# The Effects of Technique Feature Analysis on Retention of Form Recall in Written Production

Samanan Sudsa-ard

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Education

September 2023

The candidate confirms that the work submitted is her own work and that appropriate credit has been given where reference has been made to the work of others.

# Acknowledgements

My four-year Ph.D. journey which requires immense support has finally come to an end. This thesis could not have been completed without the following people.

First and foremost, I would like to express my sincerest gratitude to my supervisors, Dr. Richard Badger and Dr. Thi Ngoc Yen Dang, for their invaluable guidance, encouragement, insightful comments and great patience at every stage of the research project. Their immense knowledge and plentiful experience helped shape my academic growth and guided me through the difficult time of my study. I had received continued support and kind understanding from them during the COVID-19 pandemic although the situation brought numerous challenges and uncertainties to my Ph.D. journey. Throughout these four years of their supervision, I have gained more confidence in doing research and become a stronger researcher. Without their unwavering support and insightful feedbacks, the completion of my thesis would not be possible. I cannot envision a better supervisor than both of them.

I am also deeply grateful to both my internal and external examiners: Professor Alice Deignan, University of Leeds and Dr. Beatriz Gonzalez-Fernandez, University of Sheffield for their invaluable time giving insightful comments in the Viva Voce Examination. My thesis would not be well-improved and completed without their kind support and suggestions.

I would also like to thank Associate Professor Matt Homer for his invaluable time in giving me advices on Statistics. Learning new knowledge and skills on Statistics was extremely daunting. Before receiving his guidance, I always felt I had never read enough Statistics books and done good enough analysis.

Special appreciation goes to all the teachers, students and raters at Thammasat University in Thailand for their generosity in helping me during the Pilot Study and Main Study. I am especially indebted to Thammasat University, my workplace and scholarship sponsor for supporting my academic growth.

My deepest gratitude to my family and friends in Thailand who are always inspiring me to follow my dream. I would not have thought of starting my Ph.D. five years ago without one inspiring sentence from my father: "*You promised me that we would pursue a Ph.D. together, and now I nearly make it to the finish line*!" I am profoundly grateful for everything my family did in keeping me on track academically. Their love also keeps me motivated and warm even in winter. Their warmest words have encouraged me to continue writing up from the first winter in 2019 till finally completing the thesis in 2023.

I also thank all my friends in Leeds, the ones who always be by my side and be the laughter trying to make me smile or even laugh together with me during my rough journey. I would not have made it through my Ph.D. study without them.

Last but not least, I would like to thank the School of Education, University of Leeds for a generous bursary fund awarded to aid in producing my peer-reviewed journal article on my Ph.D. project.

I dedicate this work to them all for their contributions, encouragement, kind heart and belief in me during these times and always.

## Abstract

A lack of productive vocabulary knowledge tends to cause failures in communicative competence among EFL learners. Effective materials and vocabulary frameworks can help these learners to overcome the challenge. Technique Feature Analysis (TFA) framework has been used as a guidance for vocabulary material development. Yet, research into its effectiveness is limited and the impact of its five components (*Motivation, Noticing, Retrieval* and *Generative Use*) towards word learning is still unclear.

To fill the gap, this study examines the effectiveness of the TFA framework for two main purposes: 1) to investigate the predictive power of the TFA framework on form recall knowledge and 2) to explore the support of each TFA component (*Motivation, Noticing, Retrieval* and *Generative Use*) on retention (short-term retention and long-term retention) of form recall.

This is a quasi-experimental research design study with three repeated measurement. The constructed vocabulary test as a main research tool was implemented across three testing times (as Pre-test, Immediate Posttest, and Delayed Posttest). Quantitative data analysis through linear Mixed-Effects Models was employed to address the research questions.

The findings have shed light on the predictive power of the framework on form recall and give a clearer picture of potential factors affecting vocabulary learning. While *Motivation, Noticing, Retrieval* and *Generative Use* can lead to learning, they differ in degree of support on short-term and long-term retention of word form. Importantly, this thesis makes valuable contributions to vocabulary framework evaluation and validation. It captured the mismatch between TFA evaluated scores and test scores, suggesting that further improvement of the criteria within the framework as well as proper training for raters should be considered. Last but not least, it provides recommendations for further research into form recall knowledge and vocabulary materials development.

# List of Figures

## List of Tables

## List of Equations

# Table of Contents

# Chapter 1
# Introduction

## 1.1 Statement of problem

The ability to use English as a second language (L2) has become an additional skill that is highly required by employers in countries where English is not the first or official language. Consequently, English is widely taught as a compulsory subject in countries where the language is not their mother tongue (e.g., Thailand and Taiwan) to prepare L2 learners for the global job market. In English as a Foreign Language (EFL) context, these learners, however, have still encountered with difficulties in using English due to a lack of enough exposure to the target language both inside and outside of classroom (Benzies, 2013; Sawir, 2005). This insufficient opportunity to interact with the language appears to bring about failures in using the language to communicate effectively. While other factors such as English grammatical knowledge might affect EFL learners' proficiency in communicative skills (writing and/or speaking), vocabulary knowledge is one of the major problems. Studies with university students in various EFL contexts such as Thailand (Rattanadilok Na Phuket & Othman, 2015), Iran (Shokrpour & Fallahzadeh, 2007), Taiwan (Chen, 2002), Pakistan (Fareed et al., 2016), and Sudan (Alfaki, 2015) have revealed that a common error that obstructs effective L2 production of these students is related to the lack of sufficient vocabulary knowledge.

The same problem has been found in my own teaching context where students only have two-to-three hours to study English in classrooms each week. Some students expose themselves to the language through online channels or social media such as YouTube, Facebook, and Netflix outside of classroom. Yet, they have little opportunity to use English for communication in their daily life. When studying in class, it seems to me that students who communicate well in English are those who can remember and retain English vocabulary in their memory. Even though these students often make grammatical mistakes when expressing their ideas, they are more willing to use the language in class, compared to the rest of the students. Content knowledge and English grammar do not appear to be the major problems in using the language of these students because those who keep quiet when being asked to respond to questions or express their thoughts through speaking or writing in English are eager to share their ideas if they are allowed to use their first language (L1). This assumption is supported by Simon and Taverniers' s (2011) study on EFL university students' beliefs towards grammar, pronunciation and vocabulary, which revealed that due to the lack of confidence, these students search for the meaning of unknown words more often than looking up the grammatical structures or correct pronunciation in order to make communication successful. When discussing these problems with my colleagues, who have at least seven-year teaching experience, I realised that they found the same issue with students in their Foundation English courses and English for Specific Purposes (ESP) courses. So, it might be concluded that vocabulary deficiency is one of the main obstacles to effective communication (Shokrpour & Fallahzadeh, 2007; Rattanadilok Na Phuket & Othman, 2015; Fareed et al., 2016).

One solution to this problem is to provide learners with effective learning materials. According to Tomlinson (2012), learning resources such as coursebook, flashcards, and games can facilitate learning. With regards to EFL learning, knowledge gains require at least one-hour of exposure to English in classroom per week (Unsworth, et al., 2015). Ineffective design of vocabulary learning materials that does not provide adequate of learning time may lead to negative results on language learning. It is essential to assess and/or develop effective materials to make the most of time for students in language classrooms. For vocabulary learning, some scholars have claimed that vocabulary retention that involves ability to recognise and recall target learned words can be generally stimulated by well-designed tasks that induce active rehearsal and organisation (Postman & Rau, 1957; Bahrick, 1974 cited in Bahrick, 1979). For example, when learners are encouraged to recall target learned words by a well-organised vocabulary task/material with appropriate learning sequences in classrooms, their working memory are likely to be stimulated by the process of rehearsal which can lead to long-term retention. This means that a provision of opportunities for learners to repeatedly practice tasks or encounter with target words in an effective learning material would also bring benefits to learning. Many of my colleagues also pointed out the need for developing effective vocabulary learning materials and teaching productive skills in classrooms. It is because although they were aware of the importance of productive vocabulary knowledge for learners and the limitations of the English coursebooks used in our programmes, they tended to rely on the activities in these coursebooks due to time constraints in class and insufficient knowledge of how to make the activities better facilitate vocabulary development. So, it is important to develop a framework to evaluate the effectiveness of vocabulary learning tasks. The tasks should significantly support productive vocabulary learning in class that time is limited. This is because both receptive and productive word knowledge are suggested by earlier research (DeKeyser & Sokalski, 1996) to be sufficiently practised and properly learned. The support of proper materials validated by an effective framework may lead to the development of word knowledge to some extent. Such framework would help EFL learners to make the most of learning and for language teachers to manage classroom effectively.

In recognition of this need, several frameworks such as Involvement Load Hypothesis (ILH) (Laufer & Hulstijn, 2001) and Technique Feature Analysis (TFA) (Nation & Webb, 2011) have been developed. The current study, however, will pay attention to the TFA framework since it has emerged as a new framework to operationalise the construct of deep processing for L2 vocabulary learning and is suggested by previous research (e.g., Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018; Gohar et al, 2018; Zou & Xie, 2018) to be a more valid tool to evaluate vocabulary activities than the Involvement Load Hypothesis (ILH). By means of a checklist consisting of five main components: *Motivation, Noticing, Retrieval, Generative Use* and *Retention* based on memory research and theories, activities for vocabulary learning can be evaluated. However, previous studies (e.g., Hu and Nassaji, 2016; Nakata and Webb, 2016; Zou and Xie, 2018) seem to apply this framework to evaluate only receptive vocabulary knowledge. Little is known about the effects of each TFA component towards retention of vocabulary in written production. It is also unclear how far all of the five TFA components (*Motivation*, *Noticing*, *Retrieval*, *Generative Use* and *Retention*) including eighteen criteria as a checklist of evaluation need to be highly rated in vocabulary materials to facilitate word retention.Therefore, this study will investigate the effectiveness of TFA framework in evaluating vocabulary learning materials by emphasising the different components which are believed to promote word retention in written production.

## 1.2 Research aim

The present study evaluates the  Technique Feature Analysis (TFA) framework on four language learning components (*Motivation, Noticing, Retrieval* and *Generative Use*) as a way of evaluating vocabulary learning materials with a focus on written production of single-word form. The research aim has been identified primarily from the need for developing effective vocabulary learning materials for EFL learners. This has led me to seek for a useful vocabulary framework to develop learning materials that can enhance productive vocabulary knowledge. Based on the evidence found from previous empirical research (e.g., Hu and Nassaji, 2016; Chaharlang and Farvardin, 2018; Gohar et al, 2018; Zou and Xie, 201) the TFA framework has been recognised as an effective vocabulary tool for measuring predictive power of vocabulary materials (see details in Chapter 2). Yet, its five components (*Motivation, Noticing, Retrieval, Generative Use* and *Retention*) have not been explored whether they can support learning and retention of form recall to the same extent. So, this study pays attention to the overall effectiveness of the framework and the extent to which its components, as influential learning factors, can potentially encourage short-term and long-term retention of form recall of EFL learners.

There are two main research aims. The first aim is to explore the predictive power of the TFA framework in evaluating form recall knowledge in written production. This aim is important because research related to the effects of TFA on vocabulary activities focusing on form recall is still scarce. Previous research on TFA tended to rely mostly on word recognition or receptive word knowledge.  The second aim is to investigate degrees of effectiveness of each TFA components. However,  it should be noted that *Retention* component in the TFA framework was not included as a learning factor for the investigation (see Section 2.3.3.5 for rationales), and the term ‘retention’ used in the current study refers to the participants’ ability to retain productive word knowledge in controlled written production. The second aim is essential because prior research did not compare the effects of four TFA components (*Motivation, Noticing, Retrieval* and *Generative Use*) on short-term and long-term retention of form recall while they may give different degrees of support on vocabulary learning. This study can provide an insight into the effects of TFA components which helps the development of effective productive vocabulary materials and the validation of the framework. Section 1.3 below provides theoretical, methodological and pedagogical contributions of the current study based on the two main research aims.

## 1.3 Research contribution

For theoretical and empirical contributions, the present study would provide further insight into the value of TFA to evaluate and discern the potential of different vocabulary activities with different TFA support on *Motivation, Noticing, Retrieval* and *Generative Use* for developing knowledge and retention of productive vocabulary. This is meaningfully given that studies validating the TFA are limited in number and no studies have investigated the validity of the TFA as a framework to evaluate the activities for encouraging motivation, noticing, retrieval, and generative use in written production. This study may also contribute to the investigation of the most effective components for retaining written words in ESL/EFL learners' memory. This is important because memory retention tends to enhance learning (Vanichvasin, 2021). Unlike this study, earlier research related to TFA relied heavily on self-evaluation which might cause bias. Also, previous research (i.e., Kamali et al., 2020) on word recall tend to employ only strict scoring system to assess word knowledge. Avoiding self-evaluation and using both scoring scheme: strict and sensitive would provide more precise findings that can benefit other studies in the field of vocabulary assessment. These could bring about methodological and pedagogical contributions. The investigation may also unveil more constraints on word retention and language learning and performing as well as effective ways for vocabulary instruction. In terms of pedagogy, the findings can shed some light on the effective ways that support word gain and retention in classroom learning and how language teachers make use of the framework to develop materials that support recall of vocabulary.

## 1.4 Organisation of the chapter

The present study covers seven chapters: 1) Introduction, 2) Literature Review, 3) Research Methodology and Pilot Study, 4) Main Study, 5) Main Study Findings, 6) Discussion, and 7) Conclusion. Chapter 1 provides research background and highlights the statement of problems as well as the research aim of this study. Chapter 2 gives information regarding word knowledge and the components (*Motivation, Noticing, Retrieval, and Generative Use*) that may facilitate learning and retention. It also presents related frameworks (ILH and TFA) that can be used to evaluate vocabulary learning materials. In Chapter 3, rationales for the research questions are outlined before describing the research design, context of the study, participants, the data collection instruments and the analysis of the Pilot Study. After that, Chapter 4 and Chapter 5 report the data analysis and the findings of the Main Study, respectively. Finally, I discuss the results from the findings in Chapter 6 and give a conclusion of the present study in and Chapter 7.

# Chapter 2
# Literature Review

To investigate word gains and retention, information related to word knowledge and word learning should be reviewed. In this chapter, details about word knowledge was provided in Section 2.1 to understand what it means to know a word. The current study had a specific aim on the investigation of word knowledge gained from written production. Due to this, I also sought for relevant information on types of word knowledge in this section. Then, I explain in Section 2.2 the need of EFL learners on vocabulary development and how teachers can facilitate them to gain word knowledge. In Section 2.3, I explored vocabulary frameworks that could promote the construct of vocabulary materials as well as the gains of word knowledge. Related vocabulary learning frameworks: ILH (Hulstijn & Laufer, 2001) and TFA (Nation & Webb, 2011) that are believed to have potential to improve vocabulary materials for gaining word knowledge and retention are reviewed in this section. Before moving onto the methodology section, I presented the scope of the current study and research questions that this study aims to explore in Section 2.4 and Section 2.5, respectively.

## 2.1 What is involved in knowing a word?

Word knowledge can be generally defined as the knowledge of word elements: form and meaning as well as word associations used in both spoken and written language. However, the fundamental conceptions of vocabulary knowledge construction are still unclear (González-Fernández & Schmitt, 2020). There is no single well-accepted definition of word knowledge. Due to this, researchers have endeavored to better understand what involves in knowing a word because word knowledge is central to communicative competence, which is the main aim of acquiring/learning a new second or foreign language.

A solid vocabulary knowledge facilitates all areas of communication, namely reading, listening, speaking and writing. It is important for EFL learners for several reasons. First, knowledge of vocabulary significantly corelated to reading and listening comprehension (Nation, 2001; Zhang & Zhang, 2020).

