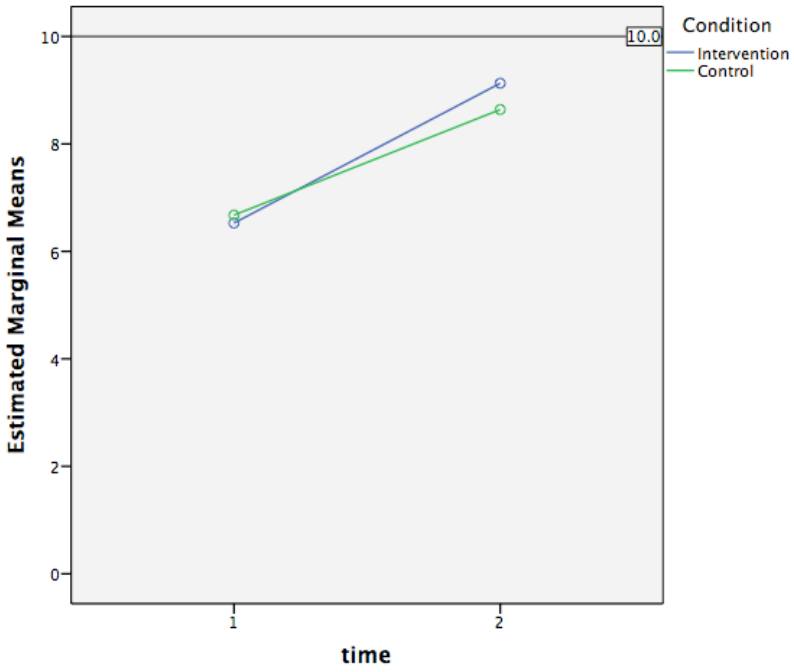
[49] t1 *z*-score of skewness = -3.48, *z*-score of kurtosis = -.92; t2 *z*-score of skewness = -9.78, *z*-score of kurtosis = 1
[50] t1 $F(1,178) = .30$, $p = .58$; t2 $F(1,178) = 13.75$, $p < .001$

Firstly, a Wilcoxon Signed Rank test was conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for both the intervention participants ($Z$= -4.21, $p < .001$, $r$ = -.31) and the comparison groups ($Z$ = -3.48, $p < .001$, $r$ = -.26). As the written aural recognition scores were higher at t2 compared to t1 for both groups, this suggests that both the intervention participants and the comparison groups in all three universities made significant gains in the written recognition test. Then, a Mann-Whitney test was also conducted to examine the differences between the two groups' written recognition scores at each time point. It was found that the scores of the two groups were neither significantly different at t1 ($U$ = 3860, $Z$ = -.53, $p$ = .60) nor significantly different at t2 ($U$ = 3544, $Z$ = -1.40, $p$ = .16).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 178) = 230.03$, $p < .001$, $r$ = .57. The effect of condition was non-significant, $F(1, 178) = .48$, $p$ = .50. The interaction between time and condition was also non-significant, $F(1, 178) = 4.48$, $p$ = .06. The interaction graph is presented in Figure 7.20. These results support the findings of the non-parametric tests, indicating that the phonics instruction did not lead to significantly greater progress in terms of the written recognition test results than the phonology instruction. This is in contrast with the finding of the analyses of the other three tests, leaving the written recognition test the only one of the four test forms

where the phonics instruction programme did not lead to significantly greater gains for the intervention participants.

*Figure 7.20 Estimated marginal means of written recognition scores of all three universities*

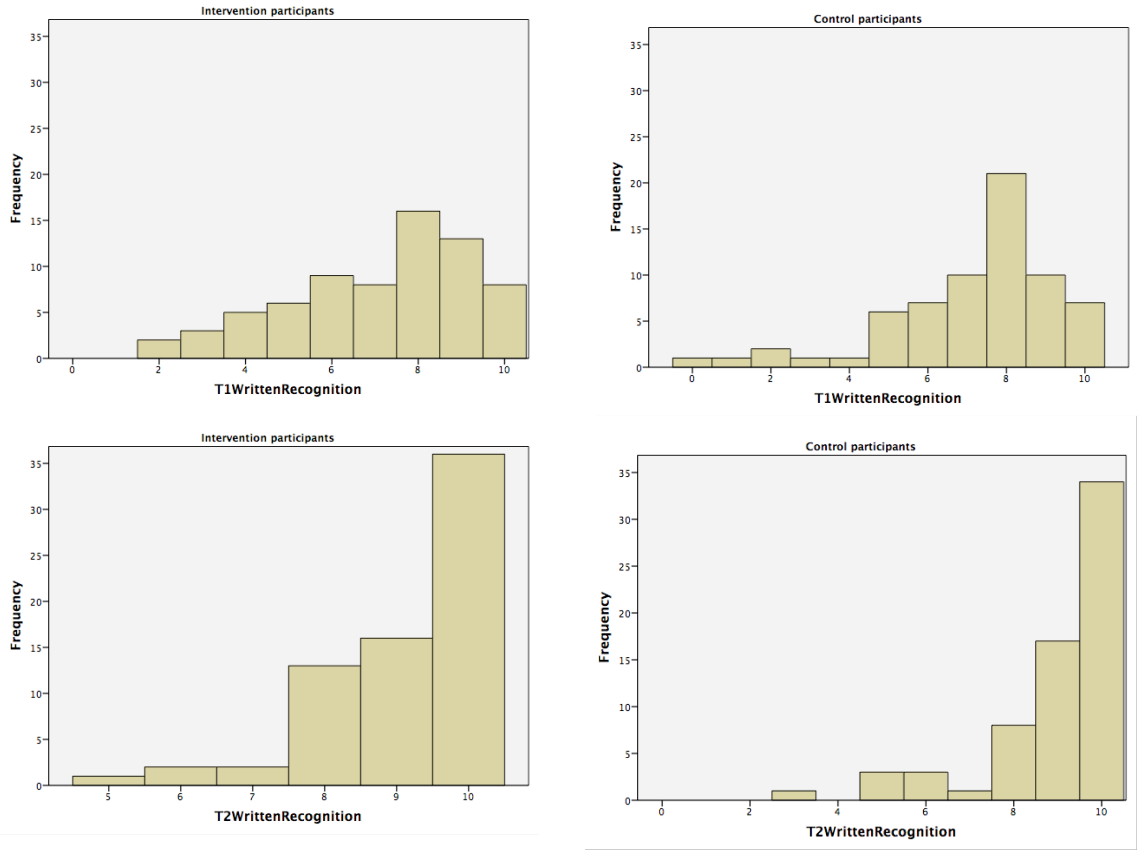



The aggregated written recognition test results of Universities A and B were then analysed separately. The descriptive statistics of the written recognition test results of Universities A and B are summarised in Table 7.14, and the histograms are presented in Figure 7.21.

*Table 7.14 Written recognition scores of Universities A and B (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=71)** | 8.00 | 7.1 | 2.1 | 2 | 10 | 10.00 | 9.1 | 1.1 | 5 | 10 |
| **Comparison (N=67)** | 8.00 | 7.2 | 2.2 | 0 | 10 | 10.00 | 9.0 | 1.5 | 3 | 10 |

*Figure 7.21 Histograms of written recognition scores of Universities A and B*




The assumptions of a two-way mixed factorial ANOVA with one within subjects variable (time) and one between-subjects variable (condition) were checked. Firstly, the assumption of Normality was violated at both t1 and t2[51]. Secondly, the issue of

[51] t1 *z*-score of skewness = -4.45, *z*-score of kurtosis = 1.21; t2 *z*-score of skewness = -9.01, *z*-score of kurtosis = 9.27

sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2). Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at both t1 and t2[52]. As a result, both non-parametric and parametric tests were conducted as a point of comparison.

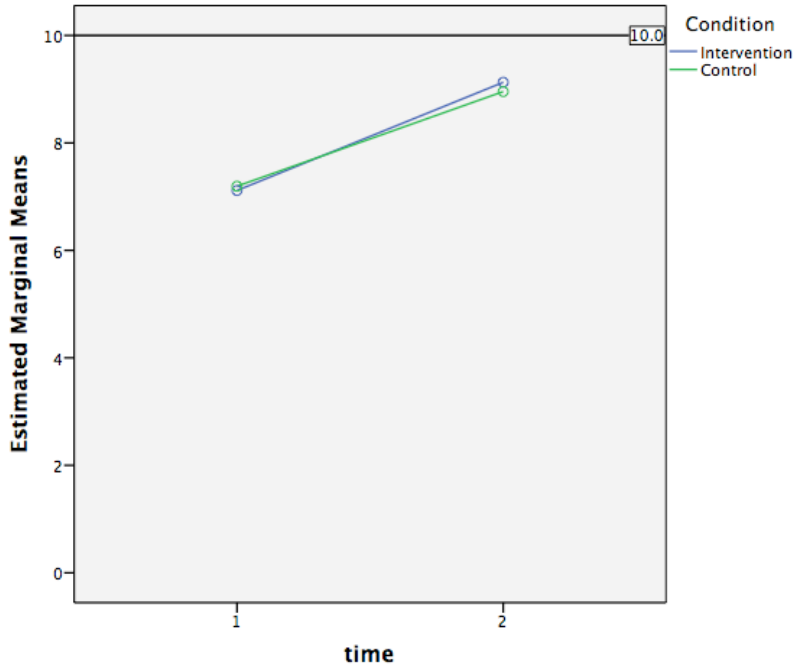
A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -6.52, $p$ < .001, $r$ = -.56) and the comparison group ($Z$ = -6.01, $p$ < .001, $r$ = -.50). This indicates that both groups (based on participants in Universities A and B) made significant gains after the instruction programmes at t2. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' written recognition scores at each time point. It was found that the scores of the two groups were neither significantly different at t1 ($U$ = 2292, $Z$ = -.37, $p$ = .71) nor significantly different at t2 ($U$ = 2313.5, $Z$ = -.15, $p$ = .88).

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 136) = 161.79$, $p$ < .001, $r$ = .74. However, the main effect of condition was not found to be non-significant, $F(1, 136) = .03$, $p$ = .86. The interaction between time and condition was also non-significant, $F(1, 136) = .73$, $p$ = .40. The interaction graph is presented in Figure 7.22. These results support the

[52] t1 $F(1, 136) = .36$, $p$ = .55, t2 $F(1, 136) = 1.44$, $p$ = .23

findings of the non-parametric tests, indicating that the phonics instruction did not lead to significantly greater progress in terms of the written recognition test results than the phonology instruction, for participants in Universities A and B. This is also in line with the findings of the analysis of the written recognition test results of the sample as a whole (all three universities).

*Figure 7.22 Estimated marginal means of written recognition scores of Universities A and B*

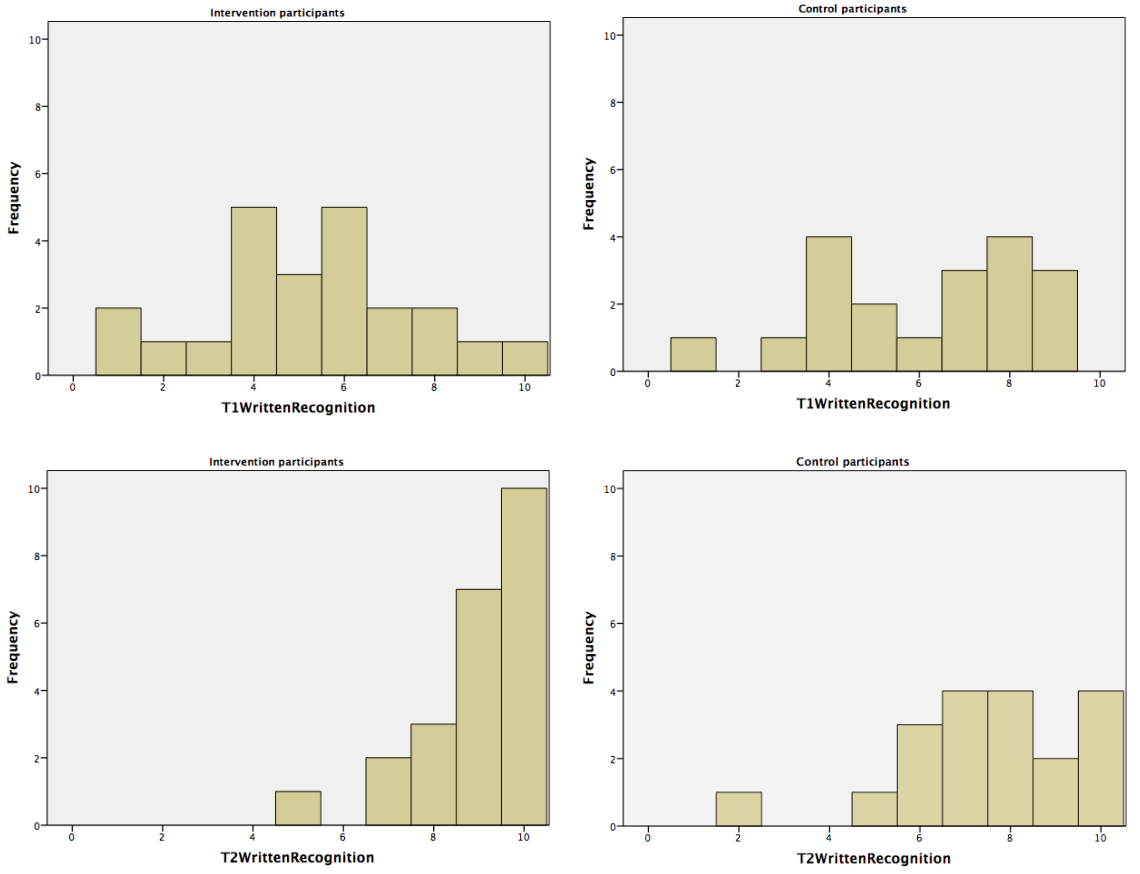



Finally, the written recognition test results of University C were analysed. The descriptive statistics of the written recognition test results of Universities C are summarised in Table 7.15, and the histograms are presented in Figure 7.23.

*Table 7.15 Written recognition scores of University C (out of 10)*

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Median** | **Mean** | **S.D.** | **Min** | **Max** | **Median** | **Mean** | **S.D.** | **Min** | **Max** |
| **Intervention (N=23)** | 5 | 5.3 | 2.3 | 1 | 10 | 7 | 9.0 | 1.3 | 5 | 10 |
| **Comparison (N=19)** | 7 | 6.1 | 2.3 | 1 | 9 | 8 | 7.5 | 2.0 | 2 | 10 |

*Figure 7.23 Histograms of written recognition scores of University C (out of 10)*




The assumptions of a two-way factorial ANOVA with one within subject variable (time) and one between-subject variable (condition) were firstly checked. Firstly, the assumption of Normality was violated[53]. Secondly, the issue of sphericity was not considered here as the repeated measures variable had only two levels (t1 and t2).

[53] t1 *z*-score of skewness = -.55, *z*-score of kurtosis = -.86; t2 *z*-score of skewness = -3.65, *z*-score of kurtosis = 3.15

Thirdly, Levene's test revealed that the assumption of homogeneity of variances was retained at t1 but rejected at t2[54]. Following the previous analyses, both non-parametric and parametric tests were conducted as a point of comparison.

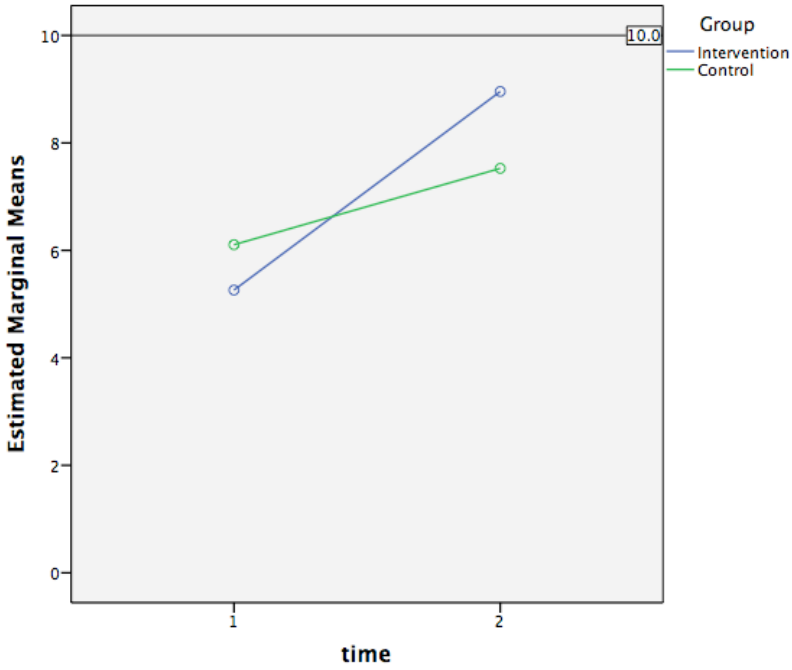
A Wilcoxon Signed Rank test was firstly conducted to examine whether or not, for each individual group, there was a significant difference in their scores between t1 and t2. It was found that the scores at t1 and t2 were significantly different for the intervention group ($Z$ = -4.04, $p < .001$, $r$ = -.62) as well as the comparison group ($Z$ = -2.29, $p < .05$, $r$ = -.35). As the mean written recognition scores were higher at t2 compared to t1 for both groups, this indicates that both the intervention participants and the comparison groups in University C made significant gains after the instruction programmes. Then, a Mann-Whitney test was conducted to examine the differences between the two groups' written recognition scores at each time point. It was found that the scores of the two groups were not significantly different at t1 ($U$ = 172.5, $Z$ = -1.17, $p$ = .24) but were significantly different at t2 ($U$ = 120.5, $Z$ = -2.55, $p < .05$, $r$ = -.39). Thus, at t2, the intervention group obtained a significantly higher score on the written recognition test than the comparison group.

An ANOVA was also conducted as a point of comparison. There was a main effect of time with a large effect size, $F(1, 40) = 48.75$, $p < .001$, $r = .74$. However, there was no significant main effect of condition, $F(1, 40) = .33$, $p = .57$. The interaction

[54] t1 $F(1,40) = 1.16$, $p = .29$ ; t2 $F(1,136) = 14.29$, $p < .01$

between time and condition was significant with a medium effect size, $F(1, 40) = 9.63$, $p < .01$, $r = .44$. The interaction graph is presented in Figure 7.24. These results support the findings of the non-parametric tests, indicating that the phonics group showed significantly greater gains in terms of the written recognition test results than the comparison group, for participants in University C. This will be further discussed in section 8.5.2.

*Figure 7.24 Estimated marginal means of written recognition scores of University C*




In summary, the analysis of the written recognition test yielded contrasting findings to those of the analyses of the other three test forms. The phonics instruction programme did not lead to significantly greater gains for the intervention participants than the comparison groups, regardless of whether University C was included or not. Conversely, a separate analysis of University C alone revealed that the intervention

participants in University C did achieve significantly greater gains than the comparison groups at t2 in the written recognition test.

## 7.5 Summary

This chapter has compared the results of the intervention and comparison groups in terms of the four vocabulary recall and recognition tests. The intervention participants showed significantly more gains than the comparison groups in three of the tests, regardless of whether or not University C was included: namely the oral recall test, the written recall test and the aural recognition test. This provides an affirmative answer to RQ 4, at least with respect to the effectiveness of phonics instruction in promoting the recall and recognition of phonological forms, as well as the recall of written forms, in vocabulary learning.

Error analyses of the written recall test results of participants in Universities A and B revealed that the intervention participants made more phonologically plausible errors after the phonics instruction. Correlation tests also revealed that both intervention and comparison groups' decoding scores were positively correlated with the number of phonologically plausible errors at both time points, suggesting that those with higher decoding proficiency were more likely to make phonologically plausible errors regardless of whether they received phonics instruction. A significant negative correlation was also found between intervention participants' decoding scores and the

number of positional errors they made at t2, suggesting that after the phonics instruction, those with higher decoding proficiency in the intervention group were less likely to make positional errors.

The intervention participants did not make significantly greater gains compared to the comparison groups in the written recognition test at t2, suggesting that the phonics instruction did not facilitate the recognition of written forms in vocabulary learning.

The results of University C were analysed separately, as they performed differently from the other two universities in the decoding test as revealed in Chapter 4, possibly due to the confounding effect of learning another foreign language (French or Spanish). It was found that the results of University C were in contrast to the results of the other two universities in all four tests: the intervention participants in University C did not make significantly greater improvements than the comparison groups at t2 in the oral recall test, written recall test or aural recognition test, but they did make significantly greater improvements than the comparison groups in the written recognition test at t2. This contrasting finding will be further discussed in Chapter 8.

# Chapter 8. Discussion

## 8.1 Baseline data

At t1, participants' NCEEE (National College Entrance English Exam) and BPVS (British Picture Vocabulary Scale) scores were gathered in order to determine whether the intervention and comparison groups were matched in terms of English proficiency and vocabulary knowledge. The results demonstrate that intervention and comparison groups – both in the sample as a whole, and within each university individually – did not demonstrate any significant differences in terms of their NCEEE or BPVS scores. This suggests that participants' English proficiency and receptive vocabulary knowledge can be excluded as confounding variables in this study.

One interesting observation from the analysis of the baseline data is the 'mismatch' between these NCEEE and BPVS scores. As mentioned in section 3.2, participants came from three different universities with varying admission criteria, namely one first-tier university with the highest admission criteria (University A), one second-tier university (University B) and one third-tier university with the lowest admission criteria (University C). Given this difference, one would have hypothesised that participants in University A would achieve the highest mean NCEEE and BPVS scores among the three universities, and those in University C would achieve the lowest mean NCEEE and BPVS scores, with University B lying in between. This was

indeed true for the NCEEE scores, where University A achieved significantly higher mean score than University B and University C, and University B achieved significantly higher mean scores than University C. However, the analysis of the BPVS scores revealed a different picture, whereby (on a descriptive level) University B achieved the highest mean scores of the three universities. (Though participants in University B did not achieve significantly higher mean scores than those in University A, they did score significantly higher than the participants in University C).

It is natural to wonder why the participants in University B, who did not achieve the highest mean scores in the NCEEE, an examination of English proficiency (listening, reading and writing), nonetheless received the highest mean scores in the BPVS, a test of English vocabulary breadth. One possible reason for University B's higher vocabulary scores may lie in the background of the participants. Though the participants in this study mostly come from the same region in China, University B has the highest percentage of local students (35%), who were born and raised in Wuhan, the biggest city in central China, compared to University A (10%) and University C (20%). Though a comprehensive social-economic status survey was not conducted, the interaction with the participants over the four months of this study suggested that the majority of the participants in University B came from cities, while nearly half of the participants in Universities A and C come from small towns and rural areas. It is conjectured that the participants who are from cities are likely to get more exposure to English input through daily life (e.g. bilingual advertisements) and

entertainment (e.g. English movies and songs), and thus to pick up more English vocabulary than the participants from small towns and rural areas. Though this is only a conjecture rather than a conclusion led by statistical analysis, the pattern does echo the findings of previous studies of L2 vocabulary learning, where the importance of frequent exposure in the incidental acquisition of vocabulary is identified (Eckerth & Tavakoli, 2012). Nonetheless, this does not explain why their higher vocabulary knowledge was not translated into higher proficiency scores.

Viewed from a different perspective, it is interesting to note that University B participants, who achieved significantly higher mean scores in BPVS than University C participants, also achieved significantly higher NCEEE scores than them. This was not surprising, as vocabulary is generally perceived as an important indicator of L2 English proficiency (Milton, 2013). A series of studies have found vocabulary knowledge strongly correlates to different aspects of language use including reading, writing, listening and speaking (e.g. Stæhr, 2008; Milton, 2010), supporting the Lexical Learning Hypothesis (Ellis, 1997), which argues that vocabulary knowledge is the cornerstone of L2 learning. However, it should also be noted that University A participants achieved significantly higher mean NCEEE scores compared to University B participants, despite the fact that University B participants achieved higher (but not significantly higher) mean BPVS scores. This may reflect the fact that, though the importance of vocabulary knowledge is evident, other factors also contribute to the success of L2 learning, such as motivation, self-efficacy and strategy

use (Gan, Humphreys & Hamp-Lyons, 2004). Moreover, other factors beyond language proficiency per se, such as test taking strategies and preparedness for examinations, might also have contributed to the significantly higher NCEEE scores for participants in University A.

Though the NCEEE and BPVS were baseline tests in this study, the analysis of the results of three universities of different levels provides interesting insights into the relationship between vocabulary knowledge and language proficiency in L2 English. Vocabulary is unarguably an integral part of L2 proficiency. It is natural to assume that a more proficient L2 learner will have a larger vocabulary. However, the relationship between vocabulary knowledge and L2 proficiency may not be this straightforward. A more complex model is required to accurately assess the contribution of vocabulary in L2 proficiency.

## 8.2 Research Question 1

*RQ 1: Does a programme of systematic phonics instruction lead to improvement in Chinese university EFL learners' English decoding?*

This question was intended to probe whether Chinese EFL learners' English decoding proficiency can be enhanced through a systematic phonics instruction programme. This question was addressed using the word attack section of the Woodcock Reading

Mastery Tests, containing 28 pseudo words varying from one single syllable to multiple syllables. Participants' pronunciations were scored at both grapheme level and word level as a point of comparison. Intra- and inter-rater reliability checks were also conducted, which revealed that the scoring of this test was highly reliable.

### 8.2.1 Word-level phonological decoding scores

At t1, both intervention and comparison groups (across all three universities) correctly decoded approximately 10 words out of the 28 pseudo words on average in the decoding test, achieving a mean accuracy of roughly 37%. According to the Woodcock Reading Mastery Tests manual (Woodcock, 2011), this corresponds to the decoding proficiency of a first grade native English student. This echoes the results of the BPVS, as the mean score of the two groups (approximately 56) corresponds to the vocabulary breadth of a five-year old native English student (Dunn & Dunn, 2009). This reveals the close relationship between vocabulary breadth and decoding proficiency, even in a second language. This is intuitively plausible, as EFL learners can draw on their existing vocabulary knowledge to make analogies to known words, in order to work out the pronunciation of unknown English words. This also echoes the findings of research on decoding proficiency in L1 English, as native English speaking students with larger vocabularies also tend to achieve higher scores in English decoding tests (e.g. White, Graves & Slater, 1990). From a different perspective, it is also possible that the relationship runs in the other direction – i.e.

that decoding proficiency directly or indirectly facilitates vocabulary acquisition; or indeed that the relationship is reciprocal.

At t2, the intervention participants (across all three universities) correctly decoded 14 words out of the 28 pseudo words on average (50% accuracy) as opposed to the 10.3 words at t1 (37% accuracy). The comparison groups (across all three universities) correctly decoded 11.3 words on average at t2 (40% accuracy), as opposed to 10.7 words at t1 (38% accuracy). Statistical analysis revealed that the intervention participants made significantly more progress between t1 and t2 compared to the comparison groups, providing an affirmative answer to RQ1.

Though the progress made by intervention participants at word level did not seem to be marked and they still had plenty of room for progress, these results are still encouraging, considering that they made this progress over only 12 weeks. Given that no delayed-post test was conducted, it remains unknown whether the intervention participants' advantage would have been retained a few weeks after completing the phonics programme, or indeed whether this advantage would have continued to increase. However, it is possible that intervention participants would actually demonstrate more progress in a delayed post-test, given that it may take time to internalise the phonics knowledge and achieve consistently correct decoding. Similar observations were also made in Woore (2010, 2011), that L2 learners' decoding firstly gets destabilised as they are not sure how to pronounce unfamiliar words, and

only later do they manage to work out (and eventually automatize) the correct pronunciations.

One thing worth noting is that both intervention and comparison groups scored significantly higher at t2 than at t1, regardless of whether university C was included. In other words, both groups had made significant progress in word-level decoding by t2. This was unexpected, as previous evidence demonstrates that significant progress in foreign language decoding may be rare in the absence of explicit phonics instruction, at least in the context where learners' L1 and L2 are both alphabetic languages (e.g. Woore, 2009, where learners' L1 is English and L2 is French). The reasons for the significant increase in word-level decoding scores for the comparison groups may be threefold.

Firstly, the comparison groups learned the concept of syllables and received some intra-word analysis exercises on word segmenting and syllable counting at the beginning of the phonology instruction programme. This kind of exercise was also carried out every week in the phonics instruction programme, to raise intervention participants' English phonological awareness. Studies of L1 English reading have consistently demonstrated that phonological awareness is a key predictor of literacy development (e.g. August, McCardle & Shanahan, 2014). It is possible that the knowledge of word segmenting and syllable counting helped the comparison groups raise their English phonological awareness, which may have contributed to the

significantly higher word-level decoding scores compared to t1.

Secondly, though the comparison groups did not receive any explicit instruction in English GPCs, it can be argued that some incidental learning may have taken place in the phonology instruction programme. This learning may have been facilitated by the focused input. For instance, when learning the phoneme /i:/, participants were given some examples that contain this phoneme to practice, such as *fee*, *feed*, *feet* and *meet*. The very high prevalence of /i/ = <ee> in this input compared to its occurrence in natural language may have made this GPC more salient, and thus more likely to be 'noticed' (Schmidt, 1990) by the comparison groups as a result.

Thirdly, as mentioned in the previous section, higher decoding proficiency seemed to be associated with more vocabulary knowledge at t1. It is possible that after 12 weeks of intensive English learning, the comparison groups had accumulated more vocabulary knowledge, which could lead to higher decoding scores. However, this is only a hypothesis, as BPVS was only used as a baseline test at t1 and was not administered at t2. This will be further discussed later.

Participants' word-level decoding scores were also analysed by university. One thing worth noting is that, out of the six groups of participants (one intervention and one comparison group in each university), four groups scored significantly higher at t2 than t1 (both intervention and comparison groups in Universities A and B); one group

(intervention group in University C) scored higher at a descriptive level (but not significantly so) at t2 than t1; and one group (comparison group in University C) actually scored lower at t2 than t1. In other words, the performance of both groups in University C was atypical.

Considering that the intervention group in University C was learning French while they received the phonics instruction, the results were understandable, since learning a new foreign language with the same Roman alphabet as English might be expected to influence the effectiveness of the English phonics instruction programme. This would be an instance of cross-linguistic interference ('negative transfer' in the terms of Odlin, 1989) of a third language on the L2. However, it is interesting to note that the comparison groups in University C, who were learning Japanese during the period of the intervention, also failed to increase in English decoding proficiency between t1 and t2. It is possible that the learning of Japanese, a non-alphabetic language, also hindered participants' progress in L2 (English) decoding. This could possibly be explained by the fact that there is interference between learning the GPCs in English and learning the grapheme-syllable correspondences in the hiragana and katakana syllabaries in Japanese. The fact that learning another foreign language – be it an alphabetic language or a non-alphabetic language – might interfere with the development of English decoding proficiency suggests that introducing a new foreign language at the same time as providing instruction in decoding the first foreign language (L2) may not a good idea. Considering the crucial role of decoding in L2

English learning, it can be argued that L3 should ideally be taught after proficient English decoding skills are acquired.

However, learning an L3 may not be the only reason why participants in University C demonstrated difference performance from particiapnts in Universities A and B. As mentioned in Chapter 4.1, participants in University C achieved significantly lower NCEEE scores than those in the other two universities. It is possible that, different levels of English proficiency may have an impact on how much the participants could benefit from the phonics/phonology lessons, especially when considering that the lessons were conducted in English. In addition, other factors beyond language proficiency per se, such as attitude towards the phonics/phonology lessons, and the time spent after the phonics/phonology lessons reviewing and practising English decoding, may also influenced how participants perform in the decoding test. However, given the scope of this study, these factors were not probed into, thus making it difficult to draw any convicing conlusion on why neither the intervention nor the comparison participants in University C made siginificant progress in the decoding test after the phonics/phonology instruction programmes.

### 8.2.2 Grapheme-level decoding scores

Participants' phonological decoding test results were also scored at the grapheme level. Both the intervention and comparison groups (across the sample as a whole, i.e.

all three universities) achieved roughly 68% accuracy at t1, meaning they correctly decoded 80 out of the approximately 120 graphemes. At t2, the intervention participants achieved roughly 83% accuracy (99 graphemes), while the comparison groups achieved roughly 77% accuracy (90 graphemes). Non-parametric tests showed that the grapheme-level decoding scores were significantly higher at t2 than t1 for both groups, but that the intervention participants made significantly more gains than the comparison groups, echoing the results of the analysis of word-level decoding scores; these results were supported by an ANOVA, although the assumptions for this were not fully met. This, from a different perspective, again points to the effectiveness of the phonics instruction programme in promoting L2 learners' English decoding proficiency, providing an affirmative answer to RQ1.

Given that the decoding test was scored at both word level and grapheme level, it is of interest to compare the two sets of results. It can be seen that there is a clear discrepancy between the mean accuracy achieved at the word level and that achieved at the grapheme level at both time points: at t1, both intervention and comparison groups (across all three universities) achieved 37% accuracy in word-level decoding, but 68% accuracy in grapheme-level decoding; at t2, the word-level decoding accuracy rose to 50% for the intervention participants and 40% for the comparison groups (across all three universities), while the grapheme-level decoding accuracy reached 83% for the intervention participants and 77% for the comparison groups (across all three universities). Such a discrepancy can be explained by the fact that

word-level scoring is stricter, in that one wrongly decoded grapheme in a word results in a score of zero.

The discrepancy between word-level and grapheme-level accuracy naturally invites a question, that of which scoring system better reflects decoding proficiency. It can be argued that the binary word-level scoring system is not sufficiently granular, for two reasons. Firstly, if a participant only wrongly decoded one grapheme in a word with multiple graphemes, for instance wrongly decoding the grapheme <i> in the word *cigbet*, their correct understanding of the other graphemes in the word (<c>, <g>, <b>, <e>, <t>) would not be acknowledged. In this regard, grapheme-level scoring is a more sensitive scoring system. Secondly, from a practical point of view, one wrongly decoded grapheme in a word with only two graphemes may not have the same implications for communication as one wrongly decoded grapheme in a word with nine graphemes (if this were a real word). For instance, wrong decoding of the grapheme <ay> in the pseudoword *tay* (such as / taɪ / or /tiː/) would certainly hinder the participants from recognising its phonological form, but omitting the grapheme <f> in the pseudoword *bafmotbem* may not necessarily lead to the same result, as the percentage of graphemes decoded correctly for the latter pseudoword is much higher than for *tay.*

This leads to another question, which is whether different graphemes carry the same communicative value. It is a known fact that some graphemes appear more often in

English than others. For instance, according to Gontijo et al.'s (2003) study of the probabilities of occurrence of different English graphemes based on a corpus of 160,000 word forms, the grapheme <e-e> (as in *theme*) appears in 3.35% of these English words; the grapheme <ay> appears in 0.27% of these English words, and grapheme <au> only appears in 0.04% of these English words. From a practical point of view, it is obviously more important to master the grapheme <e-e> than graphemes <ay> and <au>, because the former is much more likely to appear in English words than the latter two. This was reflected in the phonological decoding test used in this study, as various graphemes were represented by different numbers of tokens. In most cases, graphemes with higher probabilities of occurrence in the language were also represented by more tokens, as strong correlations between the probabilities of occurrence in English and the numbers of tokens in the test were observed for both consonant graphemes and vowel graphemes (consonant graphemes: $\rho = .78$, $p < .001$, vowel graphemes: $\rho = .70$, $p < .001$). Taking the previously mentioned graphemes as examples, <e-e> had three tokens, <ay> had two and <au> had only one. As a result, if a participant did not know how to decode <au>, they would lose one point at the grapheme-level; but if they did not know how to decode <e-e>, they would lose three points. This, in fact, can be argued as a more accurate way of assessing decoding proficiency than having the same number of tokens for each grapheme, as a participant would lose more points if they did not know how to decode a more frequently appearing grapheme. Such differences cannot be reflected by word-level decoding scores. Similar observations were also made in Woore (2009, 2011), where

a grapheme-level scoring system was used to assess participants' L2 French decoding proficiency. By contrast, in most previous studies, only word-level scoring has been administered (e.g. Hamada and Koda, 2011). Hence, it is recommended that a grapheme-level scoring system be used in conjunction with a word-level scoring system in future studies, to provide a more comprehensive and realistic picture of participants' decoding proficiency.

Looking back at the word-level and grapheme-level decoding results, though the intervention participants (across all three universities) only corrected decoded 14 out of the 28 pseudowords at t2, they correctly decoded 99 out of the approximately 120 graphemes, which is a quite promising result. In other words, though they were 14 words short of achieving the full score (50% correct), they only wrongly decoded 21 graphemes out of 120 (83% correct). This suggests that the 'completion rate' of the phonological decoding test was actually quite high at t2, especially considering that the intervention was only 12 weeks long.

It is also worth noting that not only did the intervention participants in University C not make significantly more gains than the comparison group at the word level but they also failed to do so at grapheme level. This again suggests that learning another foreign language compromised the effect of the English phonics instruction programme.

### 8.2.3 Overall time of decoding

Before going on to discuss the findings on the analysis of overall time of decoding, it is worth pointing out that the measure of the overall time of decoding is a very rough, broad-brush measure compared to the millisecond timing that are usually reported for individual items in this area of study (e.g. Coltheart & Rastle, 1994; Ferrand, 2000); hence the results reported in this section need to be interpreted with caution. Nonetheless, the analysis in this section still provides a rough indicator of the overall time spent on processing the pronunciations of the test items, and may help complement the picture of participants' decoding.

The first impression regarding overall decoding time is that both intervention and comparison groups spent less time completing the test at t2 compared to t1, which is understandable, as some degree of decrease in time might be expected to arise through a practice effect – e.g. as the participants were used to the test form the second time around. However, the comparison group seemed to spend much less time than the intervention group at t2. Statistical analysis also confirmed that the comparison groups, who did not follow the programme of phonics instruction, showed a significantly greater decrease in overall decoding time compared to the intervention participants at t2. To address RQ1, these results demonstrate that the phonics instruction programme did not lead to greater improvement in terms of decoding speed than did the phonology instruction. This was not originally hypothesised, as it seemed natural to assume that after the phonics instruction, the

intervention participants would not only achieve higher scores but also become more fluent in decoding. The comparison groups, on the other hand, were hypothesised to spend roughly the same amount of time completing the decoding test at the two time points.

It is interesting to explore why the intervention participants, who made significantly more gains at both word level and grapheme level than the comparison group, actually took a longer time to complete the decoding test. One possible explanation is that the intervention participants had not fully automatized the newly acquired GPCs knowledge, and thus still needed time to consciously work out the pronunciations. This is in line with the controlled versus automatic processing theory proposed by Schneider and Shiffrin (1977). According to the theory, automatic processing 'operate[s] through a relatively permanent set of associative connections… and require[s] an appreciable amount of consistent training to develop fully' (Schneider and Chein, 2003). It seems that the intervention participants had not achieved automatic processing, possibly due to the fact that the phonics instruction lasted only 12 weeks. In contrast, they had to employ a controlled processing mechanism, which was 'activated under control of, and through attention by, the subject' (Schneider and Chein, 2003). This argument can arguably be supported by the finding that the standard deviation for the overall time of decoding was noticeably larger for the intervention participants than the comparison groups at t2, as greater variations would be expected in controlled processing than in automatic processing. Informal

observations during the administration of the decoding test at t2 also supports this argument, as many of the intervention participants actually revised their decoding of a word several times before deciding on the final answer. In contrast, most comparison groups just made one attempt in decoding each word without any further revision. One comparison group, who finished the test almost with no hesitation at all, was asked why he finished so quickly. He commented that 'I don't know how to read the words anyway, so I just said the first thing that came into my mind'. Though it is not certain whether this comment represented the typical thinking of other comparison groups, it is possible that the reason why they finished more quickly at t2 was not because they were more confident about their answers, but because they just wanted to get it over with.

When further examining the data, it transpired that those intervention participants in Universities A and B who spent a longer time completing the decoding test than most others actually produced the best results. For instance, one participant in the intervention group in University A completed the decoding test in 75 seconds at t1 and spent 100 seconds at t2, but her word-level decoding scores rose from 10 to 23 (out of a possible 28), and her grapheme-level decoding results rose from 80 to 108 (out of a possible total of 120 graphemes), which were among the best scores achieved at t2. This is in accordance with the finding in Erler (2003), in which those Year 7 English learners of French who spent the longest time in a timed reading-aloud task, actually produced the best results. Similar findings were also reported in Li

(2012), where those Chinese advanced EFL learners who spent the longest time in an untimed task of reading a list of unfamiliar real English words produced the best pronunciations. This points to the value of using untimed (or at least generously timed) decoding tests, at least in some contexts, as tightly-timed decoding tests might not have allowed participants to consciously work out the pronunciations. Of course, the ultimate aim is for these learners to be able to decode both accurately and fluently, which may not be the case after only 12 weeks of phonics instruction, but may possibly be achieved given more time to practise the instructed GPC knowledge. From a different perspective, the best test to use in terms of measuring the speed of decoding may depend on the participants' decoding proficiency at the time of testing.

### 8.2.4 Summary for RQ1

In summary, the analysis of the decoding test results provided a clear answer to RQ1: the phonics instruction programme did lead to improvement in the accuracy of English decoding, but not in the speed of English decoding, for the university-level Chinese EFL learners. The finding that a 12-week intervention led to such results provides encouraging evidence for the effectiveness of phonics instruction in promoting L2 English decoding accuracy, and could hopefully bring confidence to future L2 English phonics instruction programmes. However, it should be noted that learning an L3 concurrently might compromise the results of L2 English phonics instruction, and thus should not be recommended. It should also be noted that further

research is needed with a wider range of outcome measures, to ensure that the programme of phonics instruction does not come at the expense of progress in other aspects of English learning.

### 8.3 Research Question 2

*RQ2: Is the programme of phonics instruction more effective for some GPCs than others?*

This question was intended to examine the effects of the phonics instruction programme on different GPCs, in order to determine whether certain GPCs are more difficult than others for the target population, and thereby potentially merit particular instructional attention. The question was addressed by comparing participants' accuracy percentages for individual GPCs before and after the instruction. For the purpose of this RQ, only the data of Universities A and B were analysed. The consonant GPCs and the vowel GPCs were analysed separately, based on the fact that English phonemes are conventionally divided into these two articulatory categories (Chomsky & Halle, 1968), and also on the impressionistic observation (during data collection) that participants seemed to encounter more difficulty in the decoding of vowel graphemes than consonant graphemes.

### 8.3.1 Consonant GPCs

The first impression on examining the accuracy percentage of individual consonant GPCs at t1 is that most of these were already realised with high levels of accuracy before the instruction programmes. This is good to observe, as according to the Lingua Franca Core (LFC) proposed by Jenkins (2000, 2007), consonants are extremely important for non-native speakers of English to master. Dauer (2005: 546) also argues that mastering consonants is essential, as they are 'quite stable across all varieties of English'. It is natural to wonder why the participants already had good knowledge of most consonant GPCs at t1. However, as many English consonant graphemes also exist in Pinyin, and many English consonant phonemes are also shared by the Chinese phonological system, two potential factors contributing to the decoding of English consonant GPCs naturally reveal themselves. These are: 1) whether the English grapheme exists in Pinyin; and 2) whether the corresponding phoneme is pronounced similarly in Chinese phonology. Based on this, the consonant GPCs in the test may be divided into the following four categories:

(a) Pinyin-congruent GPCs: namely consonant graphemes that exist in Pinyin and have the same corresponding phonemes in Chinese (or closely similar phonemes; this will be elaborated on in 9.3.1.1). This applies to 20 of the 35 consonant GPCs in the test:

<p> = /p/ , <k> = /k/, <h> = /h/, <d> = /d/, <b> = /b/, <sh> = /ʃ/, <f> = /f/, <l> = /l/, <t> = /t/, <c> = /s/, <j> = /dʒ/, <m> = /m/, <z> = /z/, <r> = /r/, <ch> = /tʃ/, <w> =

/w/, <g> = /g/, <n> = /n/, <s> = /s/ , <y> = /j/

(b) Pinyin-incongruent GPCs: namely consonant graphemes that exist in Pinyin but have different corresponding phonemes in Chinese (3 out of 35 consonant GPCs in the test):

<s> = /z/, <c> = /k/, <x> = /ks/

(c) Pinyin-absent GPCs: namely consonant graphemes that do not exist in Pinyin but whose corresponding phonemes exist in Chinese (9 out of 35 consonant GPCs in the test):

<ed> =/d/, <ss> = /s/, <ff> = /f/, <wh> = /w/, <ck> = /k/, <wr> = /r/, <kn> = /n/, <ng> =/ŋ/, <mb> = /m/

(d) Chinese-absent GPCs: namely consonant graphemes that do not exist in Pinyin and whose corresponding phonemes do not exist in Chinese either (3 out of 35 consonant GPCs in the test):

<th> = /θ/, <qu> = /kw/[55], <dge> = /dʒ/

Based on this categorisation, these four types of consonant GPCs are analysed separately below.

[55] Phonemes /k/ and /w/ both in exist in Chinese, but consonant clusters are not permitted in Chinese; therefore they <qu> = /kw/ are categorised as Chinese-absent GPCs here.