The development of receptive vocabulary knowledge improves lexical coverage rate required for unassisted reading and listening comprehension (Hu & Nation, 2000; Nation, 2006), so that it can lead to achievement in language learning. Second, previous studies (Shokrpour & Fallahzadeh, 2007; Alfaki, 2015; Rattanadilok Na Phuket & Othman, 2015; Chen, 2002; Fareed et al., 2016) found that insufficient productive vocabulary knowledge obstructs effective language production. A robust productive vocabulary knowledge would improve speaking and writing abilities of EFL learners. Last but not least, adequate knowledge of vocabulary enables language use in various contexts, resulting in effective communication. Given the importance of word knowledge, considerable attempts have been made to operationalise what it means as knowing a word. Richards (1976) is probably the one who made the first attempt.

In Richards' s framework, eight assumptions were proposed as components of vocabulary knowledge as shown below:

> ***ASSUMPTION 1****: "The native speaker of a language continues to expand his vocabulary in adulthood, whereas there is comparatively little development of syntax in adult life."*
>
> ***ASSUMPTION 2****: "Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words we also "know" the sort of words most likely to be found associated with the word."*
>
> ***ASSUMPTION 3****: "Knowing a word implies knowing the limitations imposed on the use of the word according to variation of function and situation."*
>
> ***ASSUMPTION 4****: "Knowing a word means knowing the syntactic behavior associated with that word."*
>
> ***ASSUMPTION 5****: "Knowing a word entails knowledge of the underlying form of a word and the derivations that can be made from it."*
>
> ***ASSUMPTION 6****: "Knowing a word entails knowledge of the network of associations between that word and other words in language."*
>
> ***ASSUMPTION 7****: "Knowing a word means knowing the semantic value of a word."*

> ***ASSUMPTION 8**: "Knowing a word means knowing many of the different meanings associated with the word."*
>
> (Richards, 1976, pp. 78 - 82)

These assumptions demonstrate that word knowledge tends to have a complex structure associated with several components such as word frequency, word association and semantic structure. Thus, Richard' s framework has been credited as the first detailed conceptualisation of word knowledge and raised an awareness of scholars (e.g., Nation, 1990; 2001; 2013; Milton & Fitzpatrick, 2014) to conduct more research on vocabulary. For example, the second assumption explains that the most frequent words appeared in spoken or written language are likely to be more familiar to the language learners/users than rarely encountered words. This concept has influenced to the increase of research on word frequency (e.g., Nation, 1997; Coxhead, 2000), leading to the development of the up-to-date tools containing word frequency data for English such as VocabProfilers (www.lextutor.ca). Moreover, some assumptions which explained that knowing a word means knowing the word form with its derivation as well as semantic value, or core meaning of a word with other possible meanings connected to the word (Richards, 1976, pp. 80-82) has been well-accepted and developed further by later studies (e.g., Nation, 1990; 2013). However, Richards's framework has been criticized for inappropriate ordering of the list of assumptions by some scholars (Meara, 1996c; Milton & Fitzpatrick, 2014). Meara (1996c) argued that the list of the assumptions is not appropriate because the seventh assumption concerning word meaning should have been considered as a primary feature of word knowledge and explained in the top of the list. Despite this, subsequence studies (i.e., Nation, 1990, 2001, 2013; Meara, 1996; Milton, 2009; Nation & Webb, 2011) have expanded Richard's framework in various views. This might be because the criticisms did not mainly relate to the core concepts of word knowledge.

Richards' s assumptions were elaborated more clearly later by Nation's taxonomy as presented in Table 2.1 below (1990; 2001; 2013). To know a word, Nation (2013) claims that learners need the knowledge of word form, word meaning, and how to use the word correctly in various contexts. While word knowledge is a complex concept for which no simple all-inclusive description has been given, a division between productive and receptive knowledge is widely accepted and even included as separate aspects under each component of Nation’s (1990; 2001; 2013) taxonomy. Having receptive word knowledge could mean being able to recognise a word from listening or reading and to retrieve appropriate meaning of the word form. Productive knowledge, on the other hand, could refer to the ability to produce word forms through speaking or writing (Nation, 2013).

***Table 2.1. Components of word knowledge (Nation, 2001, p.27; 2013, p. 49)***

| | | | |
|---|---|---|---|
| Form | spoken | R | What does the word sound like? |
| | | P | How is the word pronounced? |
| | written | R | What does the word look like?` |
| | | P | How is the word written and spelled? |
| | word parts | R | What parts are recognisable in this word? |
| | | P | What word parts are needed to express the meaning? |
| Meaning | form and meaning | R | What meaning does this word form signal? |
| | | P | What word form can be used to express this meaning? |
| | concept and referents | R | What is included in the concept? |
| | | P | What item can the concept refer to? |
| | associations | R | What other words does this make us think of? |
| | | P | What other words could we use instead of this one? |
| Use | grammatical functions | R | In what patterns does the word occur? |
| | | P | In what patterns must we use this word? |
| | collocations | R | What words or types of words occur with this one? |
| | | P | What words or types of words must we use with this one? |
| | constraints on use (register, frequency…) | R | Where, when, and how often would we expect to meet this word? |

The relationship between receptive and productive areas of knowledge has been categorised differently into two broad views: dimensional conceptualisation and developmental conceptualisation (Schmitt, 2010; Milton & Fitzpatrick, 2014). Both the dimensional approach and developmental approach put an emphasis on aspects of knowing a word as well as receptive and productive word knowledge (Milton & Fitzpatrick, 2014). However, the dimensional approach views receptive and productive knowledge of a word as separate sets of knowledge while the developmental approach tends to believe that one aspect of word knowledge leads to the development of other aspects (Milton & Fitzpatrick, 2014), and degree of mastery is developed primarily from partial knowledge (Corson, 1995; Schmitt, 2000; Pigada & Schmitt, 2006).

For the dimensional approach, Meara (1996) views receptive word knowledge and productive word knowledge as a separate part of lexical knowledge. However, these different notions remain controversial. Nation (2001; 2013) elaborates the terms receptive and productive knowledge by identifying three dimensions of knowing a word: "form" (knowing how the word is pronounced or written and recognising its word parts), "meaning" (knowing how the word expresses or refers to a particular concept and its associations to other words) and "use" (knowing how the word can be used in different grammatical patterns with various collocations, and knowing in which context the word is used). Each dimension includes both receptive and productive knowledge (see Table 2.1). It should be noted that in the field of vocabulary the term 'form' is defined differently from that of grammar. When we identify a word form we pay attention to how the word is written or spelled. So, 'form' can be categorised into three aspects: spoken form, written forms, and word parts (Nation, 1990; 2001; 2013). Word knowledge of spoken form includes not only receptive knowledge of word sound, but also productive knowledge of pronunciation of a word. Likewise, word knowledge of written form involves both receptive and productive knowledge. If we have receptive knowledge of a written word, we need to be able to answer the question *what does it look like?* (Nation, 1990, p. 31) to ensure that we can recognise its written form. Moreover, we need to be able to spell the word correctly to claim that we have adequate productive knowledge of form of that word. Thus, in this view receptive and productive word knowledge are seen as separate dimensions. Both receptive and productive knowledge of each component (form, meaning and use) are unlikely to develop together at the same time and pace. Laufer and Goldstein's (2004) and Laufer, et al. (2004) studies also support this view in that word knowledge consists of four separate dimensions: passive (meaning) recognition, active (form) recognition, passive (meaning) recall and active (form) recall. Thus, they suggested that to know a word, it is essential to be able to recognise and recall both meaning and form of that word. Also, knowledge on form and meaning may not be constantly developed to the same extent.

On the contrary, some scholars (e.g., Henriksen, 1999; Milton, 2009) who rely on the developmental approach believe that receptive vocabulary knowledge is a foundation for productive vocabulary knowledge. Henriksen (1999) claims that word knowledge is developed along a continuum from three dimensions: partial-to-full comprehension, depth, and receptive-to-productive. That is, if the partial-to-precise (or breadth) dimension is considered, learners need to comprehend the meaning of a word and be able to write its form correctly to be regarded as having full knowledge of that word. It is also believed that meaning and form recognition need only partial knowledge. To gain full knowledge, learners are required to recall form and meaning of the newly learned word correctly. The second dimension concerns the depth of knowledge of word associations, meaning that leaners need to know different meanings of a word and its associations such as other related words. Lastly, the receptive-to-productive dimension puts an emphasis on the comprehension (receptive) and use (productive) of a word. These three dimensions are common in that word knowledge is a movement along a continuum in which partial and receptive knowledge is developed first, followed by full and productive knowledge.

According to this, receptive and productive word knowledge, which is the most well-known classification of word knowledge proposed by Palmer (1921) and elaborated by Nation (1990; 2001; 2013), can be seen as both a dichotomy (Meara, 1996) and a continuum (Milton, 2009) from different perspectives. Although some recent studies (e.g., Gonzalez-Fernández & Schmitt, 2020; Laufer & Goldstein 2004; Schmitt, 2019) have tried to find more explanation, a common agreement has not yet been made whether they are a dichotomous or continuing process (AbManan et al., 2017) since the process in which receptive moves to productive has still not been made clear.

In the current study, I focused specifically on aspects under two main dimensions presented in Nation's (2013) components of word knowledge: 'form' and 'meaning', following the dimensional approach. However, the starting point was the ‘form’ as this is what is produced in written language. ‘Meaning’ was investigated in this study for the purpose of data interpretation to check when the participants can write the correct form of the word. Assessing the meaning of the correct form should ensure that the participants have knowledge of the newly learned words. According to Schmitt (2010), meaning recall is more central to investigations of receptive vocabulary knowledge while form is more important in studies of productive knowledge. Concerning the methodological purposes, the measurement of meaning recall could help to ensure the results when test takers spell a word correctly even though this study paid attention to productive knowledge of form recall. This means that meaning recall results will only help to check the partial knowledge (sensitive scoring) results of form recall, but they will not be included in the analysis of the main study.

Furthermore, types of productive knowledge other than ‘form’ needed to be reviewed in this section as I had an intention to explore written production. According to Laufer (1998), and Laufer and Paribakht (1998), productive knowledge can be categorised into two main types: controlled productive knowledge and free productive knowledge. Controlled productive vocabulary knowledge could mean the ability to recall word forms (spellings) with the help of cues such as semantic cues (L1 translation) and graphic cues (pictures) in written production. Contrarily, if a writing task is freely produced without a provision of cues, it is to assess free recall or free productive knowledge, which is the ability to retrieve words from memory without any help. In the current study, I put an emphasis on controlled productive vocabulary knowledge as I had an intention to measure the target learning words in a reading passage of the material used in a Foundation English course. Free productive knowledge might not be suitable to measure retention of the target learning words for three major reasons. First, it involves grammatical aspects that might affect the results of this study. Findings from previous studies (e.g., Phonna, 2014; Fareed et al. 2016) affirmed that grammatical mistakes tend to be common errors in free writing of L2 learners in higher education. These learners have been dealing with problems when using grammar in writing. If the students use the target learning words incorrectly in sentences, it might be because they have inadequate knowledge of English grammar. Another reason is that free writing is less controlled. As this type of writing tends to allow the students to freely use any words that are not in the target learning list, it seems difficult to interpret the result whether they do not know the words or they are just familiar with the other words they have already known and chosen to use in their writing if the students in this study do not use some of the target learning words in a free writing task. Lastly, free writing seems to require several skills and academic conventions such as essay structure. According to this, I might not be able to measure knowledge of the target learning words if they do not use it in a controlled written task.

To conclude, controlled productive knowledge was the main focus because free writing tends to contain grammatical elements which may result in misleading interpretation, so it is beyond the scope of the present study. This was not an investigation into 'use' because the component seems to require more knowledge such as grammars and word associations than knowledge on spelling (see Table 2.1: Nation, 2001; 2013). While free productive knowledge involves grammatical functions related to use, controlled productive knowledge involves mainly knowledge of word 'forms' (where 'forms' refer to lexical items). Although it is likely that the students may look at grammatical structure to help them guess the correct form of the testing words, this does not affect the results of the present study since other derived or inflected forms of the target words were not be tested, and the letters of each word are controlled by numbers of underline for each blank (see Figure 3.9). Because previous research on retention measured written production through recall, the focus was mainly on recall rather than recognition. For instance, Myers's (1914) study on vocabulary retention used a recall test by requiring the participants to spell the target words they had encountered half an hour before taking the test without being informed earlier. Moreover, both well-accepted productive tests such as the Productive Vocabulary Levels Test, or PVLT (Laufer & Nation, 1990; 1999) and Computer Adaptive Test of Size and Strength, or CATSS (Aviad-Levitzky et al., 2019) tend to measure controlled productive knowledge through meaning and form (spellings) recall. Since the PVLT have been widely employed to elicit productive knowledge by researchers (e.g., Alonso & Garcia, 2014; Gibriel, 2017; Kiliç, 2019), the recall test formats used in this test should be appropriate for controlled productive measurement (see more detail in Section 3.5.3.3). Several researchers also maintain that recall results in better vocabulary learning (Karpicke & Roediger, 2008) and ability to recall a word seems to be the most challenging one, especially in long-term memory (Laufer & Goldstein, 2004; Gonzalez & Schmitt, 2020). Thus, it might be necessary to explore recall knowledge and productive vocabulary from form (spelling) recall of the target learning words in order to measure controlled productive knowledge. In addition to form recall, I only pay attention to meanings that link to the target word forms for the purpose of data checking to ensure the form recall results of the main study as mentioned earlier in this section.

## 2.2 EFL learners' vocabulary knowledge

The prevision section has shown that word knowledge entails both receptive and productive word knowledge (Nation, 2001; 2013, Milton, 2009). Both aspects tend to be required for the improvement of EFL learners' vocabulary knowledge as suggested by the list of what it means to know a word from Nation's (2001; 2013) taxonomy shown in Table 2.1. Given the complex nature of vocabulary knowledge, it is important to investigate EFL learners' receptive and productive vocabulary. Finding of such investigation would enable language researchers and teachers to identify learners' lexical gaps and provide relevant support to help them narrow these gaps. Therefore, numerous studies have been conducted to measure vocabulary knowledge of EFL learners. Most of these studies have focused on receptive vocabulary knowledge in various ways (e.g., Nguyen & Webb, 2016; Özönder, 2016; Mungkonwong & Wudthayagorn, 2017; Snoder & Laufer, 2022). For instance, Nguyen and Webb's (2016) study focused mainly on measuring Vietnamese EFL learners' receptive knowledge of word collocations: verb-noun and adjective-noun pairs at 1,000, 2,000 and 3,000 word frequency levels. The COCA corpus (Davies, 2008) was employed to select collocates for each frequency level for constructing the collocation test. They also used a new version of the Vocabulary Levels Test, or VLT (Webb & Sasao, 2013; Webb, et al., 2016) which can measure single-word knowledge receptively to compare with the results of the designed collocation test. The VLT results showed that the participants reached 93.57% word knowledge of K1 level, but tended to have limited word knowledge at K2 and K3 levels, showing 77.83% and 65.03%, respectively. Also, they still had insufficient receptive knowledge of both types of collocations measured by their collocation test. More studies on receptive knowledge were conducted widely in other EFL contexts by using various research instruments to match their research aim. Some of them employed the VLT, similar to that of Nguyen and Webb' s (2016) study while the others used the Vocabulary Size Test (VST) by Nation and Beglar (2007). In Turkey, an investigation into receptive vocabulary size through the Vocabulary Levels Test, or VLT (Schmitt, et al., 2001) was investigated by Özönder. The researcher employed the test to measure receptive word knowledge of 104 EFL undergraduate students to compare with their GPAs. While the study reported the irrelevant relationship between the two measures (VLT and GPA), it highlights that these EFL learners have large and adequate receptive vocabulary size. Another research project (Mungkonwong & Wudthayagorn, 2017) conducted in a Thai context adapted Nation and Beglar's (2007) Vocabulary Size Test (VST) to measure vocabulary size of Thai freshmen (n = 484) in both public and private universities across Thailand. They also focused on only receptive word knowledge. This study found that Thai EFL learners have larger vocabulary size, which is about 4,200 word families, than the size required by the core curriculum (approximately 3,600 word families). This figure is also higher than the vocabulary size (at least 3,000 high-frequency words) to deal with tasks in university level as suggested by Nation (1990). A recent study (Snoder & Laufer, 2022) from Sweden, an English-input rich country, also explored only receptive word knowledge of Swedish EFL learners (n = 88) through the use of the VST 14K version. They found that EFL learners at intermediate (9th grade of compulsory school) and advanced (12th grade of upper secondary school) proficiency levels tended to have sufficient receptive knowledge of the target basewords and their derived forms. This might not be surprising since these EFL learners have a Germanic L1 closely related to English. Yet, the study reported similar findings to prior studies conducted in other EFL contexts such as Thailand and Turkey where the opportunity of exposure to English is limited mostly in classroom. It can be inferred from the earlier results that receptive knowledge of these learners does not seem to be the major concern.

A smaller number of studies have measured both receptive and productive vocabulary. Laufer and Goldstein’s (2004) study, for example, paid attention not only to receptive, but also productive word knowledge. The findings from a large group of learners (n = 435) showed that productive knowledge is superior than receptive knowledge in terms of difficulty. They claim that:

> *“if active knowledge is more difficult to achieve than passive knowledge, and if recall is more difficult than recognition, then the most advanced degree of knowledge is reflected in active recall and the least advanced knowledge is passive recognition” (2004, p. 408).*

Later in 2008, Webb’s study explored the relationship between receptive and productive word knowledge by measuring vocabulary size of EFL Japanese learners (n = 83) both receptively and productively. He employed receptive and productive translation tests to evaluate vocabulary size of these learners. Each participant has to write definition of ninety target words in their L1 for receptive test. In the productive test, they had to write L2 form of the target words given L1 meanings as a clue. Despite of different EFL contexts and research instruments, the result was similar to Laufer and Goldstein’s (2004) study in that EFL learners’ receptive word knowledge size was larger than their productive word knowledge. Webb also discovered that the receptive-productive ratio was related to word frequency (band). When the word frequency increased from word band 3 (3,401st to 6,600th most frequent word) to word band 2 (1,901st to 3,400th), this ratio also raised from 65% to 73%, respectively.

In addition to Laufer and Goldstein’s (2004) and Webb (2008) findings, an unpublished doctoral research of Utsajit (2022) showed similar results between receptive and productive knowledge. A part of her study examined the relationship between receptive and productive vocabulary size of Thai EFL advanced learners through the use of VKS receptive test, controlled productive test, and free productive test. It was found that receptive vocabulary size is larger than controlled productive vocabulary, followed by free productive vocabulary.

Both Mungkonwong and Wudthayagorn (2017) and Utsajit’s (2022) studies in Thai EFL context aligns with another study (Boonyarattanasoontorn, 2017) of

Thai learners' perceptions on writing problems in that receptive vocabulary knowledge was not seen as a major problem in writing. The students perceived themselves as having adequate vocabulary for writing tasks as they believed that they had enough receptive word knowledge in order to complete the tasks.

Taken together, it can be seen that the findings of previous studies are similar despite of the variation in EFL contexts and research designs. The findings from different EFL contexts (i.e., Turkey, Thailand, and Sweden) indicated similar positive results of receptive knowledge, meaning that EFL learners tend to have sufficient receptive vocabulary knowledge (Özönder, 2016; Mungkonwong & Wudthayagorn, 2017; Snoder & Laufer, 2022). Also, previous studies (e.g., Laufer & Goldstein, 2004; Utsajit, 2022) on both receptive and productive knowledge revealed one common pattern in that learners' productive vocabulary was always smaller than their receptive knowledge. It suggested that productive vocabulary knowledge is the aspect that EFL learners needs support from language teachers. This tends to support the argument of former researchers (e.g., Waring, 1997a; Webb, 2005) in that receptive is easier to acquire than productive knowledge. Based on these arguments and empirical findings, the need of EFL learners to improve productive vocabulary can be captured. EFL learners' receptive knowledge is likely to meet the level of satisfaction while their productive knowledge should be improved.

To gain EFL learners' productive vocabulary knowledge, vocabulary learning activities should be designed effectively and organized in a principled way so that the learning time is well spent. For more than a decade, several frameworks have been proposed to help teachers to design and evaluate vocabulary learning strategies. Section 2.3 below provides information related to two vocabulary framework (The Involvement Load Hypothesis and the Technique Feature Analysis) that have been well-accepted in the field of vocabulary to facilitate language learning. Theories influencing to the construct of and critiques on these frameworks are also discussed.

## 2.3 Frameworks to evaluate vocabulary learning activities

The Involvement Load Hypothesis (ILH) (Laufer & Hulstijn, 2001) and the Technique Feature Analysis (TFA) (Nation & Webb, 2011) frameworks have been developed as checklists for designing effective vocabulary tasks. Both frameworks are based on Craik and Lockhart' (1972) memory model – Depth of Processing Level. According to Craik and Lockhart, there are two processes involving in short-term and long-term memory: *shallow processing* and *deep processing*. While shallow processing results in short-term memories that can decay easily, long-term memory stems from deeply and meaningful processing of information. They stated that:

> *later stages of memory…[are called] "depth of processing" where greater "depth" implies a greater degree of semantic or cognitive analysis. After the stimulus has been recognised, it may undergo further processing by enrichment or elaboration. (1972, p. 675).*

Besides, they asserted that elaborate processing, a part of deep processing, links with meaningful semantic associations. These associations lead to long-lasting memories. For language learning, tasks that requires deeper processing can result in learners' higher performance (Cermak & Craik, 2014). The Depth of Processing Level have supported with evidence from empirical studies (e.g., Craik & Tulving, 1975; Laufer & Hulstijn, 2001; Hu & Nassaji, 2016; Long & Kahana, 2017). Therefore, subsequent vocabulary learning studies have put an emphasis on pedagogical tasks that require deep processing and involvement of learners performing the tasks (e.g., Laufer & Hulstijn, 2001; Jarhangiri & Abilipour, 2014; Soleimana & Ramanian, 2014; Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018). This is because deep processing is believed to be associated with word recognition and recall that can enhance retention of learners. Importantly, two frameworks which based on this hypothesis, the ILH and TFA, have been proposed to evaluate the effectiveness of vocabulary learning activities. These frameworks are explained in detail in Section 2.3.1 and Section 2.3.2, respectively.

### 2.3.1 Involvement Load Hypothesis (ILH)

The construct of Involvement Load Hypothesis (ILH) lies in the line of second language acquisition (also known as SLA) and information processing which is related to cognitive processes. The term was coined by Laufer and Hulstijn (2001) to emphasise the importance of depth of processing that has been debated by cognitive phycologists for over a decade as it is believed to stimulate better memory performance in terms of retention (Anderson, 1995; Baddeley, 1997). According to Laufer and Hulstijn (2001), the improvement of memory and vocabulary learning are heavily associated with cognitive and motivational dimensions. They were inspired by Craik and Lockhart's (1972) in-depth memory processing theory which believes that better learning and understanding is likely to result from deeper memory processing. Due to this, Laufer and Hulstijn attempted to construct the task-induced involvement framework in order to stimulate these two dimensions by introducing three domains consisting of *need* (n), *search* (s), and *evaluation* (e) for designing tasks in vocabulary learning. While the *need* component concerns motivation in understanding and using the target words, *search* and *evaluation* put an emphasis on memory retrieval and generative use of the words, respectively. There are three levels of evaluation for each domain: (a) none of an involvement factor, indicated by a minus (-), (b) a moderate involvement, marked by a plus (+), and (c) a strong involvement, signified by a double plus (++). Each level determines different degrees of achievement. All of these proposed aspects are driven from motivational and cognitive involvement views of psychology and SLA studies. As the distinction between moderate and strong is sometimes ambiguous, details regarding the three involvement domains, together with examples of moderate and strong involvement levels given to different tasks is explained below.

In terms of *need*, this component associates solely with motivational aspect, and is 'concerned with the need to achieve' (Laufer & Hulstijn 2001, p. 14). Laufer and Hulstijn (2001) believe that not only can information processing devices drive successful language learning, but motivation also plays a crucial role in achieving language learning goals in a boarder sense. That means if learners have a strong motivation or need to achieve a goal, they can perform better than those who suffer from a lack of motivation. Degrees of *need* involvement in performing a task can be assessed by applying the level of involvement criteria. For example, a reading with glosses activity could be rated as weak in terms of *need* (-) because there is a lack of learners' effort and need for learning or exploring new words. However, reading with a dictionary activity could be rated as moderate in terms of *need* (+) because learners are motivated by external factors (e.g., teacher's encouragement in this case). It cannot be rated as strong because although the unknown words in the reading passage are needed to learn, it is often that the activity is not intrinsically motivated by learners themselves. That means the same activity—reading with using a dictionary could be stronger in terms of *need* (++) if learners want to look up for the meaning of unknown words in a dictionary by themselves. Strong *need* (++) involves learner's intention or intrinsic motivation that drives to higher degree of motivation in achieving a word learning goal which will result in better word retention.

As for *search*, this domain is derived from cognitive views of vocabulary learning and retention. *Search* is strong (++) when learners have their own attempt to look up for both meanings and forms of the target learning words by the use of a dictionary or asking for help from the instructor. This is because searching for meanings and forms can trigger receptive retrieval and productive retrieval, respectively in memory process (Nation & Webb, 2011; Hu & Nassaji, 2016). However, it is moderate (+) when learners are assigned by the instructor to consult a dictionary in order to find only meanings, not form of the words they learn. *Search* is weak (-) when the meanings are provided for the unknown words.

In terms of *evaluation*, it is another cognition domain concerned with a need to assess words in specific contexts. It puts an emphasis on an association between forms and meanings. For example, one word might have various meanings which can be used differently depending on contexts. That means if the task such as wordcards plus sentence writing stimulates learners to notice and assess whether a word fits with its context and allows them to determine appropriate word choice by themselves, the level of involvement for *evaluation* is high, or strong (++). However, *evaluation* is rated as moderate (+) when different choices of words are provided for learners to choose (Laufer & Hulstijn, 2001). If meanings are provided in margins, there is no need for assessing the words. Reading with glossary, for example, can be rated as low (-) for *evaluation.*

The validity of the ILH as a framework to evaluate vocabulary learning activities has been confirmed by findings of numerous studies (e.g., Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001; Keating, 2008; Kim, 2008; Zou, 2017; Huang, 2018). Yet, several researchers (e.g., Laufer & Hulstijn, 2001; Nation & Webb, 2011; Zou, 2017; Huang, 2018) have pointed out several limitations of this framework. First, the mismatch between the ILH scores and test results (Folse, 2006; Zou, 2017; Huang, 2018). According to Huang's study (2018), for instance, ILH scores did not correlate with delayed posttest results, and tasks with higher power or higher ILH scores failed to promote long-term vocabulary retention. This might be because ILH focuses more on the final score than the value of each component, so that factors affecting to the effectiveness of each component on vocabulary learning gain might not be able to identify accurately. As the sum of the pluses indicates the efficacy of ILH for vocabulary learning, tasks with the same degree of involvement load such as moderate *need* (+), no *search* (-), moderate *evaluation* (+) (ILH score of 2) and moderate *need* (+), no *search* (-), no *evaluation* (-) (ILH score of 2) show the same involvement load result, which is hard to identify the most influential factor in supporting vocabulary learning and the most effective word learning tasks when they are compared.

Second, the value of each component might not contribute to the same degree of efficacy. Laufer and Hulstijn (2001) stated that two components, namely *need* and *evaluation* might have higher weight in predictive vocabulary gain than that of *search*. However, *search* has the same evaluation criteria as the other two (*need* and *evaluation*) components.

Third, the evaluation criteria of the ILH tend to be limited. Nation and Webb (2011) pointed out three concerning issues of the ILH which include 1.) small scales of ILH domains, 2.) restricted number of response categories, and 3.) absence of descriptions or 'anchoring labels' of each category (Nation and Webb, 2011, p. 5). They argued that these issues could lead to discrimination and low reliability in activity evaluation as supported by the evidence from the studies of Folse (2006), Zou (2017) and Huang (2018).

Several attempts have been made to refined the ILH or to replace it with another framework. In the next sections, I will review two frameworks that expand on the ILH (the refined Involvement Load Hypothesis and the Involvement Load Hypothesis Plus) and one new framework (Technique Features Analysis).

### 2.3.2 Refined Involvement Load Hypothesis and Involvement Load Hypothesis Plus

New studies (e.g., Hazrat, 2020; Yanaginawa & Webb, 2021; 2022; Hazrat and Read, 2022) has also reported some weaknesses of the original ILH as similar to the limitations mentioned in the previous section. Hazrat (2020) conducted a P.H.D. study to explore the impacts of Involvement Load Hypothesis on word learning. Following the ILH assumption, her hypothesis was that a higher involvement load task better develop word knowledge than a lower induced vocabulary task. She included ten groups (n = 168) of intermediate level learners (one control and nine experimental groups). Before the experiment, she distributed a pre-test including productive tests, translation tests and matching tests to assess form recall, meaning recall and form recognition, respectively to all groups. Then, different vocabulary activities with the same moderate involvement load for *need* (+), but various involvement load for *search* and *evaluation* were assigned to the experimental groups. Need was rated to be moderate (+) as similar to most previous studies on ILH (e.g., Keating, 2008; Kim, 2008). Hazrat divided eight experimental groups into four main categories for *evaluation*: receptive (meaning) retrieval (+) for Group 1 ang Group 3, productive (form) retrieval (++) for Group 2 and Group 4, sentence writing (++) for Group 5 and Group 7 and composition writing (++) for Group 6 ang Group 8. Among these eight groups, Group 1, Group 2, Group 5 and Group 6 were rated as weak for *search* (-) while the other groups involves moderate *search* (+). However, another experimental group (Group 9) was assigned with an activity that involves *search*, but was rated as weak for *evaluation* (-). After the learning, two tests: immediate and delayed posttest were used to measure form recall, meaning recall and form recognition again. She employed the Mann-Whitney U test to analyse the data. According to the results of multiple comparisons, the results confirmed the ILH predictions in that task with high involvement load led to better word learning and retention. However, sentence writing and writing composition activities showed significantly higher effects on word gains than the productive retrieval activities even though they had the same high involvement load for *evaluation* (++). Thus, she suggested revisiting the framework to improve its predictive power or hypothesis assumptions which is in line with the arguments made by previous researchers (e.g., Laufer&Hulstijn, 2001; Nation & Webb, 2011; Zou, 2017; Huang, 2018).

These issues had influenced to the development of the Involvement Load Plus proposed by Yanaginawa and Webb (2022). The researchers conducted a meta-analysis of 42 empirical studies to explore whether the ILH could provide an effective predictive power to the tasks and incidental L2 vocabulary learning and the different effects of each ILH domain. According to their meta-analysis results, the ILH tends to be an effective framework to measure both word gains and short- and long-term retention. Yet, the influence of *search* on word retention seems to be very little compared to the other two domains: *need* and *evaluation*. Therefore, they argued that the degree of prominence given to the *evaluation* domain of the ILH was superior than *need* and *search* and suggested that the ILH criteria should be refined by adding one extra plus to *evaluation* to differentiate its power to the other domains: *need* and *search*. To improve the effectiveness of the ILH, in 2022, Yanaginawa and Webb proposed a new framework called the ILH Plus. Drawing from the comprehensive reviews of their prior meta-analysis study (Yanaginawa & Webb, 2021), they realised that the three ILH domains (*need, search* and *evaluation*) might be improved by (a) weighting or adding pluses for the 'evaluation' domain, (b) separating types of evaluation (e.g., receptive/productive retrievals, sentence writing, composition writing), and (c) including influential learning factors such as test format, frequency, number of target words.