#### 8.3.1.1 Pinyin-congruent GPCs

It can be seen that this category covers more than half of the English consonant GPCs in the decoding test. It is worth noting that not all the consonant graphemes in this category have the same or nearly the same pronunciations as in Chinese. For instance, grapheme <h> is rendered as /h/ in English (a voiceless glottal fricative), and as /x/ in Pinyin (a voiceless velar fricative). However, though the two phonemes are not exactly the same, they are similar (both are voiceless fricatives), and do not resemble any other phonemes in either English or Chinese. As a result, the grapheme <h> is categorised here as a grapheme sharing similar pronunciations in English and Chinese. Other graphemes that are rendered as similar but not identical phonemes in English and Chinese are listed below.

| Grapheme | English pronunciation | Pinyin pronunciation |
|---|---|---|
| <b> | /b/ | /p/ |
| <p> | /p/ | /pʰ/ |
| <d> | /d/ | /t/ |
| <t> | /t/ | /tʰ/ |
| <g> | /g/ | /k/ |
| <k> | /k/ | /kʰ/ |
| <sh> | /ʃ/ | /ʂ/ |
| <ch> | /tʃ/ | /tʂ/ |
| <r> | /r/ | /ɻ/ |
| <z> | /z/ | /ts/ |
| <c> | /s/ | /tsʰ/ |

*Time 1*

The analysis of the accuracy percentages of individual GPCs at t1 revealed that most consonant GPCs in this category were realised with more than 80% accuracy before the instruction programmes. This was expected, as the Contrastive Analysis Hypothesis (Lado, 1957; Ellis, 1994) argues for positive transfer from L1 to L2 when similar properties are shared by the two languages, such as GPCs in this case.

However, it is interesting to note that the facilitative role of L1 in L2 decoding seems to be documented only in learners whose L1 and L2 are typologically similar languages. In contrast, a typologically distant L1 has always been argued to hinder the development of L2 decoding proficiency. On this basis, Koda proposed the Orthographic Distance Effect (2005), arguing that the more similar the L1 and L2 writing systems are, the more easily learners can develop decoding proficiency in L2. Within this theoretical framework, Chinese L1 learners are typically characterised as having a logographic (or morphemic) L1 writing system, which is therefore considered typologically distant from the English writing system (Koda, 2007). However, it is worth noting that though Chinese has a morphemic writing system, its romanized spelling system, Pinyin, uses the same alphabet as English, and shares some of the same (or closely similar) GPCs with English. As a result, a facilitative role for Pinyin in decoding these consonant GPCs can be hypothesised, and this hypothesis is supported by the high accuracy achieved for the consonant GPCs in this category at t1. This would suggests that, in L2 writing system research, it is inadequate simply to characterise all Chinese L1 learners as ‘logographic’, since this

fails to capture an important aspect of their existing literacy experience. Considering that the consonant graphemes in this category share the same (or closely similar) corresponding phonemes in English and Chinese, it is natural to wonder why these GPCs were not realised with nearly 100% accuracy at t1. Some of these GPCs still had quite some room for improvement. For instance, <s> = /s/ only had an accuracy of 60%. The reason for this, as discussed in Chapter 7, is that participants omitted some graphemes in consonant clusters (e.g. <s> in *untroikest*).

*Time 2*

At t2, the intervention participants achieved almost 100% accuracy for most GPCs in this category. The GPCs that were not well decoded at t1 showed clear improvement at t2. For instance, <s> = /s/ reached 99% accuracy for the intervention participants, as compared with 90% accuracy for the comparison groups. This shows that the phonics instruction programme was more effective than the phonology instruction programme in terms of promoting the decoding of the GPCs in this category.

Indeed, the comparison of mean accuracy at t1 and t2 reveals that some of the GPCs in this category showed the most progress of all the consonant graphemes for the intervention participants; these included <g> = /g/ (30 percentage points of progress), <n> = /n/ (25 percentage points of progress), <s>= /s/ (35 percentage points of progress). When looking at the words containing these graphemes (e.g. *monglustamer*,

*untroikest*, *ceisminadolt*), it can be seen that these graphemes are even found in consonant clusters in which consonants were omitted or pronounced in the wrong order at t1. It may therefore be argued that the marked improvement in accuracy percentages for these GPCs indicates the effectiveness of the phonics instruction programme in terms of promoting participants' intra-word analysis.

#### 8.3.1.2 Pinyin-incongruent GPCs

The three consonant GPCs in this category are <s> = /z/, <c> = /k/ and <x> = /ks/.

*Time 1*

Examination of the accuracy percentages at t1 reveals that both intervention and comparison groups achieved more than 90% accuracy in decoding <s> = /z/. The reason why <s> = /z/ demonstrated high accuracy is probably because the two pseudo words containing this GPC (*dud's*, *ful's*) are both morphology-related, and the pronunciation rules of the possessive <s> are covered by the high school syllabus in China. This may be why most participants accurately decoded this grapheme at t1, even though the grapheme <s> is pronounced as /s/ in Pinyin, which is also a possible pronunciation of this grapheme in English.

In contrast, <c> = /k/ showed lower accuracy (approximately 70% for both

intervention and comparison groups), because some participants failed to distinguish it from <c> = /s/ and mistakenly pronounced <c> as /s/ in the word *byrcal*.

Finally, the <x> = /ks/ was among those demonstrating the lowest accuracy at t1 (approximately 30% for both intervention and comparison groups), which was expected, as its Pinyin pronunciation /ɕ/ does not exist in English and its canonical English pronunciation /ks/ does not exist in Chinese.

*Time 2*

The comparison between the accuracy percentages at t1 and t2 reveals that the intervention participants made improvement in decoding all three consonant GPCs in this category. <s> = /z/ witnessed almost 100% accuracy at t2. <c>= /k/ also saw some improvement at t2, and reached an accuracy of 80%. In contrast, the comparison groups achieved similar accuracy at the two time points.

However, it can be seen that <x> = /ks/ saw only very little progress after the phonics instruction and remained one of the most poorly decoded consonant GPCs (with an accuracy of 40% at t2). One possible explanation is that words containing the grapheme <x> are limited in number, with only 1.9% of English words containing this grapheme. As a result, the participants were not given enough chance to practise this grapheme. Pinyin GPCs also seemed to influence the decoding of this grapheme, as some participants decoded the grapheme <x> as /s/, which sounds similar to its

Pinyin pronunciation /ɛ/. This echoes Woore's (2014:170) conclusion that 'automatically-activated L1 processing mechanisms may also be disruptive'.

#### 8.3.1.3 Pinyin-absent GPCs

The nine consonant GPCs in category are <ed> =/d/, <ss> = /s/, <ff> = /f/, <wh> = /w/, <ck> = /k/, <wr> = /r/, <kn> = /n/, <ng> = /ŋ/, <mb> = /m/.

*Time 1*

Though all the graphemes in this category have corresponding phonemes in Chinese, their accuracy varied greatly at t1. The GPCs <ed> = /d/, <ss> = /s/ and <ff> = /f/ were generally well decoded at t1, with an accuracy of over 80%. In contrast, other GPCs in this category were relatively poorly decoded at t1.

One common problem in decoding these graphemes was that many participants failed to recognise them as single graphemes and treated them as two separate consonants, for example decoding the grapheme <kn> in *knaf* as /kən/, and decoding <wh> in *whumb* as /wəh/, as shown in section 7.1. This is because <kn> and <wh> are not Pinyin graphemes, while <k> <n> <w> and <h> are Pinyin graphemes. Unfamiliar with the English grapheme system, many participants wrongly perceived these graphemes as two separate graphemes. As consonant clusters are not permitted in

Pinyin syllables, many participants may also have added a vowel between them in decoding for this reason (often an epenthetic schwa).

*Time 2*

After the phonics instruction, intervention participants showed clear improvement in decoding all the consonant GPCs in this category, except for <mb> = /m/. Their advantage over the comparison groups in decoding these graphemes was very clear at time 2. For instance, intervention participants made marked progress of 40 percentage points in decoding <kn> as /n/, while comparison groups made little progress, improving by less than 10 percentage points. This suggests that the phonics instruction appears to be effective in promoting the knowledge of Pinyin-absent GPCs. This is easy to understand. As the phonics instruction provided systematic training in common English graphemes, intervention participants learned how to correctly segment English words by graphemes, rather than falling back on the Pinyin system, which could be disruptive. The only exception was <mb> = /m/, for which both groups of participants showed little progress. One possible explanation is that words containing this GPC are limited in number both inside and outside the intervention; as a result, the participants were not given enough chance to practise this GPC.

An interesting observation is that Pinyin-absent GPCs seemed to show more progress than Pinyin-incongruent GPCs, perhaps indicating that establishing the link between

an unfamiliar English grapheme and a familiar Chinese phoneme is somewhat easier than establishing the link between an unfamiliar Pinyin grapheme and a familiar English phoneme. This again provides evidence that L1 processing mechanisms can be a double-edged sword in L2 decoding. If the phonemes already exist in L1, phonics instruction may be helpful in mapping these phonemes onto the new graphemes in L2. However, if the graphemes already exist in L1, automatically-activated L1 processing mechanisms may be difficult to combat even after the phonics instruction, leading to somewhat L1-resembling pronunciations.

#### 8.3.1.4 Chinese-absent GPCs

The three consonant GPCs in this category are <th> = /θ/, <qu> = /kw/, <dge> = /dʒ/.

*Time 1*

As these graphemes do not exist in Pinyin and their corresponding phonemes do not exist in Chinese either, it is not surprising that these graphemes were all poorly decoded at t1, as participants had no existing knowledge base on which to draw when decoding them. At t1, both intervention and comparison groups only achieved approximately 20% accuracy in decoding graphemes <qu> and <dge>, and approximately 30% accuracy in decoding grapheme <th>.

*Time 2*

After the phonics instruction programme, it can be seen that intervention participants made considerable progress in decoding graphemes <qu> and <dge>, with the accuracy increasing from 20% to 70%, compared to 30% for the comparison group. By contrast, the grapheme <th> seemed to resist the effects of instruction, with both groups of participants showing very limited progress (40% accuracy for the intervention group, and 35% accuracy for the comparison group). This might be because the sound /θ/ is difficult to pronounce for Chinese students, as evidenced by many previous studies (e.g. Rau, Chang & Tarone, 2009). As a result, many participants still fell back on their Chinese phoneme inventory and decoded grapheme <th> as the similar-sounding /s/.

The marked progress on the graphemes <qu> and <dge> seems to suggest that the phonics instruction programme was especially effective in promoting the knowledge of Chinese-absent GPCs. This is understandable, as participants could learn these GPCs 'from scratch'; that is to say, they are less likely to make false analogies to their repertoire of L1 GPCs, and are more likely to establish new GPCs.

It is interesting to note that the comparison groups, who received systematic phonology instruction, also made visible progress in two of these GPCs. This suggests that familiarising learners with the English phoneme system can be beneficial to the

decoding of some English graphemes, even without the explicit phonics instruction. This is presumably because for some of these graphemes, the challenge may lie in having an acceptable phoneme in place onto which the graphemes can be mapped, and a systematic phonology instruction programme can pave the way for that.

### 8.3.2 Vowel GPCs

Compared to the consonant graphemes, most of which were already pronounced with high accuracy at t1, the vowel graphemes were decoded less accurately at t1. This is as expected, based on two factors. Firstly, pinyin vowel graphemes and phonemes have strict one-to-one mappings (Lin, McBride-Chang, Shu, Zhang, Li, Zhang & Levin, 2010). In other words, each vowel grapheme in pinyin has one single pronunciation, and each vowel phoneme has one single written representation. By contrast, many English vowel graphemes are decoded differently according to their orthographic context (e.g. compare the realisations of <ou> in the English words *out, touch, four, route, enormous,* cited in Gontijo et al., 2003). Needless to say, this makes the decoding of English vowel graphemes challenging for Chinese students. Secondly, many English vowel graphemes are also Pinyin graphemes, which complicates the picture even more. As discussed in the previous section, L1 processing mechanisms are often automatically activated in decoding, which can lead to problems. Based on these two factors, the vowel graphemes in the decoding test are divided into three categories:

(a) Pinyin-incongruent graphemes with only one realisation, namely those vowel graphemes that exist in Pinyin, whose English realisation is different from the Pinyin one, and which are associated with only one English phoneme in the phonics instruction programme (4 out of 28 vowel GPCs in the test):

<ie> = /aɪ/, <ai> = /eɪ/, <ei> = /iː/, <ou> = /aʊ/

(b) Pinyin-incongruent graphemes with multiple realisations, namely those vowel graphemes that exist in Pinyin, whose English realization is different from the Pinyin one, and which are associated with more than one English phoneme in the phonics instruction programme (10 out of 28 vowel GPCs in the test):

<e> = /ə/; <e> =/iː/, <o-e> = /əʊ/; <a> = /ə/; <u> = /ə/, <o> = /ɒ/, <i> = /ɪ/, <u> = /ʌ/, <e> = /e/, <a> =/æ/

(c) Pinyin-absent graphemes with only one realisation, namely those vowel graphemes that do not exist in Pinyin with only one corresponding phoneme instructed in the phonics instruction programme (14 out of 28 vowel GPCs in the test):

<a-e> = /eɪ/, <u-e> = /uː/, <i-e> = /aɪ/, <ir> = /ɜː/, <ey>= /eɪ/, <au> = /ɔː/, <ar> = /ɑː/; <ow> = /əʊ/, <ay> = /eɪ/; <oi> = /ɒɪ/, <er> = /ə/; <ea> = /iː/; <ee>= /iː/; <y>= /iː/

It is worth mentioning that in categories (a) and (c), ‘graphemes with only one

realisation' does not necessarily mean that those graphemes only have one possible realisation in the English language. Here 'graphemes with only one realisation' indicates that only one realisation was covered in the phonics instruction programme in the current study, with other possible realisations interpreted as 'exceptions'. As discussed in Chapter 4, given that English has a relatively opaque orthography (Koda, 2007), the phonics instruction programme did not attempt to (and could not possibly) cover all GPCs in English. As a compromise, for some graphemes with multiple possible realisations, only the most common realisation was instructed.

Based on this categorisation, the three types of vowel GPCs are analysed separately below.

#### 8.3.2.1 Pinyin-incongruent graphemes with only one realisation

The four GPCs in this category are <ie> = /aɪ/, <ai> = /eɪ/, <ei> = /iː/, <ou> = /aʊ/.

*Time 1*

The analysis reveals that the vowel GPCs in this category were all among the most poorly decoded vowel GPCs at t1. For both intervention and comparison groups, <ei>= /iː/ had an accuracy percentage of 50%, while <ie>= /aɪ/, <ai>= /eɪ/ and <ou>= /aʊ/ only had an accuracy percentage of roughly 30%.

As revealed in section 7.2.3, many participants decoded the graphemes in the same way as in Pinyin (e.g. pronouncing <ai> in *laip* as /ai/). This is similar to the finding in Woore (2014), where the majority of English participants pronounced the French grapheme <ée> as /i/, probably triggered by their English GPCs knowledge that <ee> is pronounced as /i:/. This, again, shows that the L1 processing mechanisms may result in errors in L2 decoding.

*Time 2*

After the phonics instruction, these GPCs only showed limited improvement for the intervention participants, with the accuracy of <ei>= /i:/ rising from 50% to roughly 65%, the accuracy of <ou>= /aʊ/ rising from 30% to roughly 40%, while the accuracy of <ie>= /aɪ/, <ai>= /eɪ/ remained roughly the same. The comparison groups achieved similar accuracy in all four GPCs at the two time points. These results suggest that Pinyin-incongruent GPCs, even when they had a consistent realisation, seemed to resist the effects of the phonics instruction. This was not originally hypothesized. Though these graphemes also exist in Pinyin, their English correspondence rules were clear in the phonics instruction programme (even though, as noted above, the real picture was more complicated than what was covered in the instruction programme). therefore, one might have expected them to be highly teachable.

Taking a look at the pseudo words containing these graphemes (*whie*, *laip*, *gouch*,

*ceisminadolt*), it can be seen that three out of the four graphemes are highly congruent with Pinyin syllables: both *lai* and *gou* are valid Pinyin syllables; though *whie* is not a valid Pinyin syllable, the structure of the word (onset + rime) is the same as Pinyin. Muljani et al. (1998) have demonstrated that spelling patterns have an important impact on L2 word recognition, where L2 spelling patterns which are congruent with those found in the L1 are more likely to activate L1 processing mechanisms. This might explain why the phonics instruction had little effect on these three GPCs (<ie>= /aɪ/, <ai>= /eɪ/ and <ou>= /aʊ/), as the deeply-embedded L1 processing mechanisms are not easily overcome, especially when it comes to L2 spellings that are congruent with L1. In contrast, <ei>= /i:/ showed more improvement than these three GPCs, probably because *cei* is not a valid Pinyin syllable, and *ceisminadolt* is a multiple-syllable word that does not resemble the spelling patterns of Pinyin.

It is interesting to note that in Muljani et al. (1998), Chinese participants (defined as logographic readers in that study) were found to be less influenced by incongruent spelling forms than the Indonesian participants (defined as alphabetic readers), as 'the Chinese readers were accustomed to processing primarily logographic characters in their L1' (p. 109). The results in the current study, on the other hand, show that it might be inaccurate simply to characterize Chinese students as logographic readers, without previous experience of an alphabetic writing system before learning English. Their Pinyin knowledge seems to have an impact on their decoding of unfamiliar English words which share the same graphemes as Pinyin.

#### 8.3.2.2 Pinyin-incongruent graphemes with multiple realisations

The ten GPCs in this category are: <e> = /ə/; <e> =/iː/, <o-e> = /əʊ/; <a> = /ə/; <u> = /ə/, <o> = /ɒ/, <i> = /ɪ/, <u> = /ʌ/, <e> = /e/, <a> = /æ/.

*Time 1*

At t1, the Pinyin-incongruent graphemes with inconsistent realisations, namely <a>, <e>, <i>, <o> and <u>, were realised with varying degrees of accuracy at t1. It is worth noting that <e> = /ə/, <a> = /ə/ and <u> = /ə/ – in other words, when <e> <a> and <u> correspond to schwa – demonstrated lower accuracy than their other phonemic realisations for both intervention and comparison groups, as shown in section 5.2.1. For instance, both intervention and comparison groups achieved approximately 50% accuracy when grapheme <u> is decoded as /ʌ/, but only 30% accuracy when it is decoded as /ə/. This suggests that the schwa was a common challenge for the participants at t1.

*Time 2*

After the phonics instruction, the intervention participants made moderate progress in all the possible phonemic realisations of these graphemes except for <e> = /ə/, <a> = /ə/ and <u> = /ə/. The comparison groups made remained roughly the same across all the GPCs in this category.

The lack of improvement in correctly realising the schwa is understandable. A possible explanation is that the rules determining when these graphemes should be decoded as schwa are not very clear. In the phonics instruction programme, there were no sessions dedicated to the explanation of which graphemes should be pronounced as the schwa, and in what cases; instead, schwa was only briefly covered as part of the instruction of common prefixes (un-) and word endings (-el, -al), where participants were taught that graphemes <u> <e> and <a> in these cases should be pronounced as schwa. Clearly, these are not the only cases when graphemes <u> <e> and <a> should be decoded as schwa. The lack of clear and easy-to-state rules, undoubtedly, could lead to the lack of improvement in these cases. However, this might not be the only reason why the schwa was poorly realised even after the phonics instruction. As schwa is essentially an unaccented vowel (Henry, 1989), correctly realising it requires knowledge of lexical stress. This, regrettably, was not covered in the phonics instruction programme.

In contrast to the realisation of schwa, intervention participants made good progress in other phonemic realisations of the graphemes in this category. One possible reason is that the rules are clear and easy to state in these cases. For instance, participants were clearly taught that <a> should be decoded as /æ/ in closed syllables; and that <u> should be decoded as /ʌ/ in closed syllables. Straightforward rules such as these seemed to make the instruction more effective.

Another interesting observation is that, though graphemes in this category have multiple phonemic realisations, they were actually better decoded than the Pinyin-incongruent graphemes with only one realisation at both t1 and t2, and also showed more progress after the phonics instruction programme. It is natural to wonder why this was the case, considering that both the two types of graphemes also exist in Pinyin. One possible explanation is that the Pinyin-incongruent vowel graphemes are so common that they can be seen almost every English word (Gontijo et al., 2003). Moreover, they are sometimes pronounced differently in the same word (e.g. the two realizations of <a> in *animal*). The frequent exposure to these graphemes might to some degree curb the tendency to fall back on L1 processing mechanisms. However, this remains a conjecture, as how participants achieved these pronunciations is unclear. In order to get a better idea of why participants decoded certain graphemes in certain ways, self-report techniques might be useful; these were beyond the scope of the current study.

#### 8.3.2.3 Pinyin-absent graphemes with consistent realisations

The 14 GPCs in this category are: <a-e> = /eɪ/, <u-e> = /uː/, <i-e> = /aɪ/, <ir> = /ɜː/, <ey> = /eɪ/, <au> = /ɔː/, <ar> = /ɑː/; <ow> = /əʊ/, <ay> = /eɪ/; <oi> = /ɒɪ/, <er> =/ə/; <ea> = /iː/; <ee> = /iː/; <y> = /iː/.

*Time 1*

The GPCs in this category were realised with varying accuracy at t1, and the two groups of participants achieved similar results across them. As the graphemes in this category do not exist in Pinyin, they were hypothesized to be the graphemes that would respond the best to the phonics instruction programme, since there would be no interference from L1-entrenched representations or processing mechanisms.

*Time 2*

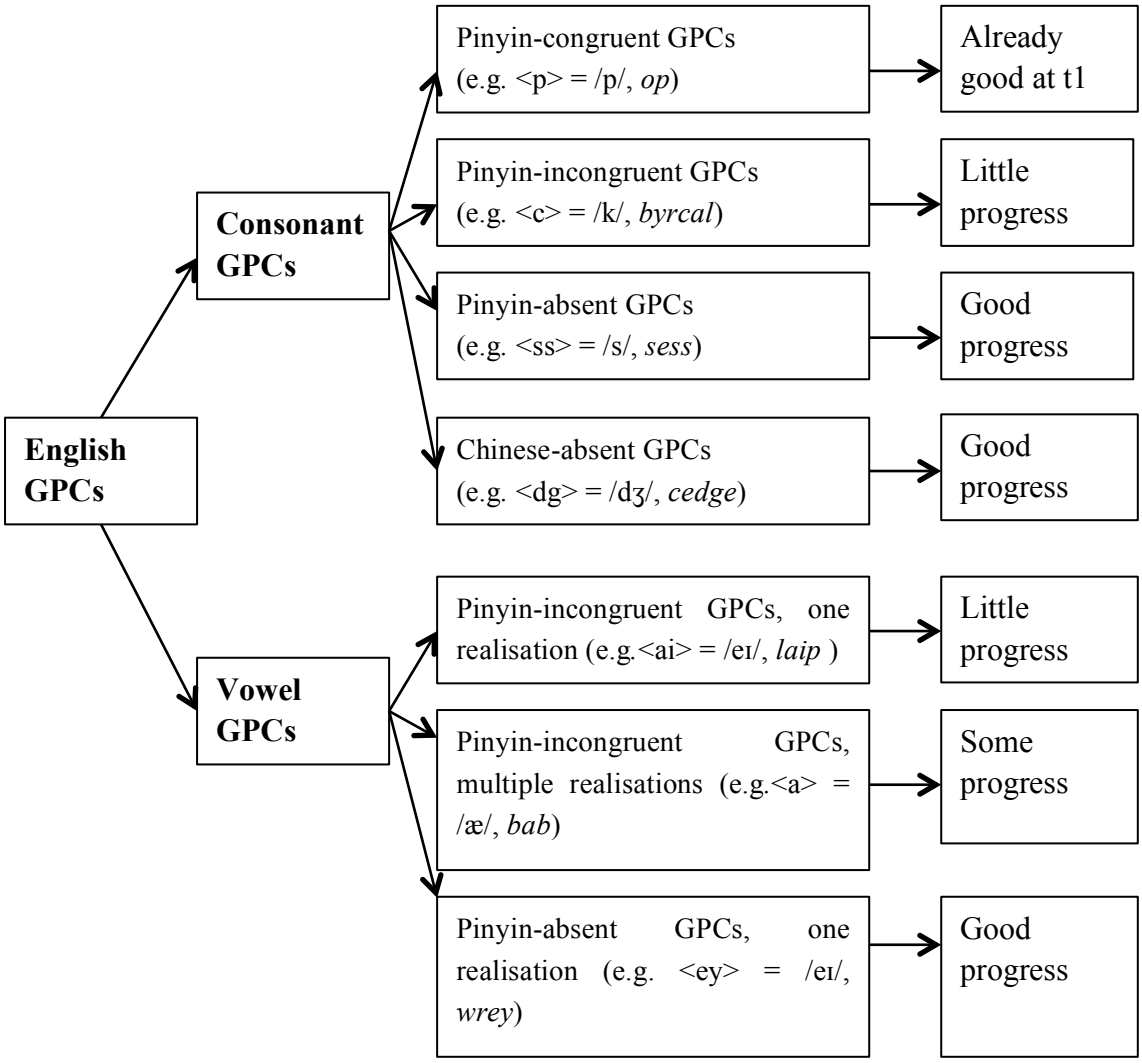
At t2, the comparison groups achieved similar accuracy in these GPCs as at t1. In contrast, some of the GPCs in this category are the ones which showed the most substantial improvement for intervention participants, such as <ey> = /eɪ/ (40 percentage points of progress), <ay> = /eɪ/ and <ir> = /ɜ:/ (30 percentage points of progress in each case). The progress is especially evident when comparing the GPCs in this category to some other GPCs with similar accuracy at t1. For instance, <ey> = /eɪ/ and <ai> = /eɪ/ showed similar accuracy at t1, but after the phonics instruction, intervention participants made 40 percentage points of progress in decoding <ey> = /eɪ/, but less than 10 percentage points of progress in decoding <ai> = /eɪ/. This is again consistent with the view that automatically-activated L1 processing mechanisms can be disruptive in L2 decoding, as <ey> is not a Pinyin grapheme but <ai> is.

### 8.3.3 Phonics instruction- rethinking the role of L1 for Chinese students

Section 8.3 provides a clear affirmative answer to RQ2*: Is the programme of phonics instruction more effective for some GPCs than others?*

Based on the analysis above, the following diagram is tentatively suggested in order to hypothesise the effects of the phonics instruction on different categories of English GPCs according to their relationship to pinyin, as shown in Figure 8.1. The grapheme type (as previously discussed) with sample grapheme, correct phonemic realisation and sample pseudoword from the decoding test and their response to the phonics instruction are tentatively summarised below:

*Figure 8.1 Effects of the phonics instruction on different categories of English GPCs*




This diagram indicates the two types of GPCs that seem to resist the effects of the phonics instruction most stubbornly: Pinyin-incongruent consonant GPCs; and Pinyin-incongruent vowel GPCs with only one realisation.

After analysing the effects of the phonics instruction on different categories of English GPCs, overall the evidence is consistent with the hypothesis that Pinyin knowledge plays an important role in influencing participants' decoding of unfamiliar

English words. Pinyin knowledge may be facilitative when the same (or similar) GPCs are shared by Pinyin and English, but disruptive under other circumstances. This recalls the Contrastive Analysis Hypothesis (Lado, 1957; Ellis, 1994), as supported by a robust line of studies on transfer effects between typologically similar L1 and L2 writing systems (e.g. Woore, 2014; Commissaire et al., 2011). However, this kind of study has seldom looked at the possibility of transfer from Pinyin to English, as Chinese students are generally conceptualised as logographic readers (Muljani et al., 1998). The results in this study have demonstrated that the true picture is probably more complex. As discussed in chapter 2, with the prevalence of mobile phones and computers, Pinyin has become a writing system that Chinese people use daily, rather than merely a tool for young children to learn Chinese characters (which they later discard as they become fluent adult readers). Because Pinyin uses the same Roman alphabet as English, it is therefore of great importance to consider the possible transfer effects of Pinyin in designing future phonics instruction programmes.

## 8.4 Research Question 3

*RQ3: What are the features and problems of Chinese university EFL learners' English decoding at each time point?*

This question was intended to explore the common problems of participants' decoding before the phonics/phonology instruction, and examine to what extent these

problems were solved as a result of the instruction programmes. In line with the previous section, the errors committed in decoding the consonant graphemes and the vowel graphemes will be examined separately. In addition, other errors beyond individual graphemes will also be analysed. In line with the previous section, only the data of Universities A and B are presented in this section.

### 8.4.1 Consonant graphemes

Three types of common errors were observed in participants' decoding of consonant graphemes at t1, namely epenthesis, omission and approximation.

#### 8.4.1.1 Epenthesis

Both intervention and comparison groups made some epenthesis errors at t1, which can be further divided into two categories, namely the epenthesis of a phoneme in an individual consonant digraph/ trigraph (<kn>, <wr>, <qu>, <wh> and <dge>), and the epenthesis of a phoneme in consonant strings (<tw>, <str>).

The reason why many participants inserted a phoneme in an individual consonant digraph/ trigraph is probably because these graphemes do not exist in Pinyin. As a result, many participants failed to recognise these graphemes as a single grapheme and instead, perceived them as two individual consonant graphemes. A piece of

evidence supporting this argument is that, at t1, none of the participants inserted a phoneme in graphemes <sh> <ch> and <ng>, which also exist in Pinyin. At t2, the intervention participants made no epenthesis errors in decoding the graphemes <wr> and <wh>, and clearly fewer errors in decoding the graphemes <kn>, <qu> and <dge>, indicating the effectiveness of the phonics instruction in combating this type of error. In contrast, comparison group participants made similar numbers of epenthesis errors in individual consonant digraphs/ trigraph at the two time points. This tentatively suggests that explicit instruction is effective in teaching these graphemes to Chinese students familiar with Pinyin.

Similarly, participants also made some epenthesis errors in consonant strings. One probable reason is also the influence of Pinyin, as there are no consonant strings in Pinyin. This recalls the findings of many speech analysis studies, where the epenthesis of a vowel, normally the schwa, is found to be a common feature in Chinese EFL learners' pronunciation of syllable codas (defined as 'the closing segment of the syllable (Crystal, 1997: 315)). For instance, Weinberger (1987) has found that Chinese speakers' production of codas at the word-final position was modified through epenthesis in roughly 10 percent of the cases. Similarly, Hansen (2001) has also found that the epenthesis of a vowel is the most common error in Chinese students' pronunciation of two-member codas. The findings of this study, where participants were asked to pronounce pseudo-words, are similar to the previous speech analysis studies, suggesting that the insertion of a vowel in consonant clusters

is a common problem in Chinese learners' English pronunciation. The only difference is that in this study, the epenthesis of a vowel in consonant strings was not only found at the word-final position (pronouncing <nk> in *knoink* as /nɪk/) or the between-syllable position (pronouncing <ft> in *bufty* as /fət/ or /fi:t/), but also found in word-initial position (e.g. pronouncing <tw> in *twem* as /təw/).

At t2, both intervention and comparison groups made fewer errors in decoding the consonant strings. Interestingly, the intervention and comparison groups made similar numbers of errors at the two time points, suggesting that the phonics instruction and the phonology instruction were equally useful in combating insertion errors in consonant strings. Even though explicit instruction on GPCs was not included in the phonology instruction, the familiarity with the English phonological system enabled the participants to understand that combining several consonants together is a common feature of English words, thus deterring them from falling back on the Pinyin phonology system. In addition, the teaching of the syllable structure in the phonology programme also potentially contributed to the decrease of vowel epenthesis at the between-syllable position, as the boundary between the syllables became clear.

It is interesting to explore why the phonology instruction was only useful in combating epenthesis errors in consonant strings, but not in individual consonant digraphs/ trigraph. This is because the origins of the two types of errors, though both

related to participants' L1, are not exactly the same. The epenthesis errors in consonant strings result from the phonotactic constraints of participants' native language, and a phonology instruction programme can help them better understand the English phonological system, which in turn leads to better decoding. The epenthesis errors in individual consonant digraphs/ trigraph on the other hand, result from failing to recognise certain English graphemes that do not exist in Pinyin, and this may only be treatable in an instruction programme teaching English graphemes. For instance, for a digraph like <wr>, learners might initially add an epenthetic vowel; but learning not to add this epenthetic vowel is not enough – they need to know that <wr> is a single grapheme with the pronunciation /r/. By contrast, in a consonant string like <tw>, simply avoiding the epenthesis would give the correct pronunciations. This, however, does not undermine the value of the phonology instruction programme. Familiarising learners with the English phonology system is still crucial in promoting their English decoding – it just might not be sufficient in itself.

#### 8.4.1.2 Omission

Another commonly observed error at t1 was that some consonant graphemes were omitted in participants' decoding. This is especially true in long words with multiple syllables, such as *bafmotbem* and *monglustamer*.

An interesting observation is that the omitted consonant graphemes are always at the start of a consonant string, while the consonant grapheme right before a vowel was always pronounced. This is probably because the consonant grapheme right before a vowel was perceived together with the vowel, which is similar to the onset + rime structure of Chinese syllables. The consonant graphemes after a vowel are frequently omitted, probably because Chinese syllables always end in a vowel.

Though the role of L1 in these errors is only a conjecture, previous speech analysis studies did find that Chinese EFL learners often omitted syllable codas in their English pronunciations. As previously mentioned, Weinberger (1987) found that Chinese EFL learners omitted approximately 10 % of codas in word-final position in their English pronunciation. Broselow, Chen & Wang (1998) also observed that the omission of syllable codas is a common pronunciation feature of Chinese students' English speech. The overlap between the findings of this study and the previous speech analysis studies is thought-provoking. If Chinese EFL learners have the problem of consonant omission in their spontaneous English speech, it is possible that the main reason leading to their omission of some consonant graphemes in their decoding is not the unfamiliarity with certain graphemes, but instead, problems with L2 phonotactics. In order to treat this problem, the most effective way might not be to teach GPCs, but to familiarise students with English phonotactics. This might explain the result that both the intervention and comparison groups made fewer errors in this category at t2.

#### 8.4.1.3 Approximation

Another type of error was approximation, defined as wrongly pronouncing a consonant grapheme as another grapheme which resembles the target grapheme either in form or in pronunciation. For instance, participants decoded the grapheme <mb> as /b/ and <wr> as /w/; and some decoded the grapheme <v> as /w/ and <th> as /s/.

After the phonics instruction, intervention participants made some progress in decoding the graphemes <v> and <wr>, but little progress in decoding the graphemes <th>, <x> and <mb>. This is probably because there are not many words containing the graphemes <x> and <mb>; as a result, the participants lacked opportunities to practise these graphemes. The picture for grapheme <th> is more complicated. As has been demonstrated in section 8.3.1.4, the effect of instruction on the pronunciation of this grapheme was limited for the intervention participants, as more than half of them still pronounced it as /s/ at t2, which was a similar result compared with the comparison groups. The problem of pronouncing the phoneme /θ/ for Chinese speakers has also been observed in many previous studies. For instance, Deterding (2010) observed that Chinese EFL learners tend to pronounce the phoneme /θ/ as /s/ even in their spontaneous English speech. It is therefore worth considering the possibility that the poor decoding of the grapheme <th> as /s/ is not a problem of decoding *per se*, but rather problem of production. Taking a step further, it might also be of interest to explore whether the participants can correctly perceive the phoneme /θ/. As /θ/ does not have a near equivalent in the Chinese phonological inventory,

participants might unconsciously classify it perceptually as the closest fricative they have, which is /s/.

If this is the case, it is not surprising that they pronounce it as /s/ as well. This recalls the Native Language Magnet Theory proposed by Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola and Nelson (2007), which argues that L1 learners develop a 'sound map' in their brain which 'functions like a magnet for other sounds' (Kuhl, 2000: 11855), making the acquisition of similar-sounding L2 phonemes somewhat difficult. Similar point is also made in Flege (1995)'s Speech Learning Model, which posits that 'the greater the perceived phonetic dissimilarity between an L2 speech sound and the closet L1 sound is, the more likely learners will be able to discern the difference between the L1 and L2 sounds and show measurable progress in production and/or perception' (Aoyama, Fledge, Guion, Akahane-Yamada & Yamada, 2004). Future work might include separate tests of whether participants can hear the difference between *thing* and *sing*; whether the same participants can produce this distinction in spontaneous speech; and finally, whether they can read *thing* and *sing* aloud accurately.

Another question naturally arises is that, whether it is important to pronounce particular sounds accurately. In the case of *thing* and *sing*, wrongly pronouncing /θ/ as /s/ would impede intelligibility, but in the case of *tooth*, wrongly pronouncing /θ/ as /s/ may not lead to comprehension problem. Similarly, there are many L2 speakers

who speak with a strong foreign accent but can nonetheless communicate effectively. Though this may be true, a phonics instruction programme should still 'set the bar high' and encourage correct, near-native pronunciations of L2 phonemes. However, in testing decoding and pronunciation, a somewhat lenient approach may be adopted to permit certain degrees of foreign colouring, as long as comprehensibility is ensured.

Looking beyond these error categories, it can be seen that many errors in decoding the consonant graphemes are probably attributable to the phonotactic constraints of participants' L1. As a result, familiarising them with the English phonology system seems to be the key to solving these problems, which can also explain why the comparison groups also made fewer errors in the decoding of many consonant graphemes after the phonology instruction programme.

### 8.4.2 Vowel graphemes

The analysis of the decoding test results demonstrated that the participants made different errors in decoding different types of vowel graphemes. As a result, the vowel graphemes were analysed based on four categories, namely split digraphs, vowel graphemes with multiple possible realisations, vowel digraphs that exist in Pinyin and vowel digraphs that do not exist in Pinyin.

#### 8.4.2.1 Split digraphs

At t1, many participants wrongly sounded out the silent <e> in split diagraphs. This is probably due to the influence of Pinyin, as the grapheme <e> is a common vowel grapheme in Pinyin and is always pronounced. Moreover, as Pinyin syllables always end in a vowel, it is not surprising that many participants sounded out the silent <e> in word-final position. The most common error is pronouncing the silent <e> as /ə/, a similar sound to its Pinyin pronunciation /ɤ/, which seems to confirm the conjecture of Pinyin influence.

Taking a further look at the errors, it can be seen that this type of error seems to originate from the incorrect segmentation of syllables. For instance, decoding *rejune* as either /rɪdʒuːnə/ or /redʒuniː/ suggests that participants segmented the word into three syllables (*re-ju-ne*), rather than two syllables (*re-june*) as it is supposed to be.

This may be due to the use of different grain sizes (Ziegler & Goswami, 2005) between readers of English and Pinyin. Pinyin is a shallow orthography with regular GPCs (Liow et al., 1998); therefore, it is efficient to decode Pinyin strings grapheme by grapheme. English, on the other hand, is a much deeper orthography, where larger spelling bodies need to be processed together in order to generate reliable decoding. The *rejune* example shows that, at t1, many participants processed the word using the small grain sizes that would be needed to process Pinyin, resulting in wrong pronunciation. This echoes Odlin's (1989) argument that moving from a shallower orthography to a deeper orthography is more difficult than in the opposite direction,

as larger chucks of information need to be processed as units.

At t2, the intervention participants almost never sounded out the silent <e> in split digraphs, indicating that they no longer had problems in seeing the split digraph as a single grapheme. This suggests that the phonics instruction was useful in helping the participants moving from the smaller grain size to larger grain size when processing English words, which is arguably necessary for processing split digraphs. In contrast, the comparison groups still had problems in decoding split digraphs as a single grapheme at t2, suggesting that the phonology instruction programme was not enough to treat this problem.

It is also worth noting that the intervention participants produced more so-called 'wild forms' at t2 than t1, such as decoding *rejune* as /rɪdʒyːn/ and /redʒyːn/ (where the final vowel is pronounced /y/, which reflects neither an English nor a Pinyin decoding of the grapheme <u> (or <u-e>). The comparison groups also produced slightly more 'wild forms' at t2 than t1.

Even though these pronunciations are still not correct, they could be deemed 'better' wrong answers compared to /rɪdʒuːnə/ or /redʒuniː/, as the former pronunciations are based on the correct syllable segmentation of the word. Further, they may indicate that participants are moving beyond the automatic application of L1-based decoding processes, even though they have not yet established the correct L2 GPC.

From a different perspective, echoing the argument made in Woore (2011) and Woore (2014), this also shows the importance of qualitative analysis in L2 decoding studies, as both /rɪdʒyːn/ and /rɪdʒuːnə/ would be treated as equivalently wrong answers in quantitative analysis; however, moving from /rɪdʒuːnə/ to /rɪdʒyːn/ may show progress in the development of L2 decoding, which should not be ignored.

#### 8.4.2.2 Vowel graphemes with multiple realisations

Many errors were observed in participants' decoding of the vowel graphemes <a> <e> <i> <o> and <u> at t1, probably because these graphemes have more than one possible phonemic realisation. The challenge is particularly understandable for Chinese students, as their Pinyin system has consistent GPCs, where <a> is always pronounced as /a/, <e> as /ɤ/, <i> as /ɪ/, <o> as /o/ and <u> as /u/. Indeed, the most common error in decoding the grapheme <a> is pronouncing it as /a/ (in contexts where it should be pronounced differently), which is its Pinyin realisation. Similarly, the most common error in decoding the grapheme <u> is also pronouncing it as its Pinyin realisation /u/, which is consistent with the interpretation of Pinyin influence. Interestingly, as the grapheme <i> is pronounced as /ɪ/ in several pseudo words in the decoding test, which is the same pronunciation as its Pinyin realisation, <i> is the grapheme showing the least number of errors out of the five graphemes, with an error rate of only 10%. This again shows that L1 influence can be positive where the GPCs are the same.

Another angle to look at the errors is that not only do these graphemes have more than one possible phonemic realisation, but also there are more phonemes in the same articulatory space in English than in Chinese. Hence, the lack of familiarity with the English phonology system might also be a source of error for these graphemes.

At t2, though the intervention participants still committed quite a number of errors in decoding these graphemes, they produced fewer Pinyin-resembling errors. This shows that the intervention participants were gradually moving from Pinyin-resembling pronunciations to more 'English-like' pronunciations, even though they had not yet achieved the standard pronunciations. This, however, was not the case for the comparison groups, who still predominantly pronounced these graphemes as they sound in Pinyin.

As mentioned in section 8.3.2.2, correctly decoding the graphemes <a> <e> and <u> as schwa seemed especially difficult for the participants. However, the findings of previous speech analysis studies might provide a different angle to interpret these results. Deterding (2006, 2010) has observed that Chinese students tend to pronounce the reduced vowel (which should be schwa) as the full vowel in their reading of an English passage, such as pronouncing <o> in 'consider' and 'confess' as /ɒ/. He argues that the reason why many Chinese students use full vowels when schwa should be pronounced is that they believe using schwa is a lazy way of speaking. It is not possible to know if this was the case for the participants in this study, as no self-report

techniques were used. However, future phonics instruction programmes should consider including stress patterns, which was not covered in the phonics instruction programme of this study, but might potentially contribute to the correct realisation of schwa in reduced (unstressed) vowels. This also shows the complexity of teaching English decoding: simply covering GPCs might not be enough; other pronunciation skills might also need to be taught.

#### 8.4.2.3 Vowel digraphs that exist in Pinyin

As demonstrated in section 8.3, the vowel digraphs that exist in Pinyin, namely <ei> <ie> <ou> and <ai>, were the ones demonstrating the lowest accuracy among all the vowel graphemes at t1, and also resisted the effects of the instruction programmes most stubbornly at t2. Error analysis suggests that the poor decoding of these graphemes may be attributable to Pinyin influence, as they were pronounced in the same way as in Pinyin in almost half of the cases at t1. At t2, both the intervention and comparison groups made similar numbers of Pinyin-resembling errors, suggesting the lack of effect for both the phonics instruction and the phonology instruction in treating these errors.

The possible reason for the lack of improvement for these graphemes was discussed in section 8.3.2.1. As these graphemes seem to be the ones that are particularly challenging for the participants, future instruction programmes may need to give more

focus to them. Possible measures to be taken might include repeated exposure to these graphemes over a longer period of time and providing more examples in words of different lengths (including both monosyllabic and polysyllabic words containing these graphemes). In this study, the graphemes were covered only for two weeks (one week of instruction plus one week of revision), and most examples given were monosyllabic; this may have been insufficient.

#### 8.4.2.4 Vowel digraphs that do not exist in Pinyin

The error pattern for the vowel digraphs that do not exist in Pinyin is less clear. This is probably because these graphemes do not have an equivalent in Pinyin; as a result, more 'wild forms' were produced for these graphemes. However, some interesting characteristics can still be observed in participants' realisation of these graphemes. For instance, the graphemes <ee> <ea> and <ey> were all wrongly pronounced as /e/, which is a possible decoding for grapheme <e>. It could be that the participants did not know how to pronounce these graphemes, so they just pronounced the first letter of these graphemes.

After the phonics instruction, intervention participants clearly made fewer errors in decoding these graphemes compared to the comparison groups, indicating that explicit instruction is helpful in learning how to decode these graphemes effectively. However, the error rate for these graphemes for the intervention participants was still

somewhat high at t2, suggesting that the successful acquisition of these graphemes is not a straightforward process and might take a longer period of time than was available for the intervention.