The most fitted combinations of predictors contributing to learning comprise (a) *need*, (b) *evaluation*, (c) *sentence-level varied use*, and (d) *composition-level varied use* (p. 1296). They also provided the incidental vocabulary learning (IVL) formulas to be calculated prior to the construct of  the ILH plus. These formulars which consist of seven predictive variables (*need, search,* *evaluation, sentence-level varied use, composition-level varied use, frequency,* and *mode*) help them to estimate the tasks that contribute to the proposed models for the ILH plus. They explained that:

> *Based on the proposed formulas, we propose an ILH Plus:*
> *1. With other factors being equal (i.e., with the same test format at the same timing, the same set of target words, and dealing with the same population of participants), language activities with a higher effectiveness index calculated with the IVL formulas will lead to greater incidental word learning than activities with a lower effectiveness index.*
> *2. Regardless of other factors that are not included in the IVL formulas, language activities with a higher effectiveness index will lead to greater incidental word learning than activities with a lower effectiveness index. (2022, p. 1300)*

Various variables and factors that might affect the evaluation results are taken into consideration in the ILH plus. However, the factors such as test day and test format that are not related to learning conditions were excluded since these factors are closely related to learning gains, not learning conditions and cannot be calculated in the same manner as the original ILH. However, the accuracy of the ILH plus should be examined more widely from empirical studies in various contexts in order to confirm its effectiveness over other vocabulary frameworks. While this ILH Plus seems to be a good alternative option, it tends to be hard to follow when applying it. This is because there is no fix pattern or criteria provided for ease of evaluation. Different raters might rely on their personal judgement for assessment which may cause bias and invalid results if no precise criteria are written or shown in a detailed table as a guidance for effective evaluation.

Another review study on Involvement Load Hypothesis was conducted by Hazrat and Read (2022) to look for issues arisen from the uncertainty about relative weight of the ILH components: *need*, *search* and *evaluation*. Their study includes two mains purposes: 1) to seek for related issues on the application of ILH framework to data analysis and 2) to study relative factors affecting the predictive power of the original ILH. They concluded that the ILH components should be refined to make the predictive power more reliable. In recognition of *need,* the component should be expanded according to Hazrat's (2020) suggestions into three sub-categories: no motivation for learning (weak, or -), task-driven (extrinsic) motivation (moderate, or +) and self-driven (intrinsic) motivation (strong, or ++) (Hazrat and Read, 2022, p. 395). This is because *need* is considered as a motivational domain. It should include not only self-driven motivation, but also task-driven motivation. A boarder view of motivation could help to enhance the degree of its predictive power.

With respect to *search*, its involvement load can be influenced by types of *evaluation* (i.e., receptive/productive retrieval, sentence writing, writing composition) as confirmed by Hazrat' s (2020) study. It is suggested that *search* should be a subordinate component of *evaluation* as it is linked to evaluation types.

In terms of *evaluation*, the evidence from previous studies (e.g., Zou, 2012; Hazrat, 2020) demonstrates that sentence-writing and composition-writing induced high evaluation to different extents while they were rated similarly as strong (++) for *evaluation*. The *evaluation* component which consists of three criteria (-, +, ++) should be refined by adding a 'very strong involvement' criterion and signified by triple plus (+++). This is because the findings showed that sentences-writing (SW) and composition-writing (CW) tasks were more effective than the other two gap-filling activities focusing on receptive (RR) and productive receival (PR) although the sentences-writing (SW) and composition-writing (CW) showed the same ILH degree (++) to the productive receival task. If the suggested criteria is applied, the SW and CW will be given 'very strong' (+++) involvement to the *evaluation*, making the result more valid and corresponding to the ILH predictive power. Therefore, they concluded that type of evaluation (i.e., receptive/productive retrieval, composition writing and sentence writing) could also have an impact on the degrees of prominence and should be carefully assessed using the four criteria (-, +, ++, +++) for *evaluation*. Although the findings from studies of Zou (2012) and Hazrat (2020) showed different superior effects of these two activities, they supported that *evaluation* should be expanded by adding an extra plus (+++) to these two activities so that it can give a valid predictive power to the evaluation. Also, they agreed with Keating's (2008) findings that time on task is considered as a crucial factor affecting the evaluation and test results.

Although recent research has an attempt to provide plausible ways to expanding on the ILH, I reviewed another vocabulary framework: Technique Feature Analysis (TFA) due to several reasons. First, while the refined ILH and the ILH Plus can be an option for the current study, there are very little empirical evidence to confirm its effectiveness as they are very recent. Second, research on the superiority of these frameworks over the original ILH and/or TFA is very limited. The previous results were concluded based on the findings from the researchers who developed and proposed the new frameworks themselves. In contrast, there is a great deal of empirical evidence supporting the superiority of the TFA over the ILH (see details in Section 2.3.5). Finally, the current study was designed prior to the publication of these two frameworks. The articles proposing the refined ILH and ILH Plus were published in 2021 and 2022 when I have already completed the data collection, within the time frame of my Ph.D. study. As a result, I cannot redo the study again due to time constrain which I am aware that this could be one of the limitations of the present study. For these reasons, I discuss another well-accepted framework in section 2.3.3 below.

### 2.3.3 Technique Feature Analysis (TFA)

The Technique Feature Analysis (Nation & Webb, 2011) is based on the idea that to stimulate deep processing in language learners' brains, activities that involve a degree of elaboration, quality of attention, richness of encoding, and linking of form and meaning are necessary (Craik & Lockhart, 1972; Nation, 2001). Therefore, this framework evaluates the effectiveness of vocabulary learning activities among 18 criteria from five components: *Motivation, Noticing, Retrieval, Generative Use* and *Retention* (see Table 2.2 below). With more precise criteria compared to the original Involvement Load Hypothesis, the TFA has been widely accepted by scholars in the field of vocabulary (e.g., Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018; Gohar et al., 2018; Zou & Xie, 2018).

The TFA is constructed based on the previous framework of Nation (2001) on vocabulary learning. He suggested that three components: *Noticing, Retrieval*, and *Generative Use* are involved in learning new words. In the modified version, two more components: *Motivation* and *Retention* are added.

Nation and Webb (2011) have presented eighteen guiding questions for each component in this latest version. The degree of effectiveness on vocabulary learning depends on the score of task evaluation. One question is equal to one point. For instance, if a task has a clear vocabulary goal, a score of one (1) is given, but if the learning goal is not clearly presented to the learners, this task will receive a score of zero (0). A task will receive a maximum score of eighteen if it meets all eighteen criterion. The higher score the task receives, the better result in vocabulary learning and retention. As presented in Table 2.2, the TFA framework contains several components that are recognised as the influential factors leading to learning. The information below presents the description of each component according to Nation and Webb (2011) and Webb and Nation (2017) with some concerning issues. In the below sub-sections, I reviewed the components of the framework in order to discover variables and research gaps that should be addressed in this study.

***Table 2.2. Technique Feature Analysis (Nation & Webb, 2011, p.7)***

| **Component** | **Criteria** | **scores** | |
|---|---|---|---|
| *Motivation* | Is there a clear vocabulary learning goal? | 0 | 1 |
| | Does the activity motivate learning? | 0 | 1 |
| | Do the learners select the words? | 0 | 1 |
| *Noticing* | Does the activity focus attention on the target words? | 0 | 1 |
| | Does the activity raise awareness of new vocabulary learning? | 0 | 1 |
| | Does the activity involve negotiation? | 0 | 1 |
| *Retrieval* | Does the activity involve retrieval of the word? | 0 | 1 |
| | Is it productive retrieval? | 0 | 1 |
| | Is it recall? | 0 | 1 |
| | Are there multiple retrievals of each word? | 0 | 1 |
| | Is there spacing between retrieval? | 0 | 1 |
| *Generative* | Does the activity involve generative use? | 0 | 1 |
| *Use* | Is it productive? | 0 | 1 |
| | Is there a marked change that involves the use of other words? | 0 | 1 |
| *Retention* | Does the activity ensure successful linking of form and meaning? | 0 | 1 |
| | Does the activity involve instantiation? | 0 | 1 |
| | Does the activity involve imagination? | 0 | 1 |
| | Does the activity avoid interference? | 0 | 1 |
| **Maximum score** | | | **18** |

#### 2.3.3.1 *Motivation*

There are three questions in the checklists of *Motivation.* The first question, *'Is there a clear vocabulary learning goal?'* aims at the goal of vocabulary learning. Nation and Webb believe that motivation could be driven when learners know what they are trying to achieve. If the proposed activity has a clear vocabulary learning goal, this criterion will be given one point. However, learners might not recognise the learning goal even though it is clear to them that they are learning about vocabulary if it is not explicitly informed. This can lead to the variation in scoring from different opinions of raters. The use of questionnaire to obtain learners' perceptions towards vocabulary learning activities could be one of the possible solutions to ensure the results. The second question *'Does the activity motivate learning?'* seems to put an emphasis on the activity design. Nation and Webb suggested that challenging or pleasant activities tend to motivate learning. This criterion also seems to be too board. Raters need background knowledge on motivation or are required to study about it in details in order to rate this. Due to this issue, a provision of trainings or workshops to make sure that raters have the same concept of understanding tends to be essential. Researchers (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, et al., 2010) suggested that rater training is essential because levels of agreement can be increased with training. It helps to promote a common understanding and eliminate the problems regarding personal beliefs and biases from different raters. The TFA score can vary if some raters rely only on their opinions without searching for findings from previous studies on how motivation can be driven. The last criterion under this component is *'Do the learners select the words?* Strong interest can increase motivation in learning (Nation & Webb, 2011). If learners are allowed to select the words they wish to learn, vocabulary learning tends to be successful, and one point is given to this criterion.

#### 2.3.3.2 *Noticing*

This component consists of three questions: *'Does the activity focus attention on the target words?*', *'Does the activity raise awareness of new vocabulary learning?*', and *'Does the activity involve negotiation?*' The first criterion focuses on attention (see more details in section 2.3.4.2). It seems that this criterion pays particular attention only to the target words. As a result, the degree of attention might be decreased if learners are given several extra words to study with a list of the target words. The Input Processing Theory (VanPatten, 2002, p.758; 2004, p.8) suggests that “Learners process content words in the input before anything else.” Extra words that are contents words may distract attention of learners away from the target words. Moreover, it could be interpreted from the second criterion that the term ‘new vocabulary’ could mean any new words that learners encounter from the learning, not only the new target words to be studied. Raters can think that if learners are aware they are learning any new words, not a set of the target new words in particular, this criterion should be given one point. If this is the case, the analysis result might not be accurate because sometimes not all new words are the focus of the learning. Thus, each rater needs to have a clear and same understanding on this in order to rate the criterion accurately. The last criterion focus is on negotiation. If learners negotiate by discussing the meaning or features of words, this criterion should be given one point. Again, some raters might not be familiar with the term ‘negotiation’. It requires background knowledge in order to rate this criterion effectively. Also, it is not clear from Nation and Webb’s (2011) explanation about the amount of time needed for negotiation to be effective. It is questionable whether only an opportunity to negotiate a small language feature should be counted. Thus, the score for this criterion seems to depend on raters’ opinions which can be bias. Organising a training to raters, as stated, can be a solution to eliminate the problems.

#### 2.3.3.3 *Retrieval*

This component is related to memory. *Retrieval* tends to play a crucial role in word learning and remembering. There are five questions as criteria for evaluation. These criteria have been developed based on research on memory and learning conditions (see Nation & Webb, 2017). According to the framework, if the activity involves either receptive retrieval (retrieving word meaning) or productive retrieval (retrieving word form) of the word, one point will be given to the first question, *‘Does the activity involve retrieval of the words?’* Because meaning tends to be learned earlier (VanPatten, 2002; 2004) and retrieved easier (Nation & Webb, 2011) than form, retrieval of word form (productive retrieval), which seems to be more demanding, will get one more score in the second question of this component, *‘Is it productive retrieval?’* Besides, the process of memory retrieval involves both recognition and recall (see more details in Section 2.3.4.3). The difference between recognition and recall concerns the provision of clues during the process of retrieval. On the one hand, it is recognition when learners have some options or clues to look at or listen to while they try to retrieve words from memory. On the other hand, it is recall when there is no clues provided to assist memory retrieval. As recall seems to be more challenging than recognition (Nation & Webb, 2011), it gets one point from the third question, *‘Is it recall?’* The next question is *‘Are there multiple retrievals of each word?* If the target words are retrieved several times from memory, it is likely that the words can be stored in learners’ memory. So, multiple retrievals are useful to vocabulary learning and gains. However, Nation and Webb did not define the term ‘multiple retrieval’ clearly. It is questionable whether only two opportunities to retrieve each word can be considered as multiple retrieval. We can only give one score to this criterion although there might be differences between having less and more chances of multiple retrievals. If raters are not in the field of memory, nor lack of background knowledge on retrieval, they might justify this criterion differently, which can affect the analysis result. This is another reason why a training to provide raters background knowledge about TFA terms seems to be crucial. Likewise, the last question, *‘Is there spacing between retrieval?’* is associated with time between the first and next retrieval. It is suggested that multiple retrievals tend to be strengthened if the target learning words are not successively learned with no space between retrievals. The question is *‘Can 5 minutes be considered as enough space?’* In order to answer this question, raters need to search for more information and rely on evidences from previous empirical studies (e.g., Cepeda et al., 2006; Nakata, 2015; Nakata & Elgort, 2021). This process can take time. If raters avoid to do this, it can affect the results of the analysis. Different raters may have different justifications if the duration of space between retrievals is not explicitly stated. A clarification of this component would help raters, especially those who are not in the field of memory, get a consistent result from the evaluation.

#### 2.3.3.4 *Generative Use*

This component, consisting of three criteria, pays attention to varied encounters and varied use of the target learning words. *Generative Use* can be divided into two types: receptive generative use and productive generative use as explained in the previous section. These features are used to form two main questions under the *generative use* component: 1) *Does the activity involve generative use?* and 2) *Is it productive?* each question gets one point if the activity provide an opportunity for learners to meet (receptive) and use (productive) the target learning words in a new way that they have never encountered it before. Productive generative use will get one more point for the second question as it is more demanding than receptive generative use. This includes generative use in both word level and sentence level. Thus, using different word forms and completing word parts also get one point from the second criterion. Nation and Webb (2011) also provide information derived from Joe (1998) about various degrees of productive generative use, starting from no generation to high generation. This information tends to shape the second and third criteria under this component. The last criterion, *Is there a marked change that involves the use of other words?* is associated with productive generative use. One more point is given to this criterion if learners have a chance to use or write the target learning words in a new different sentence or context because it involves higher degree of productive generation according to the scale proposed by Joe (1998). Despite the different degrees, we can only give one score to this criterion.