Altogether, it can be seen that compared to the consonant graphemes, the picture for the vowel graphemes is much more complicated. Participants' Pinyin knowledge seemed to be an important factor influencing their decoding of vowel graphemes. This is not only shown in some Pinyin-resembling pronunciations, but also in the way some words were segmented. In order to teach the decoding of vowel graphemes more effectively, not only should GPC knowledge be instructed, but so also should other elements of pronunciation, such as stress patterns. Though the intervention participants still produced quite a number of errors at t2, their error rate was much lower than the comparison groups' for most vowel graphemes, suggesting that the phonology instruction was effective, but not effective enough to improve the decoding of some vowel graphemes.

### 8.4.3 Other errors

Some other errors beyond individual graphemes were also observed in participants' pronunciations, namely whole-word errors and misordered graphemes.

#### 8.4.3.1 Whole-word errors

Participants made some whole-word errors in the decoding of the following words:

*bab*, *whie*, *twem*, *yeng*, *wrey*, *knoink*, *wrault*, *dat*, *quiles*, *weat*, *cyr*

It is of interest to explore why participants made these whole-word errors. An initial hypothesis, considering the obvious typological differences between the Chinese and English writing systems, would be that the participants used whole-word processing mechanisms to read these words, as they would do to read Chinese characters. This is a common opinion in research on Chinese EFL learners' reading of English words (e.g. Wang et al., 2003). These researchers, through comparing Chinese EFL learners with English learners with an alphabetic L1, argue that Chinese EFL learners lack phonological awareness because of their L1, and thus rely on orthographic information instead of phonological information in the recognition of English words. However, whole-word errors might not be specific to L2 learners. Walley (1993) argues that beginning L1 learners, who have developed some awareness of phonemes, may still perceive unknown words as single units. It is possible that the participants in this study, like some L1 learners, had some phonological awareness of the English language, but were not phonologically aware enough to combat their whole-word processing mechanisms in reading these words. It could be, as Hu argues (2003:453-454), that 'the attentional demands may be so great that students who are phonologically aware nevertheless cannot effortlessly and immediately extract the phonemic details', and instead 'rely on their capacity to store the overall shape of an unfamiliar representation temporarily as a satisfactory approximation'.

Taking a further look at the words which were associated with whole-word errors, it can be seen that all these words are monosyllabic. Based on this observation, a possible explanation is tentatively offered. The participants were influenced by the whole-word processing mechanisms of their L1 in reading pseudo English words; however, the L1 transfer only applied to short monosyllabic words, as they are more likely to bear close resemblances to real words. On the other hand, the long, polysyllabic pseudowords in this study were not associated with any whole-word errors; this may have been because they bore less resemblance to any real words that the participants knew, thus forcing them to decode them using whatever phonics knowledge they had.

At t2, the intervention participants made substantially fewer whole-word errors compared to the comparison groups. This suggests that the phonics instruction programme may have equipped them with useful knowledge of English GPCs and may have encouraged them to engage in intraword analysis, which is what previous studies (e.g. Koda, 2007) have suggested that they are less likely to do, by dint of their morphemic L1 writing system.

#### 8.4.3.2 Misordered graphemes

Another type of errors observed was misordered graphemes, such as reading *byrcal* as *brycal* (/bri:kəl/, /braɪkəl/). This shows inaccurate intra-word analysis. The lack of

competence in intra-word analysis has often been associated with the logographic nature of the Chinese writing system in previous studies (e,g. Akamatsu, 2003). However, a further look at the errors in this category suggests that the reason might be more complicated than this. For instance, the two errors in decoding the word *knoick* were /nɒki:/ and /kənɒki:/, both of which end in a vowel and follow the CVCV pattern. Similarly, the two errors in decoding the word *bufty* were /bəfjuti:/, /bəfuti:/, both of which also end in a vowel and follow the CVCVCV pattern. This tentatively suggests the influence of Chinese phonology, where a syllable always consists of an onset and a rime (with no coda). The participants who committed these errors might have been pronouncing these words in a way that fits the pattern of Chinese syllables.

Though the phonotactic constraints of the L1 may have contributed to the misordered grapheme errors, the phonology instruction programme alone did not seem to be effective in treating these problems, as the error rate remained roughly the same at the two time points for the comparison groups.

### 8.4.4 Error origins and possible implications for future instruction programmes

Section 8.4 has examined the errors in participants' decoding at the two time points, and summarised the error types at the grapheme level and beyond the grapheme level. The analysis has also tentatively indicated possible sources of these errors, and the following two sources were hypothesized.

The first source is a lack of familiarity with the English phonological system. This could be seen to account for most of the errors in the decoding of the consonant graphemes, and may also explain some of the errors beyond individual graphemes (e.g. misordered graphemes). As the Chinese and English phonological systems have many differences in both phonemes and syllable structure, it may be extremely useful for Chinese EFL learners to consciously know these differences, as a basis for phonics instruction. The importance of English phonological knowledge is also evident in comparison groups' progress at t2. Even though they did not receive explicit instruction on English GPCs, they still produced fewer errors and better decoding results at t2 for some categories of errors.

The second source is the lack of knowledge of English GPCs. As English and Pinyin share many graphemes, problematic pronunciations often arise when the corresponding phonemes are different in the two systems. This is especially true in the decoding of vowel graphemes, and is exacerbated by the contrast between the one-to-one GPC system of Pinyin and the 'many-to-many' GPC system of English: i.e. the fact that a single English vowel grapheme may have more than one conventional phonological realisation (and likewise, any one English phoneme may have more than one possible written representation). Moreover, some graphemes unique to English also showed many errors at t1, as the participants did not know how to pronounce them and may have ventured a guess. The twelve-week phonics instruction programme in this study was not able to correct all the Pinyin-resembling

pronunciations. It is worth exploring whether a longer instruction programme with repeated exposure to problematic graphemes might solve these problems.

Based on the above, the following table detailing the sources of different types of errors in participants' decoding and the possible pedagogical solutions is tentatively proposed. However, two things need to be borne in mind. Firstly, these possible error sources remain conjecture, since no insight was sought into participants' processing mechanisms in the current study (e.g. via self-report interviews). Secondly, as decoding involves intricate processing, it is possible that an error can be attributable to more than one source.

*Table 8.1 Origins of errors and possible pedagogical solutions*

<table>
<tr><th colspan="2">Possible error source</th><th>Likely error type</th><th>Possible pedagogical solutions</th></tr>
<tr><td rowspan="4">Phonology</td><td rowspan="2">phoneme inventories</td><td>Approximation of consonant graphemes</td><td rowspan="2">Familiarising learners with English phonology; practise English-specific phonemes</td></tr>
<tr><td>Epenthesis in consonant graphemes</td></tr>
<tr><td rowspan="2">phonotactics</td><td>Omission of consonant graphemes</td><td rowspan="2">Familiarising learners with English word structures; practise intraword analysis and syllable segmentation</td></tr>
<tr><td>Misordered graphemes</td></tr>
<tr><td colspan="2" rowspan="5">GPCs knowledge</td><td>Split digraphs</td><td rowspan="5">Familiarising learners with English GPCs, with special focus given to grapheme that shared by pinyin and English; Instructing lexical stress knowledge to facilitate the pronunciations of some vowel graphemes</td></tr>
<tr><td>Vowel graphemes with multiple realisation</td></tr>
<tr><td>Vowel digraphs that exist in Pinyin</td></tr>
<tr><td>Vowel digraphs that do not exist in Pinyin</td></tr>
<tr><td>Whole-word errors</td></tr>
</table>

The analysis of errors in participants' decoding shows that many of the errors seem to be L1-induced. Though Chinese EFL learners are often believed to have poorer decoding competence compared to their proficiency-matched counterparts from other L1s (e.g. Hamada & Koda, 2008), the reason is always hypothesised to be an artifact of the difference in the writing systems, where Chinese is conceptualised as a language with only a logographic system. However, the error analysis in this study suggests that Pinyin knowledge contributes to many errors in mainland Chinese participants' decoding. Future phonics instruction programmes should take this into consideration, and attach different importance to different graphemes, rather than treating mainland Chinese students merely as logographic readers, who do not have any prior knowledge of GPCs.

## 8.5 Research Question 4

*RQ4: Do participants who have followed the programme of phonics instruction also show gains in their ability to recall (productive knowledge) and recognise (receptive knowledge) new English words?*

This question was intended to investigate whether the possible causal relationship between phonological decoding and vocabulary learning in L2 English. This question was addressed by using a learning task involving ten low-frequency English words. After the learning session, participants' learning results were measured using four

tests, namely an oral recall test, a written recall test, an aural recognition test and a written recognition test. In order to acknowledge the intention to treat, the aggregated recall and recognition test results of the whole sample (all three universities) were analysed first; followed by the aggregated results of Universities A and B only; and finally the results of University C.

### 8.5.1 The oral recall and the written recall test results

*Oral recall test*

The examination of the oral recall test results showed that both intervention and comparison groups (across all three universities) correctly recalled only approximately 1 out of the 10 words at t1. The low accuracy achieved in the oral recall test at t1 naturally invites the question of why the oral recall test seemed particularly challenging for the participants. Several factors possibly contributing to these results are presented below.

Firstly, it is likely that the participants seldom had to work out the pronunciations of newly-encountered words in their previous intentional vocabulary learning experience. Instead, the pronunciations would have been provided to them either by the teacher, or by electronic dictionaries which can sound out the pronunciations of chosen words, or by the IPA symbols in a vocabulary list (Ma, 2009). As a result, working out the

pronunciations of unfamiliar words without access to audio/ IPA support was not necessarily part of their vocabulary learning routine, even when they were specifically told about the four test forms before the learning session. It should be noted that for roughly 50% of the words in the t1 test, participants made no attempt to pronounce the word at all. Given participants' lack of sounding-out experience, it is unclear whether this was due to lack of competence or lack of confidence.

Secondly, in their previous learning experience, new words were taught as sight words, with no analysis of GPCs within words (Ma, 2009). The lack of intra-word analysis practice may also have contributed to the lack of competence, or willingness, to venture a pronunciation for an unfamiliar word, which in turn may have led to the low scores in the t1 oral recall test.

At t2, the intervention participants (across all three universities) correctly recalled 3.7 out of the 10 words, while the comparison groups (across all three universities) recalled 2.5 words in the oral word recall test. Statistical analysis showed that the intervention participants made significantly more gains in the oral recall test than the comparison groups, indicating that the phonics instruction programme was effective in facilitating the recall of phonological forms in vocabulary learning. This is understandable, as the phonological decoding test and the recall of the phonological forms of unfamiliar words both entailed working out the pronunciations using GPC knowledge (and in the latter case, retaining the words in long-term memory).

The analysis of the oral recall test results of University C revealed that the intervention participants in University C did not achieve significantly greater gains than the comparison counterparts by t2. Given that the intervention participants in University C did not achieve significantly greater gains than the comparison groups in the phonological decoding test either, this contributes to the case for a connection between decoding proficiency and the productive phonological aspect of vocabulary learning.

*Written recall test*

The examination of the written recall test results showed that at t1, both intervention and comparison groups (across all three universities) correctly recalled approximately 3 out of the total 10 words. Compared with the oral recall test, both intervention and comparison groups achieved higher scores in the written recall test. As previous studies have demonstrated, Chinese EFL learners often attach more importance to the orthography of new words than to their phonology (Ma, 2009). This was confirmed by the informal observation of participants' behaviour in the vocabulary learning task, as the majority of their learning time seemed to be given to memorising the spellings of the words. Many participants chose to write down the English words after seeing them on the computer screen, sometimes repeating the spellings several times on the paper. Some participants also murmured the letter names of the words (rather than their phonological values) to themselves in the learning session. In contrast, fewer

participants were observed to pronounce the words in the learning session, even though they knew they would be tested about the pronunciations of the words later. Given this observed behaviour, the higher accuracy achieved in the written recall test compared to the oral recall test may not only be the result of different test forms, but may also reflect participants' strategic use of their learning time. This will be further discussed later. Of course, it is acknowledged that there are limitations to what the observation of external behaviours can tell us about strategic behaviour, assuming Macaro's (2006) classification of a strategy as a mental action taking place in working memory.

At t2, intervention participants (across all three universities) correctly recalled approximately 6.1 out of the 10 words, and the comparison groups (across all three universities) correctly recalled 5.1 words on average. Intervention participants made significantly more improvement compared to the comparison group in the written recall test, indicating the effectiveness of the phonics instruction programme in terms of promoting the memorisation of orthographic forms. This recalls the results of Hamada and Koda's (2008) study, where participants with higher L2 English decoding proficiency also achieved significantly higher scores in the recall of novel orthographic forms. The current study goes a step further, finding – in a controlled experimental context – that improvements in L2 decoding proficiency result in improvements in the recall of novel orthographic forms. The underlying mechanisms might be explained by the hypothesis that orthographic form and phonological form

can be mutually reinforcing in the process of vocabulary learning (Hamada & Koda, 2008). That is, reliable phonological representations of new words can be obtained through proficient decoding, and encoding the phonological representations using GPCs can facilitate the recall of orthographic forms. It is worth noting that the comparison groups also showed a significant increase in written recall scores between t1 and t2 (albeit to a lesser extent than the intervention group). Considering that they also achieved a significant increase in decoding scores between t1 and t2 (though again to a lesser extent than the intervention group), this, from a different perspective, suggests a close connection between decoding proficiency and recall of novel orthographic forms.

The analysis of the written recall test results of University C revealed that the intervention participants in University C did not achieve significantly greater gains than their comparison counterparts between t1 and t2. Given that the intervention participants in University C did not achieve significantly greater gains than the comparison groups in the phonological decoding test either, this again seems to point to the close connection between decoding proficiency and the written productive aspect of vocabulary learning.

*Comparison between the two recall test results*

As both the oral recall test and the written recall test are recall tests, which assess

productive vocabulary knowledge, it is of interest to compare the two test results at the two time points. For the sake of comparison, the mean scores achieved in the two tests at the two time points by participants across all three universities are listed below (out of a total score of 10):

| | **t1** | | **t2** | |
|---|---|---|---|---|
| | **Oral recall** | **Written recall** | **Oral recall** | **Written recall** |
| **Intervention (N=94)** | 0.8 | 2.8 | 3.8 | 5.5 |
| **Comparison (N=86)** | 0.9 | 2.8 | 2.5 | 4.7 |

It can be seen that the mean scores achieved in the written recall test were consistently higher than those in the oral recall test for both intervention and comparison groups, both before and after the instruction programmes. It should be noted that this may not exactly be a fair comparison, as the orthographic forms were clearly given in the learning task, but the phonological forms were not given and needed to be obtained through decoding. As a result, participants might have allocated more attentional resources to learning the orthographic forms than to working out the phonological forms in the learning session. However, this possible strategic allocation of their learning time may not only be the result of the task format, but may also reflect the choices of vocabulary learning strategies in their daily learning experience. Many previous studies have demonstrated that the most frequently used vocabulary learning strategy of Chinese EFL learners is to repeat the spellings, even when pronunciations of the words are given (e.g. Hong, 2008). As Gu and Johnson argue (1996: 679), 'students consistently adopt types of strategies based either on their beliefs about

vocabulary and vocabulary learning, or on other pre-existing cognitive or social factors'. From this perspective, the higher accuracy in the written recall test than the oral recall test may reflect participants' belief that learning the orthographic forms of new words is more important than learning the phonological forms. Similar results can also be found in Li (2012).

Even though the gap between oral recall scores and written recall scores can be observed at both time points, it can be seen that the gap between the two narrowed at t2 compared to t1 for the intervention participants, but not for the comparison groups. This seems to suggest that the phonics instruction programme helped intervention participants become more 'balanced' learners, whose productive phonological knowledge of new words caught up more quickly with their productive orthographic knowledge.

*Error analysis of the written recall test results*

Error analysis was also conducted for the written recall test. It can be seen that at t1, the most common type of errors made by both intervention and comparison groups were positional errors, such as recalling *ploidy* as *pliody*, *cantor* as *catron*. These errors account for approximately 50% of the total errors for both groups. The large number of positional errors seems to be in line with the previously mentioned observation that many participants wrote down or orally repeated the spellings of the

words (i.e. said the letter names out loud), indicating their heavy reliance on visual processing in the vocabulary learning task. In contrast, phonologically plausible errors, such as recalling *burlap* as *birlap* or *berlap*, *augean* as *augin*, were much less common at t1, suggesting that few participants correctly decoded the words in the learning process. This echoes the results of the oral recall test. Both intervention and comparison groups achieved low accuracy in this test at t1. At t2, intervention participants made fewer positional errors and many more phonologically plausible errors, suggesting that their progress in decoding proficiency led to greater willingness or confidence to decode unfamiliar words, which in turn led to more phonologically plausible errors in the written recall test. The comparison groups, on the other hand, made similar numbers of phonologically plausible errors at the two time points, reflecting their lack of progress in phonological decoding.

Looking beyond the errors themselves, it may be speculated that the two types of errors may indicate two types of learning strategies. Positional errors may be the outcome of more orthography-inclined strategy, whereas phonologically plausible errors indicate a more phonology-inclined strategy. This, again, may be linked with decoding proficiency. As demonstrated in Chapter 7, significant correlations were found between decoding proficiency (both word-level and grapheme-level) and the number of phonologically plausible errors for both intervention and comparison groups at both time points. This suggests that no matter what instruction programme they received, participants with higher decoding proficiency were more likely to make

phonologically plausible errors. In other words, better decoding proficiency might lead to more confidence or willingness to decode unfamiliar words, which in turn might result in more use of phonology-inclined strategies in learning new words, which might then be reflected in the phonologically plausible errors in the written recall test.

However, it is interesting to note that even though the comparison group also made significant progress in the decoding test (when measured at both word level and grapheme level) as demonstrated in Chapter 5, this did not seem to result in more phonologically plausible errors at t2. A possible explanation is that, though the phonology instruction programme helped promote comparison groups' decoding proficiency (albeit to a lesser extent than the phonics instruction), it did so in an implicit and imperceptible way. In other words, the comparison groups may not have known that they were getting better in decoding unfamiliar words, and thus were less likely to use it as a strategy to memorise new orthographic forms. It is possible that explicit instruction is required to actually convert phonics knowledge into helpful tools in memorising new orthographic forms. Another possibility is that there might exist a threshold in phonics knowledge which must be reached before this phonics knowledge is reflected in vocabulary learning strategies – i.e. though the participants did get better at decoding, they were still not 'good enough' for this to make a difference to their vocabulary learning strategies.

The results of the error analysis of the written recall test also points to the importance of qualitative analysis in vocabulary learning research. The differences between phonologically plausible errors and positional errors cannot be detected from quantitative analysis, as both errors would get a zero score on the word-level.

### 8.5.2 The aural recognition and the written recognition test results

The aural recognition test and the written recognition test were also conducted to examine the influence of the two instruction programmes on the receptive aspects of vocabulary learning.

*Aural recognition test*

At t1, the intervention and comparison groups (across three universities) recalled approximately 5.6 out of the total 10 words in the aural recognition test. The mean score in the aural recognition test was higher than in the oral recall, though both of these were designed to assess the phonological aspect of vocabulary learning. This was expected, as receptive knowledge is generally expected to be acquired prior to the productive knowledge of the same aspect of vocabulary learning (Henriksen, 1999).

However, the relatively high mean score achieved in the aural recognition test needs to be interpreted with caution. This is because the aural recognition test form permits

more strategic behaviour than the two recall test forms. For instance, correctly recognising the word *cantor* only requires establishing the link between the grapheme <c> and the phoneme /k/, as *cantor* is the only word in the test starting with /k/. Another factor worth considering is that all the words in the learning session have only six letters and two syllables, which makes such strategic behaviour more likely to work. Taking this into consideration, the higher mean score in the aural recognition test compared to the two recall tests does not necessarily suggest that the acquisition of receptive phonological knowledge is easier than the acquisition of productive phonological/orthographic knowledge in real life learning situation. In fact, many studies have documented that Chinese EFL learners encounter difficulty in listening comprehension. For instance, in Huang's (2005) study of Chinese university-level EFL learners' self-reported learning difficulties, 23% of the participants said that they had difficulty in listening comprehension, such as inability to recognise already known English words. Similar findings were also observed in Erler (2003), where some Year 7 English learners of French also reported difficulty in listening comprehension with comments like: 'when Miss says it, it doesn't sound like when I read it', 'they look like a different word than you have been repeating' (p. 165). Hence, it is important to interpret the relatively high accuracy in the aural recognition test with caution.

At t2, the intervention participants (across all three universities) recognised on average 7.7 out of the 10 words in the aural recognition test, and the comparison

groups (across all three universities) recalled approximately 6.7 words. Both the intervention and comparison groups made significant progress in the aural recognition test between t1 and t2, but the progress was significantly greater for the intervention participants than the comparison groups, indicating the advantage of the phonics instruction programme over the phonology instruction programme in promoting the development of receptive phonological knowledge in an intentional word learning task. This was expected, as the phonics instruction programme explicitly taught GPC knowledge, which fosters more reliable phonological representation of the new words. This made aural recognition of the target words more likely, even though the pronunciations of the words were not provided.

The analysis of the written recall test results of University C revealed that the intervention participants in University C did not achieve significantly greater gains than the comparison counterparts between t1 and t2. Given that the intervention participants in University C did not achieve significantly greater gains than the comparison groups in the phonological decoding test either, this again seems to point to the close connection between decoding proficiency and the receptive, phonological aspect of vocabulary learning.

*Written recognition test*

In comparison to the aural recognition test, the intervention and comparison groups

(across all three universities) scored even higher in the written recognition test at t1, with a mean score of 6.6 out of 10 for both groups. The already high score achieved at t1 was expected, as no decoding is needed in the written recognition test in order to produce a correct answer. At t2, though both intervention and comparison groups (across all three universities) scored higher than at t1 on the written recognition test, intervention participants did not make significantly greater gains than the comparison group. This suggests that the phonics instruction did not lead to better recognition of the orthographic forms compared to the phonology instruction programme. This, again, is in line with what was hypothesized. Given the lack of need for decoding in the written recognition test, it is understandable that the phonics instruction programme, which promoted decoding proficiency, did not lead to significantly greater gains in the written recognition test.

Interestingly, the analysis of the written recognition results of University C demonstrated that the intervention participants here did make significantly greater gains between t1 and t2 than the comparison groups. However, as the phonics instruction programme did not lead to significantly greater gains in phonological decoding in University C, this suggests that intervention participants' significantly greater gains in the written recognition test was not the result of improvement in decoding proficiency. This, again, seems to provide evidence to support the previous argument that decoding proficiency is less relevant for the written recognition test.

*Comparison between the two recognition test results*

As both the aural recognition test and the written recall test are recognition tests in nature, which assess receptive vocabulary knowledge, it is of interest to compare the two test results at the two time points. For the sake of comparison, the mean scores achieved in the two tests at the two time points by participants across all three universities are listed below (out of a total score of 10):

| | **t1** | | **t2** | |
|---|---|---|---|---|
| | **Aural recognition** | **Written recognition** | **Aural recognition** | **Written recognition** |
| **Intervention (N=94)** | 5.5 | 6.5 | 7.8 | 8.9 |
| **Comparison (N=86)** | 5.8 | 6.7 | 6.7 | 8.6 |

It can be seen that the mean scores achieved in the written recognition test were consistently higher than those in the aural recognition test for both intervention and comparison groups, both before and after the instruction programmes. As previously discussed, no decoding is required in the written recognition test in order to produce a correct answer; in contrast, even though aural recognition does not necessarily require correct decoding in order to produce a correct answer, some degree of decoding is still warranted, such as recognising some phonemes that could be mapped onto the graphemes in the test items.

### 8.5.3 Summary for RQ4

The analysis of the four recall and recognition test results tentatively provides an answer to the question of whether a causal relationship exists between decoding proficiency and various aspects of vocabulary learning, which has not been addressed in the previous literature. The results demonstrate that phonics instruction led to progress particularly in the phonological aspect of vocabulary learning, including both productive and receptive phonological knowledge. Phonics instruction also led to progress in the productive orthographic aspect of vocabulary learning, but not in the receptive orthographic aspect of vocabulary learning.

This section also tentatively discussed how decoding proficiency contributes to different aspects of vocabulary learning. Though the importance of decoding proficiency in vocabulary learning is well supported by the working memory model as discussed in Chapter 2, this section tentatively offered a more nuanced way to interpret the relationship between decoding proficiency and vocabulary learning: that is, decoding proficiency may directly or indirectly contribute to different aspects of vocabulary learning. For the phonological aspect of vocabulary learning, the contribution of decoding proficiency is direct, as correct recall and recognition of phonological forms both entail correct phonological representation to begin with. For the productive orthographical aspect of vocabulary learning, the contribution of decoding proficiency may be direct or indirect. Error analysis of the written recall test reveals that participants with higher decoding proficiency were more likely to make

phonologically plausible errors, whereas participants with lower decoding proficiency were more likely to make positional errors. The underlying mechanism may be that participants with different levels of decoding proficiency were likely to choose different strategies for vocabulary learning. Moreover, the intervention participants made more phonologically plausible errors after receiving the phonics instruction, perhaps because they became more likely to use phonological strategies. This suggests that decoding proficiency might not only impact the productive orthographical aspect of vocabulary learning through the quality of phonological representation of new orthographic forms, but also through vocabulary learning strategies.

# Chapter 9. Conclusions and Limitations

## 9.1 The overarching research questions

Phonological decoding, defined as the process of converting visually presented word forms into their phonological forms, has been argued to underlie different aspects of L1 and L2 learning. Traditionally, decoding has been studied in conjunction with reading comprehension (Koda, 2007). The contribution of decoding to L2 vocabulary learning, on the other hand, has received less attention from researchers, though there is strong theoretical support for a causal relationship between the two variables: correct decoding of the written forms provides reliable phonological representations of new words in phonological working memory, which plays a central role in learning new phonological forms. Moreover, knowledge of a language's GPCs allows the orthographic and the phonological representations of new words to be mutually reinforcing. Though decoding is believed to play important roles in different aspects of L2 learning, Chinese EFL learners have been argued to be less proficient in English decoding than learners with other L1 backgrounds. Previous studies (e.g. Hamada & Koda, 2008, 2010) attribute this to the possible influence of their L1 writing system, which is typologically different from English and does not necessarily require sensitivity towards intraword structures.

Currently, little has been determined about whether Chinese EFL learners' English

decoding can be promoted through systematic phonics instruction, though phonics instruction has been gaining a lot of interest and popularity in China recently based on the researcher's knowledge. Taking these points into consideration, the current study has attempted to address the following overarching research question:

> *What are the effects of a systematic programme of L2 English phonics instruction on (a) English decoding proficiency and (b) English vocabulary learning, amongst Chinese university-level EFL learners?*

This question was addressed by comparing a systematic English phonics instruction programme and a systematic English phonology instruction programme. The effects of these programmes were compared on a total of 180 first-year English majors studying in three different universities in Wuhan, China. The intervention participants (N = 94) and the comparison groups (N = 86) were closely matched in terms of their English proficiency (as measured by the National College Entrance English Exam) and their English vocabulary knowledge (as measured by the British Picture Vocabulary Scale, Dunn & Dunn, 2009). Before the intervention, both groups of participants took part in a phonological decoding test, in which they were asked to read aloud 28 pseudo words in the Word Attack section of the Woodcock Reading Mastery Test (Woodcock, 2011). They also took part in a vocabulary memorisation task, in which 10 low-frequency English words were presented with their Chinese translations for the participants to learn, followed by four recall and recognition tests

measuring (a) both productive and receptive and (b) both phonological and orthographic aspects of vocabulary learning.

The participants were then involved in 12 weeks of phonics or phonology instruction. The phonics instruction was characterised by (a) clearly explained, incremental coverage of 44 phonemes, 101 English GPCs and 27 common word endings; (b) structured reading material to help consolidate the knowledge of newly-instructed phonics knowledge; (c) feedback on participants' pronunciations. The phonology instruction was characterized by (a) clearly explained, incremental coverage of 44 phonemes and their IPA symbols; (b) pronunciation skills such as intonation, linking and stress patterns; (c) feedback on participants' pronunciations. No explicit instruction of English GPCs was provided in the phonology instruction programme, to avoid overlap between the two instruction programmes and thus to allow the effects of the Phonics instruction on the intervention group to be isolated.

After 12 weeks, participants were again involved in a phonological decoding test and a vocabulary memorisation task followed by four recall and recognition tests. Though the tests were identical in form to the ones they took before the intervention, the test items were different, to avoid a practice effect.

## 9.2 The specific research questions

Four specific research questions were addressed in this study, the findings of which are summarised below.

***RQ 1: Does a programme of systematic phonics instruction lead to improvements in Chinese university EFL learners' English decoding?***

A simple, straightforward answer to this question is yes: in the sample as a whole (i.e. combining all three universities), participants who received the systematic phonics instruction were found to make significantly more progress than the comparison groups over the course of intervention, in terms of: (a) the numbers of correctly decoded words and (b) the percentage of correctly decoded graphemes in the decoding test. This was despite the groups being matched on English proficiency and English vocabulary knowledge. Thus, the study provides evidence for the effectiveness of systematic phonics instruction in terms of promoting participants' decoding proficiency.

On average, intervention participants (across all three universities) correctly decoded 4 more words out of the 28 pseudo words after the intervention, though there was still plenty of room for progress, as they still only decoded around half of the test items correctly after the intervention. At the grapheme level, they correctly decoded less

than 70% of the total graphemes before intervention, but more than 85% of the total graphemes after the intervention. Considering that the intervention only lasted for 12 weeks, these results provide encouraging evidence for the effectiveness of the phonics instruction in terms of promoting decoding proficiency. Compared with the comparison groups who received phonology instruction, the intervention participants also received explicit instruction in the pronunciation and discrimination of the 44 English phonemes; however, the intervention participants did not receive instruction in pronunciation skills, such as linking, rhythm and intonation. Nonetheless, these pronunciation skills may still be picked up by the intervention participants in other English classes; thus, it is highly unlikely that the missing out on these pronunciation skills would jeopardise their future English learning.

Though higher accuracy of decoding is often associated with faster decoding speed (or shorter overall time of decoding, as operationalized in this study), the phonics instruction programme did not lead to greater improvement in terms of decoding speed than did the phonology instruction. In contrast, the comparison groups, who did not follow the programme of phonics instruction, showed a significantly greater decrease in overall decoding time compared to the intervention participants at t2. It was hypothesized that the reason may be attributable to the possibility that the intervention participants were engaged in conscious processing (i.e. 'figuring out' the pronunciations of the pseudowords), and had not yet achieved automaticity in decoding. Indeed, it was found that some participants who spent the longest time

completing the decoding test were actually among the ones the produced the best results. Future studies might want to consider the value of using untimed, or at least generously timed decoding test to measure learners' decoding proficiency, at least at beginner level.

The analysis by university have revealed that unlike the other two universities, the intervention participants in University C did not demonstrate greater progress decoding accuracy, either at the word level or the grapheme level. A possible reason for this is that the participants in University C were concurrently learning other foreign languages in addition to English during the intervention period. Indeed, in the case of the intervention participants, this language was French, which uses the same Roman alphabet as English and so presented particular scope for interference. Future instruction programmes might want to consider the possible confounding effects of learning other languages when learning English phonics.

***RQ2: Is the programme of phonics instruction more effective for some GPCs than others?***

Only the data of Universities A and B were analysed for the purpose of this question, given that there is no point in looking for variations where there is a lack of overall change (University C), probably due to the confound of learning another foreign language.

A simple, straightforward answer to this question is also yes. The comparison of the accuracy percentages of individual GPCs at the two time points have revealed that, the phonics instruction seemed to be the most effective in promoting the knoweldge of three types of GPCs, namely (a) Pinyin-absent consonant GPCs (e.g. <ss> = s), (b) Chinese-absent consonant GPCs (e.g. <dg> = /dʒ/), and (c) Pinyin-absent vowel GPCs (e.g. <ey> = /eɪ/).

It was also found that there appears to be a connection between participants' knowledge of Pinyin GPCs and their decoding of English words. For the Pinyin-congruent GPCs (e.g. <p> = /p/), there seems to be a facilitative influence of Pinyin knowledge, as participants already decoded them accurately before the intervention. On the contrary, Pinyin-incongruent GPCs (e.g. <ai> = /ei/) seemed to pose a challenge to participants before the intervention, and also appeared to resist the effects of the phonics instruction most stubbornly.

These results serve as a useful starting point to study Chinese university-level EFL learners' decoding of individual English GPCs. Given that participants' knowledge of different GPCs seemed to vary before the intervention, and the phonics instruction appeared to have different effects on individual GPCs, future instruction programmes might want to take this into consideration, and direct more focus on challenging GPCs (such as Pinyin-incongruent GPCs).

***RQ3: What are the features and problems of Chinese university EFL learners' English decoding at each time point?***

In line with RQ2, only the data of Universities A and B were analysed for the purpose of this question, as participants in these two universities received a 'purer' form of intervention.

Before the intervention, the two groups of participants mainly made three types of errors in the decoding of consonant graphemes, namely epenthesis, omission and approximation. Epenthesis was found in both individual consonant digraphs/ trigraph (e.g. <wr>) and in consonant strings (e.g. <tw>), where a vowel, mostly a schwa, was added in between the two consonants. Omission was found in long words with multiple syllables, where the first letter of consonant strings was often omitted. Some consonant graphemes were mispronounced as similar-sounding or a similar-looking graphemes. After the phonics instruction, the intervention participants made clearly fewer of the first two types of errors (epenthesis and omission) and showed a clear advantage over the comparison groups, but the last type of error (approximation) remained difficult to tackle.

In terms of decoding the vowel graphemes, the two groups of participants mainly made four types of errors before the intervention. The first type of error was sounding out the silent <e> in split digraphs; the second type of error was overgeneralising the

pronunciations of graphemes with multiple realisations; the third type of error was Pinyin-resembling pronunciations for vowel digraphs that also exist in Pinyin; the last type of error was wild forms for vowel digraphs that do not exist in Pinyin. After the phonics instruction, the intervention participants made fewer errors in decoding almost all the vowel graphemes, though there remained problems with the decoding of vowel digraphs that also exist in Pinyin.

In addition, some whole-word errors and misordered graphemes were also found before the intervention, but were almost eradicated by the phonics instruction, suggesting the effectiveness of phonics instruction in promoting intraword analysis.

The analysis of participants' pronunciations in the decoding test demonstrates that many of the errors appear to be L1-induced. This provides a new angle to look at the potential influence of Chinese EFL learners' L1 on their L2 learning. Many previous crosslinguistic studies seem to categorize Chinese EFL learners as logographic readers, and explain their problems in decoding English words by the typological difference between the Chinese (morphemic) and the English (alphabetic) writing systems. Indeed, some errors made by the participants seemed to be influenced by processing mechanisms associated with the morphemic Chinese writing system, for instance, some whole word errors (e.g. pronouncing *wrault* as *worried*) seemed to reflect a lack of intraword analysis and visual processing, which are often believed to be the features of the processing mechanisms associated with Chinese characters.

However, there are many other errors that cannot be explained simply by the processing mechanisms associated with the Chinese writing system. Participants' Pinyin knowledge appeared to play an active role in their decoding of English words, and could potentially have been disruptive when the same grapheme is pronounced differently in Pinyin and in English. Future crosslinguistic studies might want to further examine the influence of Pinyin knowledge on Chinese EFL learners' decoding of English words and other aspects of English learning, which remains an under-researched area that could yield fruitful findings.

***RQ4: Do participants who have followed the programme of phonics instruction also show gains in their ability to recall (productive knowledge) and recognise (receptive knowledge) new English words?***

A simple, straightforward answer to this question is yes. Looking at the sample as a whole, intervention participants showed significantly greater gains than the comparison groups in three of the four vocabulary memorisation tests after the phonics instruction. These were the oral recall test, the written recall test and the aural recognition test. This suggests the effectiveness of the phonics instruction in promoting productive phonological knowledge, receptive phonological knowledge, and productive orthographic knowledge in learning new English words. However, the phonics instruction did not lead to significantly greater gains in the written recognition test, suggesting that the receptive orthographic knowledge may not be

sensitive to phonics instruction. This seems to suggest that decoding makes different degrees of contribution in the learning of different aspects of vocabulary knowledge. Nonetheless, these findings provide evidence for the role of decoding in L2 English vocabulary learning, and points to the causal relationship between the two variables. This echoes the findings of the FLEUR study (Woore et al., 2018), where the L2 phonics instruction also seemed to promote vocabulary learning for Year 7 English learners of French.

These findings are also of great pedagogical importance. Given the crucial role of vocabulary knowledge in L2 learning, Chinese EFL learners and teachers have generally paid a lot of attention to learning English vocabulary (Ma, 2009). Past research has demonstrated that their vocabulary learning has often been characterized by orthographic processing, while the phonological information is given less focus (Li, 2012). This can also be seen in the error analysis of the written recall test results, where many participants made positional errors (e.g. spelling *ploidy* as *pliody*) before the intervention. The findings of this study show that poor decoding can impair the efficiency of different aspects of vocabulary learning. In order to help learners acquire new vocabulary more efficiently, adequate phonics knowledge may be of more importance than teachers and students have previously believed, which could make decoding a powerful tool in learning new words.

## 9.3 Limitations

The current study sought to examine the effects of systematic phonics instruction on university-level EFL learners in China. Though the findings of this study provide encouraging evidence for the effectiveness of phonics instruction in promoting both decoding proficiency and vocabulary learning, and may be relevant to other university-level Chinese EFL learners, the findings are far from being generalizable to the wider population, given the vast number of English learners in China with varying ages and English proficiency levels. Moreover, the participants in this study were all English majors, thus may have had stronger motivation and more interest in learning English phonics compared to other EFL learners, and may not be a typical sample of university-level English learners. In order to obtain a more thorough understanding of the contribution of decoding to different aspects of English learning, studies involving EFL learners of different age and levels of English proficiency in China are still needed.

The findings of this study provide clear evidence for the contribution of decoding proficiency to different aspects of vocabulary learning, and also seem to support the hypothesis of a causal relationship between decoding proficiency and vocabulary learning. However, it needs to borne in mind that the vocabulary memorisation task in this study was conducted in a highly controlled environment. Though this task may bear some similarities with learning a word list, which is arguably one of most

popular ways of learning new vocabulary among Chinese EFL learners (Gu & Johnson, 1996), this does not reflect other forms of vocabulary learning, such as incidental vocabulary learning. Hence, future studies might want to look at the impact of systematic phonics instruction on other forms of vocabulary learning, which could further contribute to the understanding of the relationship between L2 decoding and vocabulary learning,

Some limitations on the research design also need to be acknowledged. Firstly, no delayed post-tests were conducted in this study, making it difficult to know whether the effects of the phonics instruction were maintained over time. Thus, even though the current study has demonstrated that the phonics instruction programme was effective in promoting both decoding proficiency and vocabulary learning results, the magnitude of such effects (durability of learning) needs to be interpreted with caution.

Secondly, though both the phonics instruction and the phonology instruction were delivered by the researcher herself, ruling out the teacher effect as a possible confounding variable, fidelity to condition was not monitored by an impartial observer. Possible bias of the researcher, though largely not intentional, needs to be acknowledged, which may potentially contribute to the better performance of the intervention group and confound the results. In addition, there may also be a difference between the interventions as delivered by a researcher who was totally committed, and the interventions as delivered by 'regular' teachers. Therefore,

large-scale ‘effectiveness’ trials (to use Education Endowment Foundation’s parlance) may be needed next, in order to see whether the intervention is ‘scalable’.

In the design of the phonics instruction programme, the lexical stress, which is often argued to be of crucial importance in English phonology (e.g. Field, 2005), was regrettably not explicitly instructed. Instead, the reduced vowel was instructed in combination with other phonemes in the teaching of common word endings (e.g. -el, -al). It is possible that explicit instruction in lexical stress may help learners in the decoding of vowel graphemes that should be pronounced as the schwa, which appeared to be particularly challenging for the participants in this study.

The above limitations, many which were largely the results of compromise due to the difficulty in scheduling, nonetheless need to be borne in mind when interpreting the findings of this study.

### 9.4 Directions for future research

The findings of this study, for the first time, have provided encouraging evidence for the effects of phonics instruction on Chinese university-level EFL learners. Phonics instruction not only promoted English decoding proficiency measured at both the word level and the grapheme level, but also promoted productive phonological knowledge, receptive phonological knowledge, and productive orthographic

knowledge in learning new English words. Given the potential benefits of phonics knowledge in L2 learning, it is worth exploring the effects of phonics instruction on younger L2 learners; this is because, if phonics instruction were found to be useful in promoting various aspects of English learning for younger Chinese students, it would potentially be useful to teach phonics at an earlier stage of English learning.

Another thing worth noting is that, phonics is usually considered as a tool for learning to read – certainly in the L1 context and that is perhaps tacitly the assumption in L2 contexts too. But in fact it may be that in L2, the key benefit of phonics instruction is not (or not only) improved reading comprehension but other variables – e.g. vocabulary learning as shown in the current study. Of course, without any measure of reading comprehension in this study, no conclusion regarding the effects of phonics instruction on L2 reading comprehension can be drawn. This points to the need of using a wider range of outcome measure in evaluating the effects of phonics instruction. Future studies may want to consider measuring the effects of phonics instruction on other aspects of L2 learning, which may help complement the understanding of how L2 phonics knowledge contribute to L2 learning.

**References**


Aarnoutse, C., Leeuwe, J. V., Voeten, M. and Oud, H. (2001). Development of decoding, reading comprehension, vocabulary and spelling during the elementary school years. *Reading and Writing: An Interdisciplinary Journal*, 14 (1), 61-89.

Abbott, M. (2006). ESL reading strategies: Differences in Arabic and Mandarin speaker test performance. *Language Learning*, 56 (4), 633-670.

Adair, J. G., Sharpe, D., & Huynh, C. L. (1989). Placebo, Hawthorne, and other artifact controls: Researchers' opinions and practices. *The Journal of Experimental Education*, 57(4), 341-355.

Akamatsu, N. (2003). The effects of first language orthographic features on second language reading in text. *Language Learning*, 53 (2), 207-231.

Altenberg, E. P. (2005). The judgment, perception, and production of consonant clusters in a second language. *International Review of Applied Linguistics in Language Teaching*, 43(1), 53-80.

Alves, U. K., & Magro, V. (2011). Raising awareness of L2 phonology: Explicit instruction and the acquisition of aspirated/p/by Brazilian Portuguese speakers. *Letras de hoje. Porto Alegre*. Vol. 46, n. 3 (jul./set. 2011), p. 71-80.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese/r/ and English /l/ and /r/. *Journal of Phonetics*, 32(2), 233-250.

August, D., McCardle, P., & Shanahan, T. (2014). Developing literacy in English language learners: Findings from a review of the experimental research. *School Psychology Review*, 43(4), 490-498.

Baddeley, A. D. (1992). Working Memory. *Science*, 255 (5044), 556-559.

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences*, 4(11), 417-423.

Baddeley, A. D., Gathercole, S. and Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105 (1). 158-173.

Baddeley, A. D. and Logie, R. H. (1999). Working memory: the multiple component model. In Miyake, A. and P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge: Cambridge University Press.

Baddeley, A.D. and Hitch, G, J. (1974). Working memory. In: G. H. Bower (ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory*. New York: Academic Press.

Ball. E. W. and Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and development? *Reading Research Quarterly*, 26 (1), 49-66.

Baron, J. and Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 2 (3), 386-393.

Bentin, S. and Frost, R.(1987). Processing lexical ambiguity and visual word

recognition in a deep orthography. *Memory and Cognition*, 15 (1), 13-23.

Berndt, R. S., Reggia, J. A. and Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments and Computers,* 19 (1), 1-9.

Berry, C. (2015). Read Write Inc. accelerates progress for all pupils. http://www.ruthmiskin.com/en/success-stories/discovery-bay-international-school-hong-kong/ (last accessed: 30/07/15)

Biederman, I. and Tsao, Y. On processing Chinese ideographs and English words: some implications from Stroop-test results. *Cognitive Psychology*, 11 (2), 125-132.

Bowers, P. G., Golden, J., Kennedy, A., & Young, A. (1994). Limits upon orthographic knowledge due to processes indexed by naming speed. In V. W. Berninger (Ed.), *The varieties of orthographic knowledge* (Vol. 1, pp. 173–218). Dordrecht, Netherlands: Kluwer Academic.

Broselow, E., Chen, S. I., & Wang, C. (1998). The emergence of the unmarked in second language phonology. *Studies in second language acquisition*, 20(2), 261-280.

Broselow, E., & Park, H. B. (1995). Mora conservation in second language prosody. *Phonological acquisition and phonological theory*, 151-168.

Brown, G. D. A. and Hulme, C. (1992). Cognitive psychology and second-language processing: The role of short-term memory. In R. J. Harris (Ed.), *Cognitive Approaches to Bilingualism*. New York: Elsevier.

Byrne, B. and Fieldling-Barnsely, R. (1995). Evaluation of a program to teach phonemic awareness to young children: A 2- and 3- year follow-up and a new preschool trial. *Journal of Educational Psychology*, 87 (3), 488-503.

Cai, J. (2004). ESP and the direction of China's college English teaching. *Foreign Language World*, 2, 22-28.

Chen, Z. and Lee, K. F. (2000). A new statistical approach to Chinese Pinyin input. Proceedings of the 38th Annual Meeting on Association for Computational linguistics, 241-247.

Cheung, H., & Ng, L. K. H. (2003). Pinyin and phonotactics affect the development of phonemic awareness in English-Cantonese bilinguals. In C. McBride-Chang & H.C. Chen (Eds.), *Reading development in Chinese children*. Westport, CT: Greenwood Press.