#### 2.3.3.5 *Retention*

*Retention* comprises four criteria including several learning variables, related to the four questions in Table 2.2. The first variable is about the link between form and meaning. If there is a chance that the activity can ensure this link successfully, the first criterion, *Does the activity ensure successful linking of form and meaning?* gets one point. However, identifying a successful link is difficult without a test because of differences in the depth of vocabulary knowledge. As Henriksen (1999) noted, there are different degrees of word knowledge from partial to precise, and this makes it difficult to know if there is a successful link between form and meaning before testing. Although Nation and Webb (2011) suggested that some activities such as using dictionaries, word cards, and reading with glosses can ensure successful linking of form and meaning, these activities can be designed or modified differently, which can affect the degree of linking between form and meaning. This brings me to the questions ‘To what extend the degree of linking between form and meaning is considered to be successful?’ and ‘Do we need a test to ensure this link?’ If the test is not utilised, the raters need to be trained to make sure that they have background knowledge of word retention and related vocabulary activities in order to rate this criterion effectively. The second variable within retention is instantiation. Nation and Webb (2011, p.10) maintained that “Instantiation involves seeing an instance of a word such as when the word is used in a meaningful situation where the object, action, or quality referred to is visually present.” Hence, seeing a sentence containing the target word in a reading passage would be instantiation. If the activity meets this condition, it gets a point for the second criterion, *Does the activity involve instantiation?* The next criterion, *Does the activity involve imagination?* focuses on a deliberate occurrence of a mental image related to the word meaning while hearing or seeing the target word. The keyword technique, as suggested by Nation and Webb, is an example of the activity that involves making visual images. Though, it is unclear whether the activity should or should not get one point for this criterion if the list of target words in an activity contains some words such as *portray* and *paradigm* that might be difficult to visualise. Lastly, this component includes the question *Does the activity avoid interference?* to put an emphasis on the influences of negative transfers. If the activity involves L2 related words such as new vocabulary from the same lexical sets, near synonym, opposites, and cognates, it will not get a point for this criterion.

Tasks or activities can be rated through these eighteen suggested questions under the five components (see Table 2.2). Several vocabulary learning activities are evaluated using these eighteen criteria by Nation and Webb (2011) as shown in Table 2.3.

***Table 2.3. Analysis of twelve vocabulary activities by Nation and Webb (2011, p. 14)***

| **Activity** | **TFA score**<br>(Total score of 18) |
|---|---|
| Fill in the blank | 8 |
| Find the words in the text | 8 |
| Word part table | 9 |
| Write with target words | 8 |
| True/false | 6 |
| Reword the sentence | 6 |
| Multiple-choice on text | 6 |
| Wordcards/ flashcard | 11 |
| Read and choose definitions | 6 |
| Keyword | 8 |
| Reading plus fill in | 7 |
| Reading with glosses | 5 |

In conclusion, the TFA framework which has five components (*Motivation*, *Noticing*, *Retrieval*, *Generative Use*, and *Retention*) has been constructed to assist the analysis of vocabulary tasks based on notions of vocabulary learning. To provide further insights in each TFA component, in the next section, I will discuss theories driven to the construct of the framework.

### 2.3.4 Learning factors informing the development of the TFA framework

In this section, I explain theories behind the components of the TFA framework. Although a number of factors influences to successful L2 learning, there are four general components: motivation, noticing, retrieval, and generative use that seem to involve in word learning and retention (Nation & Webb, 2011; Nation, 2013), and be proposed as main vocabulary learning components in Nation and Webb's (2011) Technique Feature Analysis framework. These components were discussed in the following sub-sections.

#### 2.3.4.1 Motivation

Motivation which can be seen as a social-psychological factor is believed to be a driven mechanism of language-learning success. Researchers (e.g., Dörnyei, 2001a; Gardner, 2001; Ushioda, 2003; Shuman, 2014) have confirmed that motivation has a positive impact on second language acquisition (SLA) and language achievement. According to Dörnyei and Csizer (1998, p. 203), “Without sufficient motivation, even individuals with the most remarkable abilities cannot accomplish long-term goals, and neither are appropriate curricula and good teaching enough to ensure student achievement.” This belief leads to the question about how learners can be motivated. Motivation, which is a complex dynamic process (Garner & MacIntyre, 1993; Dörnyei, 1994), can be divided into two main types: intrinsic motivation and extrinsic motivation (Brown, 1990). While intrinsic motivation can be driven by internal stimuli such as the pleasure in doing a task or self-interest in exploring new knowledge and solving problems, extrinsic motivation can be stimulated by external stimulus such as incentives, grades or extra credits (Dörnyei, 1994; Sternberg & Williams, 2002). Previous studies (e.g., Peacock, 1997; Hamada & Kito, 2008; Sakai & Kikuchi, 2009; Yoshimura, 2017) found the relationship between coursebook, materials, and motivation. Motivation can be decreased with the use of ineffective coursebook (Hamada & Kito, 2008) or demotivated learning materials (Sakai & Kikuchi, 2009), regarded as extrinsic stimuli. However, I have realised that challenging activities that engage students to learn might also lead to intrinsic motivation. Barry and King (2000) stated that factors such as enjoyment, challenge, and interest can result in intrinsic motivation as learners may decide to engage in doing activities because they feel enjoyed or challenged. Yet, little is known about the relationship between intrinsic motivation and materials although the effects of using learning materials may affect both intrinsic and extrinsic motivation. As argued by initial research that only intrinsic motivation may not be able to drive learners to sustain language learning (Noels et al., 2000), the current study focused on external stimuli which are vocabulary learning materials that may drive extrinsic motivation to occur in classroom settings. The aim of this study therefore tends to be more related to extrinsic motivation because it can be activated by rewards after achieving challenging activities or winning games.

Several frameworks (e.g., Dörnyei, 1994; Williams & Burden, 1997; Coyle, 2011; Bower, 2019) have emerged to conceptualise motivation in the field of language acquisition and learning. Recently, Bower (2019) proposed the Process Motivation Model that includes learning environment, learner engagement and learner identities/self. This framework seems to pay much attention to the aspects of learners which could be seen as internal factors. Because I do not have an intention to explore either intrinsic motivation or the whole process in driving motivation, this model seems to be irrelevant to the current study. Likewise, Coyle's (2011) framework may not suit the main purpose of this study which focuses on analysing learning materials although it is more current than that of Dörnyei (1994) and Williams and Burden (1997). The model puts an emphasis on motivation in the Content-and-Language Integrated Learning (CLIL) setting which this study does not aim to explore.

William and Burden' s (1997, p. 24) model (see Figure 2.1) of motivation provides a concrete concept of what make learners decide to learn and how to sustain an effort in learning through motivation. Building on this, external stimuli such as well-designed learning activities, materials, and interactions with significant others such as teachers and peers are believed to have significant influences to motivation and help sustain learners' efforts in doing activities as well as learning.

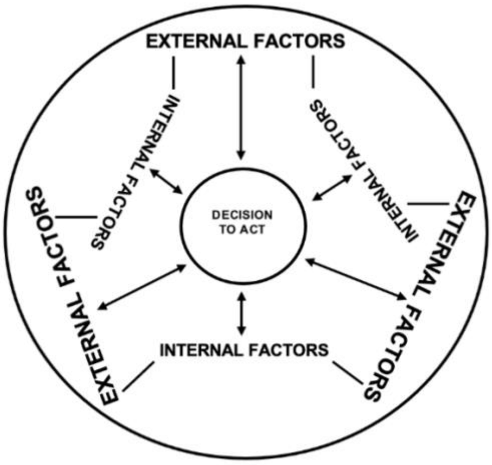


**Figure 2.1. A model of motivation adapted from Williams and Burden (1997, p. 24)**

This conceptualisation allies with what could be observed from my teaching experience. Decisions about learning and maintaining the efforts in classroom may depend greatly on extrinsic motivation that can be driven from external stimuli such as materials. William and Burden's model also relates to the Learning Situational Level of Dörnyei's (1994) Components of Foreign Language Learning Motivation Model in that external stimuli such as course, materials, teachers, and peers involve in the constructs of motivation in language learning. It is suggested in the Learning Situational Level that implementing tasks with game-like features, for example, could motivate students to have an interest and engagement in learning (Dörnyei, 1994). However, this model pays more attention to internal stimuli as it also includes two levels: Language Level and Learner Level that seem to have an influence to intrinsic motivation. This makes William and Burden’s (1997) model more relevant to the main purpose of this study on analysing materials in order to explore the effects of motivation.

Earlier studies (e.g., Elley, 1989; Bailey et al., 1999; Amiri & Salehi, 2017) revealed that motivation has also shown a strong relationship with vocabulary learning. Bailey, Hsu, and Dicarlo (1999) used extrinsically motivating vocabulary games such as crossword puzzles and word scrambles in teaching new vocabulary in a medical physiology course and found that both activities could motivate students to master the target learning words even though crossword puzzles resulted in the most effective learning outcomes. The finding is supported by a recent study by Amiri and Salehi (2017) using crossword puzzles to motivate learning and enhance word spelling ability.

The empirical evidence (e.g., Bailey et al., 1999; Amiri & Salehi, 2017) shows that effective materials can enhance greater extrinsic motivation that leads to language achievement of learners. Because learning materials, as an external stimulus, could result in more or less extrinsic motivation, they should not be neglected in any language classrooms. Although motivation could be seen as one of the influential factors affecting learners’ vocabulary development and enabling noticing, which is another significant feature of vocabulary learning (Nation, 2013), the extent to which motivation can be promoted through the TFA framework in order to support word retention in written production has not been revealed yet. As Nation and Webb (2001) claimed that three criteria (see Table 2.1) under the *motivation* component of the TFA framework can be used to evaluate motivation in vocabulary activities, I applied this framework to evaluate learning materials for word retention. The effects of the framework on motivation were explored in this study. In the following sub-section, another component that may lead to word retention was reviewed.

#### 2.3.4.2 Noticing

Intentional learning depends on noticing which is a determinant of what will be learned (Barcroft, 2009). According to Schmidt' (1990, p. 134), "the conscious/unconscious learning contrast may refer to awareness at the level of noticing." With regard to cognitive view, issues on consciousness, which refers to awareness (Battista, 1978; Schmidt, 1990) or intention (Schmidt, 1990), and unconsciousness in L2 acquisition and learning have been debated for decades and have influenced to Noticing Hypothesis proposed by Schmidt (1990; 2001). Awareness is required in early stages of procedural knowledge (Anderson, 1982; Schmidt, 1990) which is the knowledge used to develop cognitive or brain-based skills (Sharwood-Smith, 1981; Sorace, 1985). It also relates to ideas about depth of processing which is believe to stimulate better memory performance in terms of retention (Anderson, 1995; Baddeley, 1997). Although there are different notions of consciousness and information processing, according to Schmidt (1990), some common agreements are 1) consciousness tends to be equally recognised as focal awareness and short-term store (Kihlstrom, 1984) in primary memory, and 2) long-term storage requires information processing in short-term memory (working memory), where attention works as a pathway bringing information to the memory (see memory model of Waugh and Norman (1965), Atkinson and Shiffrin (1968) and Kihlstrom (1984) in Figure 2.2 below.

Schmidt' s (1990) Noticing Hypothesis also arises from his experience as a L2 learner of Brazilian Portuguese participating in a research conducted with Frota (Schmidt & Frota, 1986). They found that only some input of instructed information was produced in his speech production while some features from the input never showed up, and frequency of occurrence depends largely on frequency of input he received. He, therefore, drawn his hypothesis that "[i]ntake is that part of the input that the learner notices" (1990, p.139). This means that intake requires noticing to raise awareness of new learning information. So, Schmidt' (1990) s Noticing Hypothesis entails both attention and awareness within the concept.

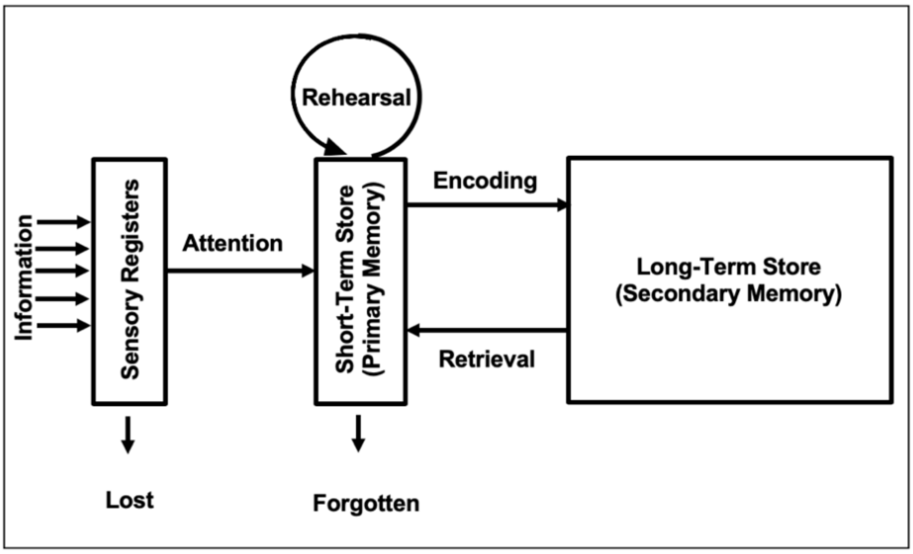


**Figure 2.2. Multi-store Model of Memory adapted from Waugh and Norman (1965), Atkinson and Shiffrin (1968), and Kihlstrom (1984)**

By linking this hypothesis to vocabulary learning, significant factors leading to noticing are frequency of input and task demands (Schmidt, 1990). Studies (e.g., Rott, 1999; Larsen-Freeman, 1976; Goldschneider & DeKeyser, 2001) found that input frequency is a powerful determinant for acquisition of language elements such as morphemes and lexical words or phrases. Furthermore, what can make input more noticeable is demand of tasks (Ericsson & Simon, 1984; Kahneman, 1973; Kilhstrom, 1984; Schmidt, 1990). Ericsson & Simon (1984) claim that information stored in memory is triggered by tasks that are processed while learning. Gu (2003) also reported that tasks such as guessing from context and note-taking are effective to vocabulary learning when the teacher intentionally makes them noticeable to learners. Target learning words should be decontexualised or taught separately from the context of learning in order to make them noticeable. This technique, which is called decontextualisation, leads noticing to occur (Nation, 2013, p. 168). However, there has been limited in number of empirical evidence to support that noticing can be encouraged by the TFA framework to promote word retention in written production. It is essential to explore the effects of the TFA framework on noticing as the results may help to strengthen vocabulary learning in classroom. It should be noted that *noticing* in this study would mean learning a set of target vocabulary items intentionally, rather than learning second language consciously in general. This is to investigate whether a provision of target vocabulary has more or less effects on noticing which may lead to word retention. Thus, in the current study *noticing* was analysed by using the TFA framework to explore the effectiveness of the framework on noticing and word retention.

Apart from noticing, retrieval and generative use are also cognitive processes that may help to increase word retention (Nation, 2013, p. 102). These components are suggested by Nation (2013) as three learning steps in which noticing occurs first, followed by retrieval and generative use, respectively. However, it is still uncertain to identify that which of these three components could better strengthen word learning and retention. Moreover, there is no empirical evidence to support the association between them. It is not clear whether retrieval and generative use require noticing in the primary stage of learning through materials analysed by using the TFA framework. Also, without a clear introduction to the target learning words it is essential to investigate the extent to which word can be retained in memory. I then view noticing, retrieval and generative use as a separate component that may lead to different degree of vocabulary retention.

#### 2.3.4.3 Retrieval

According to Webb and Nation (2011), retrieval can occur after a word is previously met by learners and is recalled from their memories. So, it is a cognitive process that influences to remembering words (Baddeley, 1990). Retrieval can be catergorised into two main types: receptive and productive retrieval. Receptive retrieval, which is a recognition process, occurs when form or meaning of a word is recognised through listening or reading after learners first encounter with the word while productive retrieval involves process in which the word is recalled and used in spoken and written modes. It can be triggered by either free recall or cued recall. Productive retrieval (recall) seems to be more powerful than receptive retrieval (recognition) in terms of enhancing vocabulary learning (Griffin & Harley, 1996) because learners have a chance to produce the language by themselves. When a word is retrieved from memory it leads to repetition, another important factor in vocabulary retention (Elley, 1989; Baddeley, 1990; Webb, 2007a; Brown et al., 2008; Vidal, 2011), because it may increase the chance for the word to be stored in long-term memory as can be seen in the memory model in the Figure

2.2 above. Webb (2007a) and Karpicke and Roediger (2008) found the relationship between vocabulary retention and various degrees of repetition. The more the learners meet with the target items, the better the results of vocabulary learning. However, repetition may not be effective if the learners lose memory of the words they previously met (Nation, 2013). Therefore, length of time between the first encounter of a newly learned word and second or subsequent encounters seems to be one of the factors affecting vocabulary retention. Nation (2013, p. 109) claims that “retrievals need to be spaced rather than massed together.” By giving more space for learners to retrieve the target learning words, the words can be retained longer in their memory (Karpicke & Bauerenschmidt, 2011).

Empirical research showed that retrieval practice shows positive results on immediate (Candry, et al., 2020) and (one-week) delayed form recall (Van den Broek, et al., 2018; Candry, et al., 2020). With regard to earlier long-term retention studies, researchers found that words can be stored in the primary (short-term) memory for one or more than two weeks (Avila & Sadoski, 1996; Waring, 1997a; Keating, 2008) or several months (Elley, 1989; Ellis, et al., 1994; Waring & Takaki, 2003) before fading away if they are not retained in long-term memory which is the secondary memory. According to this, I will define retention in the current study as the ability to store target learning words, which can be divided into two categories: short-term and long-term retention. Short-term retention is ability to retain word knowledge on form and meaning immediately after finishing the learning session while long-term retention would mean ability to store word knowledge for two weeks following previous empirical research (Avila & Sadoski, 1996; Waring, 1997a; Keating, 2008; Puimège & Peters, 2019). It should be noted that the term ‘long-term retention’ used in this study is not the same as long-term memory which is more persistent and widely used in the field of psychology. In this regard, I use the terms short-term and long-term retention mainly for methodological purposes in the current study.

While retrieval plays a crucial role in memory retention, findings from empirical research concerning the analysis of materials focusing on *Retrieval* through the TFA framework has been scarce. Therefore, this study aimed to seek for the answer whether the framework can be used effectively to encourage retrieval in vocabulary learning materials that may foster word retention.

#### 2.3.4.4 Generative use

The last component that involves cognitive processes and may lead to vocabulary retention is generative use. Various studies found that tasks that require generative use (also called creative use or creativity by Nation, 2013, pp.110-111) can strengthen word knowledge as learners have several chances to encounter with the words in different ways (Baddeley, 1990; Ellis, 1995; Joe, 1995; Newton, 2013). Webb and Nation (2011) suggest that the more a word is encountered in various ways, the better it is learned. That means generative use increases elaboration. Word knowledge can be elaborated when learners encounter with a word used for different meanings or in various forms. It leads to enrichment of a word to be learned as it requires association between the word that is previously met and the word that is presented in another context (Nation, 2013). It can be activated by several stimuli such as pictures, keyword technique, and word parts (Webb & Nation, 2011). Research indicates that pictures can facilitate vocabulary learning (Yeh and Wang, 2003; Carpenter and Olson, 2012; Strauber and Goldman, 2020) because the association between a picture and a word being learned may generate mental image that is retained in their memory. Besides, learners may memorise the target learning word because it shares similar sound with their L1 when it is learned by using key word technique (Pressley, 1977), or they may comprehend the meanings of the word more deeply when its derived forms are taught with its stem or base form (Wei & Nation, 2013). Generative use can be both receptive and productive. While meeting previously learned words again in listening and reading creates receptive generative use, using them in another spoken and written context influences to productive generative use (Nation, 2011). The degree of elaboration seems to be higher if learners meet the word again in a meaningfully different way (Nation, 2013). For example, they know that the word *plant* is a noun, but may happen to meet this word again used as a verb. They will reconceptualise their knowledge of the word, and this requires cognitive processes. Therefore, it can lead to a more powerful learning than having a chance to encounter the word, *plant* used as a noun in different sentences. However, if this word is used productively in another different way, learners' memory can be well strengthened (Joe, 1998). Although this component has been identified by a number of researchers as facilitating vocabulary learning, it remains unclear whether generative use alone can promote word retention. As mentioned earlier, it is argued that generative use may require noticing and retrieval in the primary stages of learning (Nation, 2013). Thus, empirical research should pay attention to the effects of generative use with none or little involvement of other components such as motivation, noticing, and retrieval in order to seek for the findings. Because of this, the extent to which the TFA framework can encourage generative use in vocabulary learning materials that foster word retention was explored in the current study.

In conclusion, motivation, noticing, retrieval, and generative use tend to be crucial components involving in vocabulary learning and retention. However, the focus of this study is specifically on components that may affect word retention on written output. It has been revealed from empirical research on written production that motivation (Zimmerman & Kitsantas, 1999; Hashemian & Heidari, 2013), noticing (Qi & Lapkin, 2001; Adams, 2003; Park, 2011), and retrieval (Snelling, et al., 2004) can drive better writing performance, but these components have not yet been compared to explore their degrees of effectiveness towards vocabulary retention. It seems that none of the previous studies on TFA framework (e.g., Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018; Gohar et al., 2018) regarded on this issue even though it tends to affect vocabulary learning. Besides, generative use seems to be neglected in the area of L2 written production although previous research (i.e., Smith et al., 2013) has confirmed positive effects of generative learning towards vocabulary retention. Generative use may occur in a productive condition that learners are encouraged to form sentences using the target words in different contexts or produce different forms of the target learning words. For instance, they may learn the word *influences* (plural n.) from a reading passage but may then be encouraged to form a sentence using the target word as a verb, *influence*. As generative use may lead to long-term vocabulary retention, and little has been done to explore whether it can facilitate written output, I therefore have an attempt to find out the answer in the current study.

When we know that motivation, noticing, retrieval and generative use may lead to learning, it is important to look for the effective vocabulary frameworks that can support retention of form recall. To find more gaps in using the TFA framework, I sought for findings from previous studies in order to discover research critiques related to its predictive power in facilitating vocabulary retention in the following sub-section.

### 2.3.5 Studies investigating the TFA

Previous research on the TFA has mainly compared it with the Involvement Load Hypothesis (ILH) (Hulstijn & Laufer, 2001) (e.g., Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018; Gohar, et al, 2018; Zou & Xie, 2018) (see Table 2.4). These studies have consistently showed that the TFA had greater reliability than ILH in predicting vocabulary learning gains.

***Table 2.4. Activities used in previous studies on ILH and TFA***

| Author(s) | Year | Type of source | Vocabulary tasks/learning tools | Tool(s) |
|---|---|---|---|---|
| Hu and Nassaji | 2016 | experiment | 1) multiple-choice questions, 2) choosing definitions, 3) gap-filling, and 4) rewording sentences | ILH* and TFA** |
| Nakata and Webb | 2016 | Journal report | 1) flashcards, 2) cloze exercises, and 3) crossword puzzles | TFA |
| Khoshsima and Eskandari | 2017 | experiment | 1) multiple-choice questions, 2) choosing definitions, 3) gap-filling, and 4) rewording sentences | ILH* and TFA** |
| Chaharlang and Farvardin | 2018 | experiment | 1) reading with glosses, 2) keyword techniques, 3) word card, and 4) reading and finding the words in text | ILH* and TFA** |
| Hirata | 2019 | Journal report | Multi-skill vocabulary activity: 1) reading with glosses, 2) writing by drawing a free form outline of the story, and 3) speaking (story retelling) | TFA** |

*Note: ILH* stands for Involvement Load Hypothesis and TFA** stands for Technique Feature Analysis*

Hu and Nassaji (2016) compared four different vocabulary tasks: reading with multiple-choice questions, reading and choosing definitions, reading and gap-filling, reading and rewording sentences with various scores of ILH and TFA. Each experimental group was required to perform different tasks with various ILH and TFA predictive scores. The tasks were adapted from vocabulary activities suggested by Nation and Webb (2011). For task 1 (ILH score of 3; TFA score of 6), students were assigned to read the text and answer multiple-choice questions provided after reading. Task 2 (ILH score of 3; TFA score of 6) involved reading the text and choosing definitions while Task 3 (ILH score of 2; TFA score of 7) required students to read the text with 14 blanks, look at the target words provided and fill the most appropriate word in each blank. Task 4 (ILH score of 3; TFA score of 6) encouraged students to read the text and rewording the sentences. For this task, students were assigned to rewrite the sentence from the text using another word to replace the target word for each sentence. After finishing the four tasks, a posttest measuring L1 translation or L2 synonym was distributed to the students in all groups to explore their knowledge of 14 single words at the meaning recall level (see Table 2.5). However, as the researchers did not mention the time when the test was administered, it was unclear whether the posttest in their study was an immediate or delay-posttest. The findings from the pre-test and posttest through percentage, one-way ANOVA and a hierarchical multiple-regression found that TFA showed a better predictive power in explaining vocabulary gains than ILH.

In 2008, Chaharlang and Farvardin also conducted a study to compare these two frameworks and found results corresponding with the previous findings. A similar method, L1 translation for pre- and posttest, was also used to measure meaning recall of 16 single-word items of this study (see Table 2.5). They also used the same analysis method which is one-way ANOVA and multiple regression analyses. However, they employed different vocabulary activities from the previous research. The four tasks were also taken from Nation and Webb (2011). These include reading with glosses (ILH =. 1, TFA = 5), keyword techniques (ILH = 2, TFA = 8), word card (ILH = 3, TFA = 11), and reading and finding the words in text (ILH = 4, TFA = 8) to their study. Despite this, the consistency with earlier studies suggests that the TFA is a more effective technique when designing vocabulary learning tasks than the ILH.

***Table 2.5. Details of previous studies on ILH and TFA***

| **Author(s) (year)** | **Key concepts** | **No. of tasks** | **No. of target items** | **Type of items/ word-list** | **Type of WK (pre-posttest)** |
|---|---|---|---|---|---|
| Hu & Nassaji (2016) | The effectiveness of ILH and TFA | 4 | 14 | Single/ AWL | Meaning recall (L1 translation/ L2 synonym) |
| Khoshsima & Eskandari (2017) | The effectiveness of ILH and TFA | 4 | 10 | Single/ n/a | Meaning recall (adapted VKS) |
| Chaharlang & Farvardin (2018) | The effectiveness of ILH and TFA | 4 | 16 | Single/ n/a | Meaning recall (L1 translation) |
| Gohar, et al. (2018) | The effectiveness of ILH and TFA | 3 | 10 | Single/ AWL | Meaning recall (L1 translation/ L2 synonym) |
| Zou & Xie (2018) | The effects of TFA on word learning and three e-learning methods | 20 | 40 | Single/ n/a | Meaning recall (adapted VKS) |
| Kamali et al. (2020) | The effects of TFA on productive knowledge | 2 | 40 | Single/ n/a | Form-Meaning recall (recognition and cued recall test) |

*Note: No.=number(s); n/a=not known; AWL=Academic Word List; WK=word knowledge*

A similar result was found from a study conducted in Iran. Khoshsima and Eskandari (2017) evaluated four vocabulary tasks using the ILH and TFA frameworks as checklists in order to examine vocabulary gains. They adopted

Hu and Nassaji's (2016) research design, so the vocabulary tasks proposed in their experiment involved the same ILH and TFA scores as in that of Hu and Nassaji. However, they used Folse' s (2008) adapted VKS to measure meaning recall of 10 single words before and after the experiments (see Table 2.5). Pre-test and posttest were used to measure vocabulary gains after the participants received different treatments. The study also found consistent results from one-way ANOVA analysis in that the TFA framework had a stronger predictive power than the ILH even though they used different tool that could measure meaning recall knowledge at both receptive and productive levels.

Another comparison study between these two frameworks conducted by Gohar, Rahmanian and Soleimani (2018) showed dissimilar results. The study was also taken place in the EFL context similar to other research. However, they invited high proficiency level students whose TOEFL (paper-based test) total score were between 460 and 490 (out of 677) and writing score was between 4 and 6. Unlike earlier studies, two experimental groups and one control group were assigned with different tasks. The two experimental groups were assigned to do different tasks, Task A: sentence writing and Task B: writing composition. The target words with L1 definitions were introduced to both experimental groups prior to the assigned task. Task A and Task B had the same ILH score of 3 while they received the TFA scores of 7 and 9, respectively. The control group has to do the reading comprehension task (Task C) with no other vocabulary activities, receiving the ILH and TFA score of 3 and 6, respectively. Similar analysis method from previous studies was employed. One-way ANOVA and multiple regression analyses revealed that Task A (ILH = 3, TFA = 7) was not significantly different from Task B (ILH = 3, TFA = 9) even though it had lower TFA score. It could be observed that these two selected activities were very similar in manner and their TFA scores were very close to each other. This contradict result to previous research shed light to the need in conducting more empirical studies on its effectiveness. As more research has confirmed that the TFA show superior power to the ILH, researchers and language teachers seem to rely on the consistent result. Moreover, these prior studies on the effectiveness between the ILH and TFA employed the original evaluation criteria. They did not pay attention to the concerns on giving an extra plus to ‘evaluation’ as suggested by recent research (e.g., Hazrat and Read, 2021; Yanagisawa and Webb, 2021) (see discussion in Section 2.3.1). While the new ILH could be more effective than its original, we still cannot make a conclusion that it will give more valid predictive power than the TFA. This is because no existing research compares the new ILH to the TFA frameworks. For this reason, I paid more attention to the TFA whether it can give an accurate prediction to the measurement of form recall knowledge of the current study. So, research on TFA in particular was reviewed to provide a deeper understanding of the framework and its effects based on the discussion of these studies.

While the comparison between the TFA and the ILH has attracted a great deal of attention from researchers, studies focusing only on the TFA are relatively rare. It is important to look at this framework alone because the TFA criteria and factors leading to the results of evaluation can be carefully examined. Some scholars have applied the TFA framework on its own to analyse vocabulary learning activities (Nakata & Webb, 2016; Hirata, 2019) or investigate vocabulary gains and retention through personalised word learning (Zoe & Xie, 2018). Nakata and Webb (2016) utilise the guidelines from the TFA framework to analyse the strengths and weaknesses of three vocabulary learning activities: flashcards, cloze exercises, and crossword puzzles. The result indicated that flashcards which use less time for teachers to prepare received the TFA score of 12 while cloze exercise and crossword puzzles were rated scores of 9 and 7, respectively. The score for flashcards in this study is slightly higher than that of Nation and Webb (2011) presented in the Table 2.3 (also see Table 2.6). The researchers did not provide details of the flashcard activity. So, we could not know why the flashcard score is different from that of Nation and Webb (2011). This might be because vocabulary activities can be modified according to learning purposes and learners’ needs. Based on the result, the researchers suggested that the TFA framework could benefit vocabulary learning and help instructors in the process of material selection. Nevertheless, it was suggested that vocabulary activities should be modified to suit the goal of vocabulary learning and learners’ needs. For example, some modifications in terms of generative use and retention should be considered when using crossword puzzles to facilitate vocabulary learning gains. Although this study shed some light on how to select vocabulary learning activities effectively, it provided limited insights into the efficiency of the framework as the three activities were self-evaluated through the eighteen criteria without any results from experimental studies to confirm the results of the evaluation.

With regards to the effectiveness of the TFA framework, another empirical study was conducted by Zoe and Xie (2018) by applying a “user model” which is an e-learning model to promote personalized vocabulary learning with university students from different majors in Hong Kong and mainland China. The researchers investigated the effectiveness of a personalised electronic word-learning system consisting of 20 types of vocabulary learning tasks with different TFA scores. Each task mainly focused on one particular component of the TFA framework. For instance, while Task 1 imposed recalling of the target learning word under the umbrella of the “retrieval” component, Task 2 might focus on the component “retention” that involved linking of form and meaning of the target words. The checklist for TFA was applied in their research as a theoretical framework in evaluating whether the tasks with higher score from the checklist can better facilitate individual vocabulary learning of EFL students. Similar to other studies (e.g., Nakata & Webb, 2016; Gohar, et al., 2018), it only measured meaning recall of single words (see Table 2.5) and this was a self-evaluation which may limit its reliability due to bias.