Cheung, R. L. (1966). Mandarin phonological structure. *Journal of Linguistics*, 2 (2), 135-158.

Chiesa, M., & Hobbs, S. (2008). Making sense of social research: How useful is the Hawthorne Effect?. *European Journal of Social Psychology*, 38(1), 67-74.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*.

Cisero, C. A. and Royer, J. (1995). The development and cross-language transfer of phonological awareness. *Contemporary Educational Psychology*, 20 (3), 275-303.

Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155.

Collinge, N. E. (Ed). (2002). *An Encyclopedia of Language*. London: Taylor and

Francis.

Coltheart, M., & Leahy, J. (1996). Assessment of lexical and nonlexical reading abilities in children: Some normative data. *Australian Journal of Psychology*, 48(3), 136-140.

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: human perception and performance*, 20(6), 1197.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1), 204.

Comeau, L., Cormier, P., Grandmaison, E., and Lacroix, D. (1999). A longitudinal study of phonological processing skills in children learning to read in a second language. *Journal of Educational Psychology*, 91 (1), 29-43.

Commissaire, E., Duncan, L. G., & Casalis, S. (2011). Cross-language transfer of orthographic processing skills: a study of French children who learn English at school. *Journal of Research in Reading*, 34(1), 59-76.

Cook, V. (2004). *The English Writing System*. London: Arnold.

Cook, V. and Bassetti, B. (Eds). (2005). *Second Language Writing Systems*. Clevedon: Cromwell.

Cox, N. (2011). School makes fast progress to move from satisfactory to good in under two years. http://www.ruthmiskin.com/en/success-stories/brompton-westbrook-primary-sch

ool/ (last accessed: 30/07/18)

Cruttenden, A. 2008. *Gimson's Pronunciation of English (7th ed.)*. London: Arnold.

Crystal, D. (1997). *A Dictionary of Linguistics and Phonetics (4th edn)*. Oxford: Blackwell Publishers Inc.

DaFontoura, H. A. and Siegel, L. S. (1995). Reading, syntactic and working memory skills of bilingual Portuguese-English Canadian children. *Reading and Writing: An interdisciplinary Journal*, 7 (2), 139-153.

Daniels, P. T. and Bright, W. (Eds). (1996). *The World's Writing Systems*. New York: Oxford University Press.

Dauer, R. M. (2005). The lingua franca core: A new model for pronunciation instruction?. *Tesol Quarterly*, 39(3), 543-550.

DeFrancis, J. (1989). *Visible Speech: The Diverse Oneness of Writing System*. Honolulu: University of Hawaii.

Deterding, D. (2006). The pronunciation of English by speakers from China. *English World-Wide*, 27(2), 175-198.

Deterding, D. (2010). EFL-based pronunciation teaching in China. *Chinese Journal of Applied Linguistics*, 33 (6), 3-15.

Doughty, C. J. (2003). Instructed SLA: constraints, compensation, and enhancement. In Doughty, C. J. and Long, M. H. (eds), *The Handbook of Second Language Acquisition*. Malden, MA: Blackwell.

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38, 78–103.

Dunn, L. M., & Dunn, D. M. (2009). *The British picture vocabulary scale*. GL Assessment Limited.

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of educational psychology*, 85(3), 453.

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, 16(2), 227-252.

Ehri, L. (1992). Reconceptualizing the development of sight word reading and tis relationship to recoding. In P. Gough, L. Ehri and R. Treiman (Eds.), *Reading Acquisition*. Hillsdale, NJ: Erlbaum.

Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, N. C. and Sinclair, S. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *Quarterly Journal of Experimental Psychology*, 49 (A). 234-250.

Ellis, N. (1997). Vocabulary acquisition: Word structure, collocation, word-class, and meaning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description Acquisition and Pedagogy* (pp. 122-139). Cambridge: Cambridge University Press.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in*

*second language acquisition*, 24(2), 143-188.

Erler, L. (2003). Reading in a foreign language--- near-beginner adolescents' experience of French in English secondary schools. Unpublished D.Phil Thesis, Oxford University, Department of Educational Studies.

Erler, L. (2004). Near-beginner learners of French are reading at a disability level. *Francophonie*, 9-15.

Erler, L., & Macaro, E. (2011). Decoding ability in French as a foreign language and language learning motivation. *The Modern Language Journal*, 95(4), 496-518.

Evita. (2014, May 04). The most scientific way to recite English vocabulary. Retrieved from: http://evita6804.pixnet.net/blog/post/123514427-%E8%83%8C%E5%96%AE%E8%A9%9E%E6%9C%80%E7%A7%91%E5%AD%B8%E7%9A%84%E6%96%B9%E6%B3%95

Fan, K. Y., Gao, J. Y. and Ao. X. P. (1984). Pronunciation principles of the Chinese characters and alphabetic writing scripts. *Chinese Character Reform*, 3, 23-27.

Feldman, L. B. and Siok, W. W. T. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23 (3), 776-781.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, 16(2), 234-248.

Ferrand, L. (2000). Reading aloud polysyllabic words and nonwords: The syllabic

length effect reexamined. *Psychonomic Bulletin & Review*, 7(1), 142-148.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Timonium, MD: York Press.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, 97(5), 3125-3134.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL quarterly*, 39(3), 399-423.

Frisch, S. A., Large, N. R., Zawaydeh, B., & Pisoni, D. B. (2001). Emergent phonotactic generalizations in English and Arabic. *Typological studies in Language*, 45, 159-180.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1), 104.

Fuchs, D., Fuchs, L. S., Thompson, A., Al Otaiba, S., Yen, L., & Yang, N. J. (2000). *Peer-assisted learning strategies in reading: A teacher manual* (Rev. ed.). Available from Douglas Fuchs, Box 328 Peabody, Vanderbilt University, Nashville, TN 37203 (or http://www.peerassistedlearningstrategies.net).

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese universities. *The modern language journal*,

88(2), 229-244.

Gathercole, S. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory and Cognition*, 23 (1), 83-94.

Gaterhcole, S. and Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, 28 (2), 336-360.

Gathercole. S. and Baddeley, A. D. (1990a). Phonological memory deficits in language-disordered children: Is there a casual connection? *Journal of Memory and Language*, 29 (3), 336-360.

Gathercole, S. and Baddeley, A. D. (1990b). The role of phonological memory in vocabulary acquisition: A study of young children learning arbitrary names of toys. *British Journal of Psychology*, 81 (4), 439-454.

Gathercole, S., Hitch, G. J., Service, E., and Martin, A. J. (1997). Short-term memory and new word learning in children. *Developmental Psychology*, 33 (6), 966-979.

Gathercole, S., Service, E., Hitch, G. J., Adams, A-M and Martin, A. J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology*, 13 (1), 65-77.

Gathercole, S. and Thorn, A. (1998). Phonological short-term memory and foreign language learning. In: A. F. Healy and L. E. Bourne (eds.), *Foreign Language Learning: Psycholinguistic Studies on Training and Retention*. Mahwah: New Jersey.

Gathercole, S., Tiffany, C., Briscoe, J., Thorn, A. and The ALSPAC TEAM. (2005). Developmental consequences of phonological loop deficits during early childhood: A longitudinal study. *Journal of Child Psychology and Psychiatry*, 46 (6), 598-611.

Gathercole, S., Willis, C., Emslie, H and Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Development Psychology*, 28 (5), 887-898.

Geva, E. and Siegel, L. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing: An Interdisciplinary Journal*, 12 (1), 1-30.

Geva, E., Wade-Woolley, L., & Shany, M. (1997). Development of reading efficiency in first and second language. *Scientific Studies of Reading*, 1(2), 119-144.

Geva, E., & Wang, M. (2001). The development of basic reading skills in children: A cross-language perspective. *Annual Review of Applied Linguistics*, 21, 182-204.

Goff, D. A., Pratt, C., & Ong, B. (2005). The relations between children's reading comprehension, working memory, language skills and components of reading decoding in a normal sample. *Reading and writing*, 18(7-9), 583-616.

Gottardo, A., Yan, B., Siegel, L. S. and Wade-Woolley, L. (2001). Factors related to English reading performance in children with Chinese as a first language: More evidence of cross-language transfer of phonological processing. *Journal of Educational Psychology*, 93 (3), 530-542.

Gontijo, P. F., Gontijo, I., & Shillcock, R. (2003). Grapheme—phoneme probabilities in British English. *Behavior Research Methods, Instruments & Computers*, 35(1), 136-157.

Grainger, J., & Dijkstra, T. (1992). On the representation and use of language information in bilinguals. *Advances in Psychology*, 83, 207-220.

Grant, A., Gottardo, A., & Geva, E. (2011). Reading in English as a first or second language: The case of grade 3 Spanish, Portuguese, and English speakers. *Learning Disabilities Research & Practice*, 26(2), 67-83.

Gu, Y., & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. *Language learning*, 46(4), 643-679.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *The Quarterly Journal of Experimental Psychology*, 56A (7), 1213-1236.

Hamada, M. and Koda, K. (2008). Influence of first language orthographic experience on second language decoding and word leaning. *Language Learning*, 58 (1), 1-31.

Hamada, M. and Koda, K. (2010). The role of phonological decoding in second language word-meaning inference. *Applied Linguistics*, 31 (4), 513-531.

Hansen, J. G. (2001). Linguistic constraints on the acquisition of English syllable codas by native speakers of Mandarin Chinese. *Applied Linguistics*, 22(3), 338-365.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in*

*second language acquisition*, 21(2), 303-317.

Henry, M. K. (1989). Children's word structure knowledge: Implications for decoding and spelling instruction. *Reading and Writing*, 1(2), 135-152.

Ho, C. S. H. and Bryant, P. (1997). Phonological skills are important in learning to read Chinese. *Developmental Psychology*, 33 (6), 946-951.

Hong, L. P. C. (2008). Investigating the most frequently-used and most-useful vocabulary language learning strategies among Chinese EFL postsecondary students in Hong Kong. In Global Practices of Language Teaching: Proceedings of the 2008 International Online Language Conference (IOLC 2008) (p. 209). Universal-Publishers.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing,* 2(2), 127-160.

Howlin, P., Goode, S., Hutton, J., & Rutter, M. (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry*, 45(2), 212-229.

Hsueh-Chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, 13(1), 403-30.

Hu, C. F. (2003). Phonological memory, phonological awareness, and foreign language word learning. *Language learning*, 53(3), 429-462.

Hu, G. (2005). English language education in China: Policies, progress, and problems. *Language policy*, 4(1), 5-24.

Huang, H. S. and Hanley, J. R. (1994). Phonological awareness and visual skills in learning to read Chinese and English. *Cognition*, 54 (1), 73-98.

Huang, J. (2005). A diary study of difficulties and constraints in EFL learning. *System*, 33(4), 609-621.

Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. *Journal of experimental child psychology*, 82(1), 2-28.

Hulme, C., Maughan, S. and Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Languag*e, 30 (6), 685-701.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In *Vocabulary and Applied Linguistics*, P. J. L. Arnaud & H. Bejoint, Eds. London: Macmillan, 1992, 113-125.

Hulstijn, J. H. (2003). Incidental and intentional learning. *The handbook of second language acquisition*, 349-381.

Jenkins, J. 2000. *The Phonology of English as an International Language*. Oxford, UK: Oxford University Press.

Jenkins, J. 2007. *English as a Lingua Franca: Attitude and Identity*. Oxford, UK: Oxford University Press.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A metaanalysis. *Language Learning*, 64(1), 160-212.

Jian, T., Sachs, J. D., & Warner, A. M. (1996). Trends in regional inequality in China (No. w5412). National Bureau of Economic Research.

Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second

language. *Studies in Second Language Acquisition*, 24 (4), 617-637.

Johnson, R. and Watson, J. (2005). The effects of synthetic phonics teaching on reading and spelling attainment: A seven year longitudinal study. Available at: http://dera.ioe.ac.uk/14793/1/0023582.pdf [Last accessed: 01/09/18]

Just, M. A. and Carpenter, P. A. (1975). A theory of reading: from eye fixation to comprehension. *Psychological Review*, 87 (4), 329-354.

Kanbur, R., & Zhang, X. (2005). Fifty years of regional inequality in China: a journey through central planning, reform, and openness. *Review of development Economics*, 9(1), 87-106.

Katz, L. and Frost, R. (1992). Reading in different orthographies: The orthographic depth hypothesis. In: R. frost and L. Katz (eds.), *Orthography, Phonology, Morphology, and Meaning*. Amsterdam: Elsevier.

Keenan, J. M., Betjemann, R. S. and Olsen, R. K. (2008). Reading comprehension tests vary in the skills they assess: differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12 (3), 281-300.

Kellerman, E. (1984). If at first you do succeed… in S. Gass and N. Madden (eds.) *Input in second language acquisition*, pp. 345-353. Rowley, MA: Newbury House.

Kleiman, G. M. (1975). Speech recording in reading. *Journal of Verbal Learning and Verbal Behavior*, 14 (4), 329-339.

Koda, K. (1988). Cognitive process in second language reading: Transfer of L1 reading skills and strategies. *Second Language Research*, 4 (2), 133-156.

Koda, K. (1997). Orthographic knowledge in L2 lexical processing. *Second Language Vocabulary Acquisition*, (1) 35-52.

Koda, K. (1999). Development of L2 intraword orthographic sensitivity and decoding skills. *Modern Language Journal*, 83 (1), 51-64.

Koda, K. (2000). Cross-linguistic variations in L1 morphological awareness. *Applied Psycholinguistics*, 21 (3), 297-320.

Koda, K. (2005). Learning to read across writing systems: Transfer, metalinguistic awareness and second-language reading development. In V. Cook & B. Bassetti (Eds.), *Second language writing systems*, pp.311–334. Clevedon: Multilingual Matters.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language learning*, 57, 1-44.

Koda, K., & Reddy, P. (2008). Cross-linguistic transfer in second language reading. *Language Teaching*, 41(4), 497-508.

Kuhl, P. K. (2000). A new view of language acquisition. Proceedings of the National Academy of Sciences, 97(22), 11850-11857.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2007). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). Philosophical Transactions of the Royal Society B: Biological Sciences.

Lado, R. (1957). Linguistics across cultures: Applied linguistics for language teachers. Ann Arbor: University of Michigan Press.

Laufer, B. (1988). The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, 2 (2), 113-132.

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

Lengeris, A. (2009). Individual differences in second-language vowel learning (Doctoral dissertation, UCL (University College London)).

Lam, S., Yim, P., Law, J. S. F. and Cheung, R. W. Y. (2004). The effects of competition on achievement motivation in Chinese classrooms. *British Journal of Education Psychology*, 74 (2), 281-296.

Leech, G., & Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Leong, C. K., Cheng, P. W., & Tan, L. H. (2005). The role of sensitivity to rhymes, phonemes and tones in reading English and Chinese pseudowords. *Reading and Writing*, 18(1), 1-26.

Li, C. N., & Thompson, S. A. (1981). *A grammar of spoken Chinese: A functional reference grammar*.

Li, H., & Rao, N. (2000). Parental influences on Chinese literacy development: A comparison of preschoolers in Beijing, Hong Kong and Singapore. *International Journal of Behavioral Development*, 24, 82–90.

Li, S. (2012). Decoding ability and vocabulary acquisition in second language English- A study of Chinese advanced EFL learners. Unpublished Master Dissertation, Oxford University, Department of Education.

Liaw, M. L. (2003). Integrating phonics instruction and whole language principles in an elementary school EFL classroom. *English Teaching & Learning*, 27(3), 15-34.

Lin, D., McBride-Chang, C., Shu, H., Zhang, Y., Li, H., Zhang, J. & Levin, I. (2010). Small wins big: Analytic Pinyin skills promote Chinese word reading. *Psychological Science,* 21(8), 1117-1122.

Liow, S. J. R., & Poon, K. K. (1998). Phonological awareness in multilingual Chinese children. *Applied Psycholinguistics*, 19(3), 339-362.

Livingston, M and Flaherty, J. (1997). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly*, 23 (3), 263-284.

Lord, G. (2005). (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania*, 557-567.

Lloyd S. (1992). *The jolly phonics handbook*. Essex, United Kingdom: Jolly Learning Ltd.

Lü, C. (2017). The Roles of Pinyin Skill in English-Chinese Biliteracy Learning: Evidence From Chinese Immersion Learners. *Foreign Language Annals*, 50(2), 306-322.

Ma, Q. (2009). *Two empirical studies of Chinese learners' approaches to vocabulary acquisition*. In: Q. Ma (ed.), Bern: Peter Lang.

Ma, Z. (2007). A functional analysis of the speech perception of Chinese EFL learners. *Modern Foreign Languages*, 30 (1), 80-110.

Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *The Modern Language Journal*, 90(3), 320-337.

Maehler, C. and Schuchardt, K. (2011). Working memory in children with learning disability: rethinking the criterion of discrepancy. *International Journal of Disability, Development and Education*, 58 (1), 5-17.

McBride-Chang, C., Bialystok, E., Chong, K. K., & Li, Y. (2004). Levels of phonological awareness in three cultures. *Journal of experimental child psychology*, 89(2), 93-111.

McBride-Chang, C., Lin, D., Liu, P. D., Aram, D., Levin, I., Cho, J. R., Shu, H. & Zhang, Y. (2012). The ABC's of Chinese: Maternal mediation of Pinyin for Chinese children's early literacy skills. *Reading and Writing*, 25(1), 283-300.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, 55(3), 381-401.

McLaughlin, B. (1990). Restructuring. Applied Linguistics, 11(2), 113–128.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, 34(1), 114-135.

Michas, I. C. and Henry, L. A. (1994). The link between phonological memory and vocabulary acquisition. *British Journal of Developmental Psychology*, 12 (2), 147-164.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In

I. Vedder, I. Bartning, & M. Martin (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (pp. 211- 232). Second Language Acquisition and Testing in Europe Monograph Series 1.

Milton, J (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In *L2 Vocabulary Acquisition, Knowledge and Use*. Bardel, C., Lindqvist, C. & Laufer, B. (Eds.), 57-78.

Mok, K. H., Wong, Y. C. and Zhang, X. (2009). When marketisation and privatisation clash with socialist ideas: Educational inequality in urban China. *International Journal of Educational Development*, 29 (5), 505-512.

Muljani, D., Koda, K., & Moates, D. R. (1998). The development of word recognition in a second language. *Applied Psycholinguistics*, 19(1), 99-113.

Muter, V., Hulme, C., Snowling, M. and Taylor, S. (1998). Segmentation, not rhyming, predicts early progress in learning to read: Erratum. *Journal of Experimental Child Psychology*, 71 (1), 3-27.

Nagy, W. E. (1995). *On the role of context in first-and second-language vocabulary learning*. Champaign, Ill.: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Nassaji, H. (2014). The role and importance of lower-level processes in second language reading. *Language Teaching*, 47(1), 1-37.

Nation, I. S. P. (1990). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups. National Institute of Child Health and Human Development, National Institutes of Health.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language-learning*. Cambridge: Cambridge University Press.

Olson, R. K., Wise, B. W., Johnson, M. C., & Ring, J. (1997). The etiology and remediation of phonologically based word recognition and spelling disabilities. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 305–326). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of educational psychology*, 98(3), 554.

Packard, J. L. (1990). Effects of time lag in the introduction of characters into the Chinese language curriculum. *The Modern Language Journal*, 74(2), 167-175.

Papagno, C., Valentine, T. and Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, 30 (3), 331-347.

Parrila, R., Kirby, J. R. and Mcquarrie, L. (2004). Articulation rate, naming speed,

verbal short-term memory, and phonological awareness: longitudinal predictors of early reading development? *Scientific Studies of Reading*, 8 (1), 3-26.

Pascale, M. J. E., Martine, B., Marina, L. P. and Debora, M. B. (2012). Cross-linguistic and cross-cultural effects on verbal working memory and vocabulary: Testing language minority children with an immigrant background. *Journal of Speech, Language, and Hearing Research*, 56 (2), 630-642.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford, UK: Blackwell.

Perfetti, C. A., & Lesgold, A. M. (1977). *Discourse comprehension and sources of individual differences*.

Quinn, J. G. (1994). Towards a clarification of spatial processing. *The Quarterly Journal of Experimental Psychology*, 47A (2), 465-480.

Rao, Z. (2002). Chinese students' perceptions of communicative and non-communicative activities in EFL classroom. *System*, 30(1), 85-105.

Rau, D., Chang, H. H. A., & Tarone, E. E. (2009). Think or sink: Chinese learners' acquisition of the English voiceless interdental fricative. *Language Learning*, 59(3), 581-621.

Roodenrys, S., Hulme. C. and Brown, G. (1993). The development of short-term memory span: separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology*, 56 (3), 431-442.

Rose, J. (2006). *Independent review of the teaching of early reading*.

Rowe, M. L. & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, 323(5916), 951-953.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental psychology*, 39(3), 484.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp.97– 110). New York: Guilford Press.

Schiff, R., & Calif, S. (2007). Role of phonological and morphological awareness in L2 oral word reading. *Language Learning*, 57(2), 271-298.

Schmidt, R. W. (1990). The role of consciousness in second language learning1. *Applied linguistics*, 11(2), 129-158.

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165–210). San Diego, CA: Academic Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing. I. Detection, search, and attention. *Psychological Review*, 84(1), 1–66.

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science*, 27(3), 525-559.

Senechal, M., Ouellette,.G.,& Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In S.B. Neuman & D. Dickinson (Eds.), *Handbook of early literacy research*: Vol.2 (pp. 173-182). New York: Guilford Press.

Service, E. (1992). Phonology, working memory and foreign language learning. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 45 (1). 21-50.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55(2), 151-218.

Shi, L. (2006). The successors to Confucianism or a new generation? A questionnaire study on Chinese students' culture of learning English. *Language, culture and curriculum*, 19(1), 122-147.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing. II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127–190.

Shu, H. and Anderson, R. C. (2000). Phonetic awareness: Knowledge of orthography-phonology relationships in the character acquisition of Chinese children. *Journal of Educational Psychology*, 92 (1), 56-62.

Siok, W. T., & Fletcher, P. (2001). The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. *Developmental psychology*, 37(6), 886.

Slavin, R. E., Cheung, A., Groff, C. and Lake, C. (2008). Effective reading programs

for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, 43 (3), 390-322.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, 79(4), 1391-1466.

So, D. and Siegel, L. S. (1997). Learning to read Chinese: Semantic, syntactic, phonological and working memory skills in normally achieving and poor Chinese readers. *Reading and Writing: An Interdisciplinary Journal*, 9 (1), 1-21.

Solso, R. L. and Juel, C. L. (1980). Positional frequency and versatility of bigrams for two-through nine letter English words. *Behavior Research Methods & Instrumentation*, 12 (3), 297-343.