Additionally, a recent longitudinal study conducted by Kamali and his colleagues (2020) also employed the TFA framework to evaluate their vocabulary tasks for advanced Iranian EFL learners. However, they focused on two different productive tasks: Oral Reproduction and Summary Writing with the same TFA score of 11 (out of 18). Unlike previous studies, the experiment with two different tasks aiming to teach 40 target words took place for eight weeks. A Pre-test with 40 target items designed by using the multiple-choice format was employed before the experiment to the participants to measure background vocabulary knowledge of the target words. After the treatments, two posttests: Immediate Posttest and Delayed Posttest were used to assess word recognition and recall by using multiple-choice items and cued recall items, respectively. In terms of evaluation, strict scoring scheme with one for a correct answer and zero for an incorrect response was used. The researchers found that even though both tasks facilitated vocabulary gains, the Oral Reproduction task had higher effects on long-term retention than the Summary Writing, and the TFA framework is an effective tool for assessing vocabulary tasks. Yet, the framework was suggested to be refined by adding one criterion: *task-induced generation* under the *Generative Use* component, making the total TFA score to increase from 18 to 19 points. According to the researchers, the Oral Reproduction groups had more chance to produce comprehensible output so the target words were generatively used with other words and/or grammatical structures more often. This led the Oral Reproduction group to have a better understanding, resulting in higher scores than the Summary Writing group in word retention.

Although the findings of previous studies have indicated the value of the TFA as a framework to evaluate the effectiveness of vocabulary learning activities, it should be noted that some issues need further investigation. First, these studies did not investigate the effects of TFA framework on each of its components so it is not clear whether all the TFA components support word retention. Second, previous studies seem to compare only three or four vocabulary activities without paying attention to effects of each TFA component that might affect the learning to different extents. Also, most studies paid more attention to receptive than productive recognition and recall while effective communicative competence entails knowledge of language production. Table 2.6 below represents the TFA scores analysed by different studies and raters. The scores of the same vocabulary activity might be different, depending largely on the design of the activity and raters' point of view. If an activity is modified, it can result in a change of the TFA score. For example, a sentence writing activity which has low scores on Retrieval component can be modified to have a higher total TFA score. If learners have at least a chance to retry and recall the target words included in the sentence writing activity and there is a space between retrievals, this activity can get higher overall score. Thus, the analysis scores could vary from different raters because of the design. However, most studies related to TFA (see Table 2.6) did not clarify how each activity was designed or modified. While Hu and Nassaji's (2016) study tended to rely on the analysis scores evaluated by

Nation and Webb (2011, pp. 318-319), other studies (e.g., Kamali et al., 2020; Khoshsima & Eskandari, 2017; Chaharlang & Farvardin, 2018; Gohar et al., 2018; Zou & Xie, 2018) did not provide much information about how each activity was designed to use with the learners. The lack of a lesson plan or explicit description of each activity raises doubts about the validity of self-evaluation. This is because some variables such as Motivation in this framework need a careful analysis, but some terms in the framework are not clearly defined. For example, it is not clear from the framework that the term motivation could only mean extrinsic motivation, or refer to both extrinsic and intrinsic motivation (see details in Section 2.3.3.1). Motivation is a complex and sensitive term that needs a precise description. Although vocabulary activities tend to involve extrinsic motivation, some raters might interpret the term as both intrinsic and extrinsic motivation when evaluating vocabulary activities, and this could lead to misleading results from raters' different point of view.

Furthermore, some terms such as 'receptive retrieval' in the Retrieval component and 'instantiation' and 'interference' in the Retention component tend to require specific knowledge in the fields of vocabulary and memory. It is questionable whether the use of this framework can be applicable to all English teachers in general. It seems to me that teachers or researchers who are not familiar with these terms might face with difficulty in rating. It could be time-consuming to review all the terms from the other resources or books before doing an activity analysis. Therefore, in this study, the terms that tend to be too board are briefly described in Table 2.7. This is to ensure that different raters will have the same understanding of these terms before doing the analysis in order to avoid bias from self-evaluation.

*Table 2.6. TFA analysis of vocabulary activities from previous studies*

| | Word cards | Gap-filling/ cloze exercise | Keyword techniques | Sentence or summary writing | Rewording sentences | Multiple-choice | Reading with glosses/ definitions | Oral Reproduction |
|---|---|---|---|---|---|---|---|---|
| Nation & Webb (2011) | **11** | **8** | **7** | **-** | **6** | **6** | **5** | **-** |
| Hu & Nassaji (2016) | **-** | **7** | **-** | **-** | **6** | **6** | **-** | **-** |
| Nakata & Webb (2016) | **12** | **9** | **-** | **7** | **-** | **-** | **-** | **-** |
| Khoshsima & Eskandari (2017) | **-** | **7** | **-** | **-** | **6** | **6** | **6** | **-** |
| Chaharlang &Farvardin (2018) | **11** | **-** | **8** | **-** | **-** | **-** | **5** | **-** |
| Zou & Xie (2018) | **-** | **7** | **-** | **9** | **-** | **-** | **-** | **-** |
| Gohar et al., (2018) | **-** | **-** | **-** | **7** | **-** | **-** | **-** | **-** |
| Kamali et al., (2020) | **-** | **-** | **-** | **11** | **-** | **-** | **-** | **11** |

*Note: Total TFA score = 18 points*

***Table 2.7. Description of terms in TFA framework adapted from Nation and Webb (2011; 2017) and Webb (2013) and applying to the current study***

| Terms | Description |
|---|---|
| ***Motivation*** | Extrinsic motivation that can be stimulated by external stimulus such as incentives or enjoyable/challenged activities |
| ***Vocabulary learning goal*** | The explicit purpose of the vocabulary learning/activity |
| ***Noticing*** | Paying deliberate attention to the target learning vocabulary |
| ***Awareness*** | A deliberate focus on learning a new set of vocabulary |
| ***Negotiation*** | The situation when learners discuss and clarify features of vocabulary such as its meaning or spelling to each other |
| ***Retrieval*** | The condition when learners recognize or recall vocabulary they have encountered from their memory |
| ***Recall*** | The condition when learners retrieve the meaning or form of vocabulary from their memory without seeing or hearing any choices, or clues |
| ***Receptive retrieval*** | The condition when learners retrieve the meaning of vocabulary from their memory |
| ***Productive retrieval*** | The condition when learners retrieve the form of vocabulary from their memory |
| ***Spacing between retrieval*** | A certain period of time between first and second/next retrievals of the same target vocabulary |
| ***Generative use*** | The condition of meeting the target learning vocabulary used in a new different way from the first encounter/meet |
| ***Productive generative use*** | The condition of using/writing the target learning vocabulary in a new different way that has not met before |

***Table 2.7. Description of terms in TFA framework adapted from Nation and Webb (2011; 2017) and Webb (2013) and applying to the current study (conts.)***

| Terms | Description |
|---|---|
| ***Retention*** | The ability to keep, remember or continue having information about a word |
| ***(Word) form*** | A written word that is spelled in a second language (L2) |
| ***(Word) meaning*** | An expression/translation of an L2 word form |
| ***Form-meaning link*** | A connection between L2 form and L1 meaning for comprehension of a word |
| ***Instantiation*** | An instance of a word used in a meaningful situation |
| ***Imagination*** | A vision that learners deliberately see/imagine to link a visual image to the meaning of the word |
| ***Interference*** | A negative effect/transfer of learning unknown L2 words with L2 related words (near synonyms, opposites, and cognates, etc.) at the same time |
| ***Motivation*** | Extrinsic motivation that can be stimulated by external stimulus such as incentives or enjoyable/challenged activities |

When compared to the Involvement Load Hypothesis, the TFA tends to involve more components leading to learning. For this reason, I had an attempt to apply the framework to use in the current study. *Motivation, Noticing, Retrieval, and Generative Use* (also called varied encounters and varied use by Nation & Webb, 2017) are recognised as influential components supporting language learning as mentioned earlier in Section 2.3. Therefore, these four TFA components: *Motivation*, *Noticing*, *Retrieval*, and *Generative Use* were the focus of the investigation. Including these criteria to the framework seems to rely on the previous findings of studies related vocabulary learning. Unlike other TFA components, *Retention* in this framework seems to be a broad term that includes various conditions/factors that are not being linked by theory. It should not be defined as a single learning variable promoting vocabulary learning. Because I tried to control the factors that can be used as variables, as suggested by the vocabulary learning theories and notions, to explore the effects of each TFA component that may lead to different degree of vocabulary learning, the *Retention* component in the TFA framework is excluded from this study due to the unconnected theoretical relationship between questions within the *Retention* component.

While the TFA seems to be useful for the present study, I faced with difficulties in using the framework because it seems that some terms from the four components are not defined precisely. The terms such as motivation and receptive/productive retrieval require background knowledge in the fields of motivation and memory as discussed earlier. Because of this, I realise the importance of a training for raters invited to analyse vocabulary activities (see more details in Section 3.4). Researchers (Lumley & McNamara, 1995; Johnson et al., 2008) suggested that rater training can give more consistent results and higher rates of agreement. Besides, the framework involves vocabulary learning components that may facilitate word retention.

## 2.4 Present study

Most TFA research does not show whether the presence of the individual components of Technique Feature Analysis (TFA), *Motivation, Noticing, Retrieval, and Generative Use* in instructional materials that supports both short-term and long-term vocabulary retention in written production. There was only a recent study from Kamali et al. (2020) that attempted to explore short-term and long-term productive word knowledge through oral and written tasks. However, they did not pay attention to the effects of each TFA component on productive word gains. I did not include the *Retention* component as a component leading to vocabulary retention in the present study although it may affect vocabulary learning. That is mainly because *Retention* suggested in this framework comprises more than one variable which seems to be ambiguous to identify as a single component in this study. Due to the limited number of research on TFA components, I would like to explore the effects of the framework and its components on retention of single words. The TFA framework was employed to evaluate supplementary vocabulary learning materials focusing on different TFA components. This study does not focus on MUW due to the complexity in defining MUW. Moon (1997, p.43) maintains that "There are many different forms of multi-word item, and the fields of lexicology and idiomatology have generated an unruly collocation of names for them, with confusing results." Since MWUs are defined differently based on different views of scholars, a common definition has not been generally agreed. Adding MWUs to the current study may cause excessive complexity in terms of research design which can lead to confusion in reporting data and delay in analysing the results due to time limitation in conducting this study. I am aware that a single study might not be able to fill all research gaps and not focusing on MWU is a limitation of the study.

This study focuses on knowledge of form recall. According to the existing research, little has been known whether TFA components can similarly promote retention of form recall. I had an attempt to fill the gap by comparing the effects of each TFA component on vocabulary learning and retention. Word retention was explored in terms of form recall through a controlled productive knowledge test (see details in Section 3.5.3.3) to investigate controlled productive knowledge.

## 2.5 Research questions

There are two main research questions that the present study attempts to answer:

1.) Do activities with high TFA scores result in better retention of single words in productive form recall?

   a. Do the TFA-supported groups result in better short-term retention than the Control group?

   b. Do the TFA-supported groups result in better long-term retention than the Control group?

2.) What is the individual role of the different TFA components on promoting vocabulary retention through vocabulary tasks?

   a. To what extent does *Motivation* support (short-/long-term) retention of controlled productive knowledge?

   b. To what extent does *Noticing* support (short-/long-term) retention of controlled productive knowledge?

   c. To what extent does *Retrieval* support (short-/long-term) retention of controlled productive knowledge?

d. To what extent does *Generative Use* support (short-/long-term) retention of controlled productive knowledge?

The value of the answers to these questions would be added to fill the research gap and bring contributions to the field of vocabulary learning as mentioned in Section 1.3 and Section 2.4.

# Chapter 3
# Research Methodology and Pilot Study

This section identifies how best to address the research questions. In order to do this, I first clarify the context, participants and research design in Section 3.1, Section 3.2 and Section 3.3, respectively. The research methods also cover the instruments: vocabulary activities and tests used in the experiment (see Section 3.4, Section 3.5 and Section 3.6) . This chapter also include the analysis and results of the Pilot Study and how the findings of the Pilot Study inform the design of the Main Study in Section 3.7 and Section 3.8. Then, I discuss the analytic method and tool employed to use in the current study in Section 3.9. The Main Study analyses and results will be presented later in Chapter 4 and Chapter 5, respectively.

## 3.1 The context of the study

The current study was conducted with participants at Thammasat University, which is one of the leading public universities in Thailand. It is the university where I work as a full-time lecturer. Similar to all public universities in Thailand, the language commonly used in classroom instructions is first language (L1), Thai. In English classes, the language of instruction is mainly English; however, L1 is used when the students could not understand the assigned tasks and need more clarification. English is taught as a foreign language (EFL) according to Kachru’s (1985) model of world Englishes because it is recognised as the language of instruction only in English classrooms, not a medium of communication in students’ daily life.

## 3.2 Participants

The participants in both the Pilot Study and Main Study were Thai undergraduate students (aged between 18 and 20 years old) taking the University English II course which is a compulsory English foundation course offered by the Language Institute of Thammasat University (LITU). More details of the sample size of both the Pilot Study and Main Study will be provided in Section 3.5.1 and Section 4.1, respectively. In every academic year, the students taking this course are randomly arranged into groups/sections at the beginning of each semester. They have got English test scores from the Ordinary National Educational Test (O-NET) between 50 to 75 (out of 100), or band 4 to 5 of the IELTS. When compared with the Common European Framework of Reference (CEFR), these students are between basic users (A2) and independent users (B1). Their language levels is described as intermediate based on the English (O-NET) score they received from a national test (see more details on the participants' prior knowledge on vocabulary in Section 3.5.2.1 in the Pilot Study and Section 4.4.1 and Section 5.1 in the Main Study). So, I will use the term *intermediate level students* to identify these English learners throughout the study. In terms of characteristics, the participants' language of mother tongue is Thai. They are all non-English major students. Those who were studying in English majors or international programmes were excluded from this study due to their advance level of English proficiency.

## 3.3 Research design

The Main Study aims to produce findings that can inform classroom practice so the ecological validity of the study design is important and this led me to a quasi-experimental research design similar to that adopted in Hulstijn and Laufer (2001), Keating (2008) and Hu and Nassaji (2016) to explore vocabulary growth and retention by using naturally occurring groups. In particular, a quasi-experimental research design would be conducted with 6 groups: a control group and five experimental groups. The Control group (Group 1) would not receive a designed vocabulary activity for learning. Learning of this group is based on assigned tasks in the Unit 1 of the coursebook in which no vocabulary activity is provided (see Section 3.4.1 for detail). Meanwhile, each of the experimental group would receive an activity with TFA scores higher than the other groups in one of the following components: *Motiving* (Group 2), *Noticing* (Group 3), *Retrieval* (Group 4), *Generative Use* (Group 5), or an activity with higher TFA scores in all components (Group 6). Before the treatment, the participants would complete a Pre-test to measure their knowledge of the target items before the treatment. Immediate after the treatment, they would complete an immediate post-test to measure their short-term retention after the treatment. Two weeks later, they would complete a delayed post-test to measure their long-term retention after the treatment. The pre-test and post-tests would measure controlled productive vocabulary knowledge in terms of form and meaning recall. Controlled productive vocabulary knowledge of form and meaning recall was chosen because the designed test was constructed to control for the target words in the unit of study (see more details in Section 3.5.2 and Section 3.5.3). Table 3.1 shows the research questions and design of the Main Study. To find the answer to RQ1, the pre-test and post-tests scores of the Control group would be compared with those of each of the experimental group. The independent variables are the groups with different TFA activities and scores. The dependent variables are vocabulary gains and retention scores from pre- and post-treatment vocabulary tests. To find the answer to RQ2, the pre-test and post-tests scores of the experimental groups would be compared with each other. The independent variables are the four components of the TFA framework: *Motivation, Noticing, Retrieval,* and *Generative Use.* The dependent variables are vocabulary gains and retention scores from pre- and post-treatment vocabulary tests.