Sparks, R., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term crosslinguistic transfer of skills from L1 to L2. *Language Learning*, 59(1), 203-243.

Speciale, G., Ellis, N. C. and Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25 (2), 293-321.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Sturm, J. L. (2013). Explicit phonetics instruction in L2 French: A global analysis of improvement. *System*, 41(3), 654-662.

Swain, M. (2005). The output hypothesis: Theory and research. In *Handbook of*

*research in second language teaching and learning* (pp. 495-508). Routledge.

Taft, M., Zhu, X. and Peng, D. (1999). Positional specificity of radicals in Chinese character recognition. *Journal of Memory and Language*, 9 (2), 182-198.

Tan, L. H., Spinks, J.A., Feng, C. M., Soik, W. T., Perfetti, C. A., Xiong, J., Fox, P. T. and Gao, J. H. (2003). Neural systems of second language reading are shape by native language. *Human Brain Mapping*, 18 (3), 158-166.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and language*, 26(6), 608.

Taylor, I. and Taylor, M. (1995). *Writing and Literacy in Chinese, Korean and Japanese*. Amsterdam: John Benjamins.

Thompson, G. (1996). Some misconceptions about communicative language teaching. *ELT journal*, 50(1), 9-15.

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 36(3), 326-344.

Torgerson, C. J., Torgerson, D. J., & Director, Y. T. U. (2013). *Randomised trials in education: An introductory handbook*. London: Education Endowment Foundation.

Tomlinson, B. (2011a). Glossary. In B. Tomlinson (Ed.), *Materials Development in Language Teaching* (pp. ix-xviii). Cambridge: Cambridge University Press.

Verhoeven, L. T. (1990). Acquisition of reading in a second language. *Reading research quarterly*, 90-114.

Verhoeven, L. (2000). Components in early second language reading and

spelling. *Scientific Studies of reading*, 4(4), 313-330.

Verhoeven, L. (2011). *Ethnic minority children acquiring literacy* (Vol. 4). Walter de Gruyter.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, C. A., Baker, T. A., Burgess, S. R., Donahue, J. and Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from Beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology*, 33 (3), 468-479.

Walley, A. C. (1993). The role of vocabulary development in children′ s spoken word recognition and segmentation ability. *Developmental review*, 13(3), 286-350.

Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics*, 25(3), 315-339.

Walter, C. (2007). First-to-second language reading comprehension: not transfer, but access. *International Journal of Applied Linguistics*, 17(1), 14-37.

Walter, C. (2008). Phonology in second language reading: Not an optional extra. *TESOL quarterly*, 42(3), 455-474.

Wang, M. and Geva, E. (2003a). Spelling performance of Chinese children using English as a second language: lexical and visual-orthographic processes. *Applied Psycholinguistics*, 24 (1). 1-25.

Wang, M., & Geva, E. (2003b). Spelling acquisition of novel English phonemes in Chinese children. *Reading and Writing*, 16(4), 325-348.

Wang, M., Koda, K. and Perfetti, C. A. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: a comparison of Korean and Chinese English L2 learners. *Cognition*, 87 (2), 129-149.

Wasik, B.A., & Madden, N. A. (1995). *Success for All tutoring manual*. Baltimore: Johns Hopkins University, Center for Research on the Education of Students Placed at Risk.

Weinberger, S. (1987). The influence of linguistic context on syllable simplification. *Interlanguage phonology: The acquisition of a second language sound system*, 401-417.

Weismer, S. E. and Hesketh, L. J. (1996). Lexical learning by children with specific language impairment: Effects of linguistic input presented at varying speaking rates. *Journal of Speech and Hearing Research*, 39, 177-190.

White, T. G., Graves, M. F. and Slater, W, H. (1990). Growth of reading vocabulary in diverse elementary schools: decoding and word meaning. *Journal of Educational Psychology*, 82 (2), 281-290.

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition*, 51(1), 91-103.

Wolf, M., Bowers, P. G. and Biddle, K. (2000). Naming-speed processes, timing and reading: a conceptual review. J*ournal of Learning Disabilities*, 33 (4), 387-407.

Woodcock, R. W. (2011). *Woodcock Reading Mastery Tests: WRMT-III*. Pearson.

Woore, R. (2009). Beginners' progress in decoding L2 French: Some longitudinal

evidence from English modern foreign languages classrooms. *Language Learning Journal*, 37(1), 3-18.

Woore, R. (2011). Investigating and developing beginner learners' decoding proficiency in second language French: an evaluation of two programmes of instruction (Doctoral dissertation, University of Oxford).

Woore, R. (2014). Beginner learners' progress in decoding L2 French: Transfer effects in typologically similar L1-L2 writing systems. *Writing Systems Research*, 6(2), 167-189.

Woore, R., Graham, S., Porter, A., Courtney, L., & Savory, C. (2018). Foreign Language Education: Unlocking Reading (FLEUR)-A study into the teaching of reading to beginner learners of French in secondary school.

Xu, F., & Ren, P. (2004). The relationship between Chinese children's phonological awareness and Pinyin skill. *Chinese Journal of Applied Psychology*, 10(4), 22–27.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of second language writing*, 15(3), 179-200.

Zhang, X. and Kanbur, R. (2005). Spatial inequality in education and health care in China. *China Economic Review*, 16 (2), 189-204.

Zhou, G. (2002). Statistical spoken dialog system. U.S. Patent Application No. 09/891,224.

**Appendix 1. Contents in the decoding instruction programme**

| | |
|---|---|
| Consonant graphemes | <f> <ff> <ph> (/f/); <l> <ll> <le> (/l/); <m> <mm> <mb> (/m/); <n> <nn> <kn> (/n/); <r> <rr> <wr> (/r/); <s> <ss> <c> <se> <ce> (/s/); <v> <ve> (/v/); <z> <zz> <s> <ze> (/z/); <sh> <ti> <ci> (/ʃ/); <th> (/θ/, /ð/); <ng> (/ŋ/) <nk> (ŋk/); <b> <bb> <bt> (/b/); <c> <k> <ck> <ch> (/k/); <d> <dd> <ed> (/d/); <g> <gg> (/g/); <h> (/h/); <j> <g> <ge> <dge> (/dʒ/); <p> <pp> (/p/); <qu> (/kw/); <t> <tt> <ed> (/t/); <w> <wh> (/w/); <x> (/kx/); <y> (/j/); <ch> <tch> (/ʧ/) |
| Vowel graphemes | <a> (/æ/); <e> <ea> (/e/); <i> <y> (/ɪ/); <o> (/ɒ/); <u> (/ʌ/); <ay> <a-e[56]> <ai> <ey> <aigh> <eigh> (/eɪ/); <ee> <ae> <ea> <y> <e> (/iː/); <igh> <i-e> <ie> <i> <y> (/aɪ/); <ow> <o-e> <oa> <o> (/əʊ/); <oo> <u-e> <ue> <ew> (/uː/); <oo> (/u/); <ar> (/ɑː/); <or> <oor> <ore> <aw> <au> (/ɔː/); <air> <are> <ear> (/eə/); <ir> <ur> (/ɜː /); <ou> <ow> (/aʊ/); <oy> <oi> (/ɒɪ/); <ear> <eer> (/ɪə/); <ire> (/aɪə/); <ure> (/ʊə/) |
| Common word endings | <-le> <-il> <-el> <-al> (/əl/); <-ent> <-ant> (/ənt/); <-ence> <-ance> (/əns/); <-ive> (/ɪv/) <-ism> (/ɪzəm/) <-age> (/ɪdʒ/); <-ture> (/ʧə/) <-ure> /ʊə/ <-our> <-or> <-er> (/ə/); <-ous> (/əs/) <-ious> <-eous> (/ɪəs/) <-cious> <-tious> (/ʃəs/); <-able> (/əbl/) <-ably> (/əbli/) <-ible> (/ɪbəl/) <-ibly> (/ɪblɪ/); <-tion> <-sion> <-ssion> (/ʃən/) |

56 <a> followed by another grapheme, with <e> as the last letter in a word such as 'cake'

**Appendix 2. Graphemes taught each week**

| Week number | Graphemes instructed |
|---|---|
| 1 | <b> <bb> (/b/); <p> <pp> (/p/); <t> <tt> <ed> (/t/); <d> <dd> <ed> (/d/); <k> (/k/); <g> <gg> (/g/); <h> (/h/); <s> (/s/); <m> <mm> (/m/); <r> <rr> (/r/); <f> <ff> (/f/); <l> <ll> (/l/); |
| 2 | <sh> (/ʃ/); <th> (/θ/, /ð/); <mb> (/m/); <wr> (/r/); <ss> (/s/); <ng> (/ŋ/) <nk> (/ŋk/); <ck> (/k/); <tch> (/tʃ/); <a> (/æ/); <e> (/e/); <i> (/ɪ/); <o> (/ɒ/); <u> (/ʌ/) |
| 3 | <a> (/æ/); <e> (/e/); <i> (/ɪ/); <o> (/ɒ/); <u> (/ʌ/); <ay> (/eɪ/); <ee> (/iː/); <igh> (/aɪ/); <c> <ck> (/k/) |
| 4 | <ow> (/əʊ/); <oo>(/u/); <ar> (/ɑː/); <th> (/θ/); <ch> (/tʃ/); <kn> <n> <nn> (/n/); <qu> (/kw/); <wh> <w> (/w/); <x> (/ks/); <y> (/j/) |
| 5 | <y> (/ɪ/); <or> <oor> <ore> (/ɔː/); <ch> (/tʃ/); <kn> (/n/); <se> (/s/); <ir> <ur> (/ɜː/)； <z> <zz> <s> (/z/) |
| 6 | <ou> (/aʊ/); <oy> <oi> (/ɒɪ/); <kn> (/n/); <se> (/s/); <bt> (/t/); <j> <dge> <ge> (/dʒ/); <wh> (/w/); <tch> (/tʃ/); <z> <zz> <s> <ze> (/z/) |
| 7 | <ay> <a-e> (/eɪ/); <ie> <i-e> (aɪ/); <ee> <ea> <ae> (/iː/); <ch> (/tʃ/); <se> <c> (/s/) |
| 8 | <ow> <o-e> (/əʊ/); <or> <aw> <au> (/ɔː/); <ear> <eer> (/ɪə/), <ce> <se> (/s/) |
| 9 | <air> <are> (/eə/); <ir> <ur> (/ɜː/); <ou> <ow> (/aʊ/); <sh> <ti> <ci> (/ʃ/); |
| 10 | <ai> <ey> <aigh> <eigh> (/eɪ/); <ow> <o> <oa> (/əʊ/); <oo> <u-e> <ue> <ew> (/uː/); <ire> (/aɪə/); |
| 11 | <-le> <-il> <-el> <-al> (/əl/); <-ent> <-ant> (/ənt/); <-ence> <-ance> (/əns/); <-ive> (/ɪv/) <-ism> (/ɪzəm/) <-age> (/ɪdʒ/); <-ture> (/tʃə/) <-ure> /ʊə/ <-our> <-or> <-er> (/ə/) |
| 12 | <-able> (/əbl/) <-ably> (/əbli/) <-ible> (/ɪbəl/) <-ibly> (/ɪblɪ/); <-tion> <-sion> <-ssion> (/ʃən/) |

**Appendix 3. Contents in the phonology instruction programme**

*Contents for University A and University B each week*

| Week number | Content |
|---|---|
| 1 | Concepts: syllables, stress and rhythm, phonemes: stops (/b/ /p/ /t/ /d/ /k/ /g/) |
| 2 | Fricatives and affricates (/f/ /v/ /θ/ /ð/ /s/ /z/ /h/ /tʃ/ /dʒ/) |
| 3 | Nasals, approximants and laterals (/m/ /n/ /ŋ/ /w/ /j/ /r/ /l/) |
| 4 | Front vowels and central vowels (/i:/ /ɪ/ /e/ /æ/ /ɜ:/ /ə/) |
| 5 | Back vowels (/u:/ /ʊ/ /ɔ:/ /ɒ/ /a:/ /ʌ/) |
| 6 | Diphthongs (/ɪə/ /eə/ /ʊə/ /eɪ/ /aɪ/ /ɒɪ/ /əʊ/ /aʊ/) |
| 7 | Stressed and unstressed syllables, stressed and unstressed words in a sentence |
| 8 | Strong forms and weak forms |
| 9 | Linking |
| 10 | Rhythm of English speech |
| 11 | Types of intonation and intonation units in English |
| 12 | Functions and uses of English intonation |

*Contents for University C each week*

| Week number | Content |
|---|---|
| 1 | Front vowels (/i:/ /ɪ/ /e/ /æ/) |
| 2 | Central and back vowels (/ɜ:/ /ə/ /ʌ/ /u:/ /ʊ/ /ɔ:/ /ɒ/ /a:/) |
| 3 | Diphthongs (/ɪə/ /eə/ /ʊə/ /eɪ/ /aɪ/ /ɒɪ/ əʊ/ /aʊ/) |
| 4 | Plosives, nasals and lateral (/b/ /p/ /t/ /d/ /k/ /g/ m/ /n/ /ŋ/ /l/) |
| 5 | Fricatives and affricates (/f/ /v/ /θ/ /ð/ /s/ /z/ /h/ /tʃ/ /dʒ/) |
| 6 | Approximants (/r/ /w/ /j/) |
| 7 | Syllables |
| 8 | Word stress, sentence stress and rhythm |
| 9 | Sound linking and sense group |
| 10 | Intonation |
| 11 | Speaking for success (vocal delivery, summing up and review) |
| 12 | Speaking for success (public speaking) |

**Appendix 4. Lesson plan example for the phonics instruction programme**

Lesson plan for GPC <a> = /æ/

Phonics lesson activities:

1. *Say grapheme <a>*
   - The teacher utters the sound /æ/ to the students.
   - Then the teacher holds up pictures of apple, ant, astronaut and acrobat in turn. The teacher repeats the sound /æ/ at the start when saying the name of each picture:
     a-a-a-apple, a-a-a-ant, a-a-a-astronaut, a-a-a-acrobat
   - The teacher asks the students to repeat after her. The teacher gives pronunciation tips of the phoneme /æ/, and invite some students to stand up and pronounce the phoneme /æ/. The teacher then gives feedback of students' pronunciations.

2. *Read grapheme <a>*
   - The teacher writes down the word 'apple' on the board.
   - Then the teacher runs her finger around the grapheme <*a*> and reads /æ/- /æ/- /æ/- /æ/. The students follow the teacher and read /æ/- /æ/- /æ/- /æ/.
   - The teacher shows the picture to the students and asks the students to say 'apple'. Then the teacher shows the grapheme <*a*> and asks the students to say /æ/. Repeat the procedure until the students are fluent.

3. *Write IPA* /æ/
   - The teacher writes the IPA /æ/ and reads it aloud. The students follow the teacher and do the same.

4. *Word segmenting*
   - The teacher shows a list of words containing the GPC <a> = /æ/, and spells the words phoneme by phoneme, emphasizing the phoneme /æ/ in them: /m/-/æ/-/n/, man; /m/-/æ/-/d/, mad; /h/-/æ/-/m/, ham, etc.
   - The students repeat after the teacher and spell the words phoneme by phoneme, /m/-/æ/-/n/, man; /m/-/æ/-/d/, mad; /h/-/æ/-/m/, etc.
   - The teacher asks the student to count the number of phonemes in different words: hat, map, apple, etc.

5. *Sound blending (with graphemes <m> <t> <s> <d> that have been instructed)*
   - Practice reading graphemes <a> <m> <t> <s> <d> at speed. The teacher puts the cards with<a> <m> <t> <s> <d> on the board, pointing to each card as she says the sounds and asks the students to follow her.

- Then the teacher puts the cards closer as she says the sounds more quickly. The students follow her.
- The teacher puts the cards together and forms the word *mat*. Then the teacher sweeps her pointer from left to right as she reads the word. The students repeat after her. Then the teacher practises other words with students, such as sat, mad, mat, etc.
- The students are formed into groups of two. Student 1 makes a word with the above graphemes, pointing to each grapheme and asks student 2 to pronounce the phoneme. Then student 2 says the whole word. Students take turns to make words.

Module lesson activities:

Students read structured reading materials containing the GPC <a> = /æ/ in the following three ways:

- One student is invited to read one paragraph. The teacher gives the student feedback on his/her pronunciation.
- The students read the article in chorus.
- The students work in pairs and read the article together.

The teacher encourages the students to find words containing the GPC <a> = /æ/ in the article. The teacher also invites the students to think of other examples containing the GPC <a> = /æ/.

## Appendix 5. Lesson plan example for the phonology instruction programme

Lesson plan for phoneme /æ/

1. *Learn the IPA symbol*
   - The teacher writes down the IPA symbol /æ/ on the board.
   - The teacher utters the phoneme /æ/, and the students repeat after her. Some students are invited to pronuce the phoneme /æ/ to the whole class, and the teacher gives feedback on their pronunciation. The teacher summarises the pronunciation tips of the phoneme /æ/.

2. *Compare with similar-sounding phonemes*
   - Students are asked to compare the phoneme /æ/ with similar-sounding phonemes /e/ and /ə/.
   - The teacher writes down the three IPA symbols on the board, /æ/, /e/ and /ə/. Then the teacher says one of the phonemes and asks the students to point out which phoneme it is.
   - The teacher summarises the differences of the three phonemes.
   - The students listens to a list of words containing phoenmes /æ/, /e/ or /ə/ (e.g. apple, bed, teacher, head, map, etc.). Then the students are asked to point out which of the three phonemes appears in each word.
   - The students work in pairs. Student 1 gives an example of a word containing phonemes /æ/, /e/ and /ə/. Student 2 is asked to select the correct phoneme in the word. Students take turns to give examples.

3. *Read the phoneme in a paragraph*
   - The students are given a paragraph containing words with phoneme /æ/.
   - The students listen to a recording of the paragphrah and repeat after it in chorus / one by one. The teacher gives feedback on their pronunciations.

**Appendix 6. Acceptable pronunciations for the stimuli in the decoding test**

| | | | | |
|---|---|---|---|---|
| Form A | bab /bæb/ | op /ɒp/ | dee /diː/ | bim /bɪm/ |
| | tay /teɪ/ | yee /jiː/ | pog /pɒg/ or /pɔːg/ | shum /ʃʌm/ |
| | plip /plɪp/ | dud's /dʌdz/ | whie /waɪ/ or /wiː/ | bufty /bʌftiː/ |
| | vunhip /vʌnhɪp/ | knaf /næf/ | twem /twem/ | adjex /adjeks/ |
| | yeng /yeŋ/ | laip /leɪp/ | zirdn't /zɜːdent/ or /zɜːdənt/ | straced /streɪst/ |
| | cedge /sedʒ/ | wrey /reɪ/ | whumb /wʌm/ | knoink /nɒɪŋk/ |
| | bafmotbem /bæfmɒtbem/ or /bæfməʊtbem/ or /bæfmɒtbəm/ or /bæfməʊtbəm/ | monglustamer /mɒnglʌstəmə/ or /mɒngljuːstəmə/ | pnir /niːr/ or /nɜː/ | ceiminadolt /siːzminədəʊlt/ or /saɪzminədəʊlt/ |
| Form B | bab /bæb/ | op /ɒp/ | ree /riː/ | raff /ræf/ |
| | dat /dæt/ | glack /glæk/ | hend /hend/ | weaf /wiːf/ or /wef/ |
| | chur /tʃɜː/ | tayed /teɪd/ or /teɪjəd/ | ful's /fʌlz/ | rejune /riːdʒuːn/ or /rədʒuːn/, |
| | weat /wiːt/ | sess /ses/ | depine /diːpaɪn/ or /dəpaɪn/ | wrault /rɔːlt/ |
| | throbe /θrəʊb/ | gouch /gaʊtʃ/ | brecked /brekt/ | darlanker /dɑːlankə/ |
| | cigbet /sɪgbet/ or /sɪgbɪt/ | mancingful /mansɪŋfəl/ | squow /skwəʊ/ | cyr /sər/ or /saɪr/ or /siːr/ |
| | quiles /kwaɪlz/ | untroikest /ʌntrɒɪkəst/ | pelnidlum /pelnɪdləm/ | byrcal /biːrkəl/ or /bərkəl/ |

**Appendix 7. Vocabulary memorisation test examples**

Part 1. Oral recall test

请说出以下单词的英文发音 (Please say the following words in English. The words appeared one by one on a computer screen).

太阳系仪，大纲，核糖，脑沟，老茧，但愿，修饰法，膨胀，暴怒的女人，使发红

Part 2. Written recall test

请给出以下单词的英文拼写 (Please spell the following words in English. The words were presented on a piece of paper).

暴怒的女人，老茧，使发红，太阳系仪，核糖，脑沟，但愿，大纲，修饰法，膨胀

Part 3. Aural recognition test

请听以下英文单词，并写出中文翻译 (Please listen to the following list of English words, and write down the Chinese translations. Each English word was pronounced three times).

zeugma, turgor, callus, maenad, orrery, ruddle, ribose, sulcus, precis, mayhap

Part 4. Written recognition test

请给出以下英文单词的中文翻译 (Please write down the Chinese translation of the following Englis words. The words were presented on a piece of paper).

maenad, orrery, zeugma, mayhap, sulcus, callus, turgor, precis, ribose, ruddle

## Appendix 8. L3-resembling errors made by participants in University C

When examining the decoding of participants in University C, it was found that most errors were similar to the ones that the participants in Universities A and B committed. However, as the intervention participants in University C were learning French while receiving English phonics instruction, some French-resembling pronunciation was observed in their decoding at t2. This is illustrated in Table 11.1.

*Table 11.1 L3-resembling mispronunciation of consonant graphemes by intervention participants in University C*

| Mispronounced grapheme | Target word | Error | Error rate at t1 | Error rate at t2 |
|---|---|---|---|---|
| <r> | *rejune* | /ʁ/ | 0% | 17% |
| | *raff* | /ʁ/ | 0% | 13% |
| | *ree* | /ʁ/ | 0% | 13% |
| <h> | *hend* | Silent (/end/) | 0% | 9% |

It can be seen that two errors were exclusively found in intervention participants in University C at t2, namely decoding <r> in the words *rejune*, *raff* and *ree* as its pronunciation in French, and omitting <h> in the word 'hend' which is in accordance with the French pronunciation rule. As these errors did not exist at t1 and were only observed in the decoding of intervention participants in University C, it is therefore conjectured that they were indeed L3-resembling errors. It is worth mentioning that not all the words containing graphemes <r> and <h> saw French-resembling errors. For instance, none of the participants decoded the grapheme <r> in the words *straced*, *throbe*, *brecked*, *cyr* and *byrcal* as /ʁ/; similarly, none of them omitted the grapheme <h> in the word *vunhip*. When examining the position of the mispronounced graphemes <r> and <h>, it was found that they were both at the beginning of a word.