***Table 3.1. The research questions with design***

| Research Questions | Design |
|---|---|
| **RQ1.** Do activities with high TFA scores result in better retention of single words in productive form recall? | The comparison between a Pre-test and Posttests (Immediate and Delayed) scores of the  Control group (Group 1) with those of each of the experimental group in turn: Group 2 (TFA=7), Group 3 (TFA=7), Group 4 (TFA=8), Group 5 (TFA=9), and Group 6 (TFA=15). |
| **RQ2.** What is the individual role of the different TFA components on promoting vocabulary retention through vocabulary tasks? | The comparison between a Pre-test and Posttests (Immediate and Delayed) scores among the experimental groups with high scores in each of the FTA components: Group 2 (high in *Motivation*), Group 3 (high in *Noticing*), Group 4 (high in *Retrieval*), and Group 5 (high in *Generative Use*). |

It is important to be noted that, although *Retention* is excluded from being a variable for the main analysis (see Section 2.3.3.5 for the rationale), it was controlled for in order to explore the effects of each independent variable, except for Group 6 that all components are rated as high (see Table 3.2). *Retention* was excluded because it involves many factors in its component that are not being linked by theory and so can lead to misleading results in investigating the effects of each learning component on word retention. This issue is addressed in Section 3.4 below.

## 3.4 Vocabulary learning activities for the treatment

As mentioned in Section 3.3 (research design), there were five experimental groups and one control group in the Main Study. Each group had to study using a different vocabulary activity. The process of designing these activities included three main steps: (a) choosing a unit from the participants' textbooks to develop the material for the treatment, (b) modifying the activities in this units using the TFA, and (c) rating the TFA score of each activity. In this section, I will describe each step in detail.

### 3.4.1 Choosing the unit in the participants' textbook

As the data collection for the main study would be conducted in Term 1 at Thammasat University, the main book used in the English II course at this university was selected to develop materials in the treatment of the Main study. This is an in-house book developed by both native and non-native English speaking teachers who have at least 5-year teaching experience in the field of English language learning. The book had been validated and used for at least two years prior to the experiment of this study. There are eight units designed to promote communicative competence. Each unit has the same structure, including reading, grammar and writing skills. In the reading part, vocabulary tasks were designed to facilitate either incidental learning or intentional learning. All vocabulary tasks included in the eight units were analyzed by myself using the TFA framework for the purpose of unit selection. Results of the analysis indicated that Unit 1, Unit 4, Unit 5 and Unit 8 had lower TFA scores than the other units. These units accounted for 0 (Unit 1), 6 (Unit 4), 8 (Unit 5), and 8 (Unit 8) out of 18 TFA scores (see the detailed scores in Table 3.2).

*Table 3.2. Analysis of eight units based on the TFA criteria*

| Component | Criteria | Unit | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| ***Motivation*** | Is there a clear vocabulary learning goal? | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Does the activity motivate learning? | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Do the learners select the words? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***Noticing*** | Does the activity focus attention on the target words? | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Does the activity raise awareness of new vocabulary learning? | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Does the activity involve negotiation? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Is it productive retrieval? | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| | Is it recall? | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| | Are there multiple retrievals of each word? | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| | Is there spacing between retrieval? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***Generative Use*** | Does the activity involve Generative Use? | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| | Is it productive? | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| | Is there a marked change that involves the use of other words? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| | Does the activity involve instantiation? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Does the activity involve imagination? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Does the activity avoid interference? | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| **Total score (18)** | | 0 | 9 | 11 | 6 | 8 | 9 | 10 | 8 |

In addition to analysing the TFA scores of each unit, I also examined if any unit had no vocabulary learning activities. It is because I aimed to include one Control group to the experiment to compare with the experimental groups with different designed activities. The results showed that only Unit 1 and 4 did not have any vocabulary learning activities. Unit 1 had lower TFA score for all components, which means that I could design  vocabulary activities for this unit to explore the effects of the framework on *Motivation, Noticing, Retrieval,* and *Generative Use* on vocabulary retention. Additionally, choosing the unit at the beginning of the programme helped to minimise the learning coming from other sources as students could make progress further in the course. In brief, Unit 1 was chosen to develop activities for the treatment.

### 3.4.2 Modifying the activities in this units using the TFA

Once I had chosen the unit to develop vocabulary learning actvities for the treatment, the second step is to use the TFA to modify the activities for learning in this unit. I started with reviewing the vocabulary activities suggested by Nation and Webb (2001) and employed to use in previous studies related to TFA (see Section 2.3 for details of these studies). As explained in the literature review (see also Section 2.3.3 and Section 2.3.5), the TFA framework created by Nation and Webb (2011; 2017) has been used in several experimental studies (e.g., Hu & Nassaji, 2016; Gohar et al., 2018; Zou & Xie, 2018). It is the main framework utilised to evaluate materials for teaching vocabulary with different groups of participants. The five sets of materials for the five experiment in Table 3.3 (also see Appendix 1) were developed systematically through three main stages: 1) Review, 2) Design, and 3) Evaluation. In this section, I provided details of Stage 1 (Review) and Stage 2 (Design) while in the following section (Section 3.4.3) I described how these activities are evaluated by raters to ensure the process of materials development.

***Stage 1: Review***

Reviewing vocabulary learning materials that are expected to be useful and match the purpose of this study was the initial step in developing materials for the main study. Information regarding vocabulary materials and activities from books (e.g., Nation & Webb, 2011; Nation, 2013; Webb & Nation, 2017), journals (e.g., Folse, 2008; Nakata & Webb, 2016) and empirical studies (e.g., Hulstijn & Laufer, 2001; Laufer, 2003; Kim, 2008; 2011; Hu & Nassaji, 2016; Gohar et al., 2018; Zou & Xie, 2018) had been studied in the stage of reviewing.

In this stage, I first reviewed activities that might give support for *Motivation, Noticing, Retrieval,* and *Generative Use* based on the TFA evaluation of Nation and Webb (2011). I looked for the activities that tend to have high support for these four components from the evaluation guideline table suggested in Nation and Webb' s (2011, pp. 318-319) book. Then, I selected five activities for the experimental groups (Group 2 to Group 6) as presented in Table 3.3 by using the eighteen checklist of the TFA as criteria. The selected tasks were reading plus fill-in (for Group 2: *Motivation*), glosses (for Group 3: *Noticing*), word parts (for Group 4: *Retrieval*), rewording sentences (for Group 5: *Generative Use*), and word cards and writing (for Group 6: *All TFA Components*) (see more detail in the Stage 2 and Stage 3 below). Table 3.3 below shows the first five activities that were selected as materials to match the purpose of this investigation.

***Table 3.3. Vocabulary activities used in five experimental groups***

| **Group 2**<br>*Motivation* | **Group 3**<br>*Noticing* | **Group 4**<br>*Retrieval* | **Group 5**<br>*Generative Use* | **Group 6**<br>*All TFA Components* |
|---|---|---|---|---|
| Reading plus fill-in | Reading and using glosses | Reading and identifying word parts | Reading, and rewording sentences | Word cards and writing |

The Design stage below gives information related to self-rating and modification of the selected activities by using the TFA framework.

### *Stage 2: Design*

In this stage, the TFA criteria were also used to control the amount of support they provide for *Motivation, Noticing, Retrieval*, and *Generative Use*. During the first self-rating, I realised some issues that vocabulary activities should be modified to match the research aim. I found that the amount of support from some components such as *Noticing* could affect other components in some groups. For example, rewording sentence, which was initially selected for Group 5 focusing on *Generative Use* component, also had a high score on the Noticing component if the target words in the original sentence are highlighted which would be needed to ensure that the participants would use the target words in a new sentence if I did not make them noticeable to them.

This made it difficult to examine if the rewording sentence activity led to word retention solely because of *Generative Use*. I then tried to change the design of the activities and reduce the impact of other components that might affect the determined component by modifying these tasks.

I used the TFA criteria again to modify the selected activities and analyse all activities for the six groups, including the Control group. These include an activity with low TFA scores on all components, an activity with high TFA scores on *Motivation*, *Noticing*, *Retrieval* and *Generative Use*, and an activity with high TFA scores on all components (see Table 3.6). Activities were modified by adding or changing tasks to match the purpose of learning before creating supplementary materials (see Appendix 3) for each group.

To control for the impact of *Noticing* in the *Generative Use* group (Group 5), I changed the rewording sentences task to focus on word parts with writing. I also redesigned the activity of the *Retrieval* group (Group 4) to avoid word formation that might give high support for *Generative Use*. The *Retrieval* group has to identify parts of words rather than writing word parts. Table 3.4 below shows the new design activities that were selected as materials to match the purpose of this investigation.

***Table 3.4. Vocabulary activities used in five experimental groups***

| **Group 2**<br>*Motivation* | **Group 3**<br>*Noticing* | **Group 4**<br>*Retrieval* | **Group 5**<br>*Generative Use* | **Group 6**<br>*All TFA Components* |
|---|---|---|---|---|
| Reading plus fill-in | Reading and using glosses | Reading and identifying word parts | Reading, word parts, and sentences writing | Word cards and writing using target words |

After that, I wrote descriptions of each activity and lesson plans (see also Appendix 2) for each group to make it more explicit to facilitate evaluation. Details of the description were given in the following:

In terms of practicality, each intervention group read the same reading passage containing the ten target words (see details and process of vocabulary selection in Section 3.5 and Section 3.7.2). These words, which comprise *concern* (K1)*, issued* (K1)*, shortage* (K1)*, encouraged*(K2)*, represent* (K2)*, strengthen* (K2)*, request* (K3)*, portrayed* (K4)*, venders* (K4), and *decrees* (K5)*,* are not highlighted in the Unit 1 of the book, as will the Control group. However, the participants in the intervention groups were given different supplementary handouts developed to provide support for different TFA components. All the six groups were treated the same except for the adaptations to the proposed materials.

I taught all groups the same contents, including grammar (forming questions), reading, and vocabulary, but the order of the activities (see lesson plans in Appendix 2) as well as the provision of supplementary materials (see Appendix 3) were different due to the appropriateness of doing each activity. Time on vocabulary task of all groups depended on the design of each activity, but it did not exceed 60 minutes for three reasons. First, some tasks such as matching and using glosses tended to require less time than the other tasks like writing. Also, to give high support for *Retrieval*, as suggested by the TFA, the activity for Group 4 was designed to provide space between the first and next encounter of the words. Finally, the participants also needed to learn other content (grammar and reading skills) in this three-hour class and do an Immediate Posttest that took approximately one and a half hours.

**Group 1** (reading only) is the Control group because learning in this group was based on the contents and sequence in the coursebook. There was no vocabulary activity used with this group, and the participants were only required to read a reading passage that contains the target learning words without the provision of a vocabulary activity.

**Group 2** (reading plus fill-in) worked with an additional activity which was highly rated in terms of support for *Motivation*. The reading passage was taken from the one used with the control group, but it was modified to support *Motivation*. Also, supplementary handouts and a Kahoot game were used with this group to motivate learning and to control for *Motivation*. Game-like features as suggested by previous empirical studies (e.g., Bailey et al., 1999; Amiri & Salehi, 2017) was therefore implemented with this group. In authentic classroom learning, we might not be able to expect the same results from learners with regard to their differences. Degrees of learning vary among students with different learning styles and aptitudes, and not everyone will be motivated to the same degree by learning through the same proposed material even though the material is designed to be challenging by using game-like features. However, empirical evidence (e.g., Bailey et al., 1999; Amiri & Salehi, 2017) on the effects of games towards motivation can confirm that learners are likely to be motivated by integrating games with learning. Not all of the participants may be motivated by the same learning materials when practising in class, but I may be able to control for *Motivation* through the suggested criteria of the TFA framework. So, this activity was analysed within the TFA framework to evaluate its capacity to motivate learners by three experienced teachers (see details in the Stage 3: Evaluation below). To confirm that this is a challenging activity which can promote motivation, I also included a questionnaire questions about enjoyment in learning after the participants finish doing the activity at the end of each class (see details of the questionnaire in Section 4.4.3 below). This allowed me to investigate whether *Motivation* within the framework promoted word retention effectively.

The participants in Group 2 were asked to work in pairs to read different versions of an incomplete passage: A and B (version A for student A to ask student B and vice versa). Each student had half of the target words and one extra word taken from the reading passage included in the target sentences in their version. Before reading, they had to learn about how to form questions which is the main grammatical aspect to be learned in this unit. The participants had to form questions to find out the missing information in their passage with a partner and fill the words in the blanks to complete the reading passage. This activity may focus attention on the missing words but does not raise awareness of new vocabulary learning. To avoid support for *Noticing*, I neither introduced the target words in class nor provided a separate set of the target words. Instead, they were told that they are learning how to form questions to complete the sentences in their reading version in order to comprehend the passage and win the game. That is because I tried to avoid informing them that vocabulary is the focus of learning. Although the target words were embedded in the sentences that the participants had to form questions, they were not be highlighted. One extra word was also added to each version of the passage in an attempt to avoid noticing on the ten target words. To go over the answers, I used a Kahoot game by asking each individual to match the twelve words (ten target words and two extra words) with the correct sentences taken from the passage. Also, I asked several teachers who are currently teaching English to Thai students about games that are likely to encourage their students to learn and feel enjoyable in the learning. Most of them agreed that their students tend to be obviously excited by and take pleasure in playing Kahoot games. Another good point of this game is that it requires less time for the teachers to prepare than some other games. So, I decided to include a Kahoot game to use with this group to motivate learning. The winner having the highest point from this matching game got a reward to increase extrinsic motivation in learning. This way, the activity was designed to be challenging and enjoyable so that it supports motivation in learning (Nation and Webb, 2013), but does not raise the learners' awareness of new vocabulary learning as the focus is on reading comprehension and grammar which is forming questions.

**Group 3** is the *Noticing* group (reading with glosses). The participants in this group were given an additional material which is a handout containing L2 definitions with L2 synonyms of the target words before reading the passage in order to make sure that they could be able to comprehend the meanings. Giving them a set of vocabulary can make the target learning vocabulary easily noticeable and raise awareness of vocabulary learning. However, this activity can also be rated as high in *Retention* if L1 definitions are given to ensure linking of form and meaning. Studies found that words that are highly related in meanings lead to negative interference effects (Waring, 1997a; Erten & Tekin, 2008). Using L2 synonyms or antonyms may lead to negative interference; therefore, *Retention* score which seems to be high if using L1 translation could be controlled. In class, they were asked to take a look at a list of ten target words with L2 definitions in a supplementary handout and be reminded that this is a list of new target vocabulary that can help them comprehend the reading passage of Unit 1. Then, they had to find these words in the reading passage and work in pairs by discussing and matching a correct synonym with each target word. As discussion might promote negotiation to some extent, each pair had to discuss in L2 in order to match the provided L2 synonyms with the target words that L2 definitions have been initially given.

This way, I could be able to observe the effects of the TFA framework specifically on *Noticing* with only little potential effects on other components such as *Retention* in this group.

For **Group 4**, the focus is on *Retrieval* (reading and identifying word parts). Before reading, the participants had to learn vocabulary through a word parts table in a supplementary handout designed to focus on part of speech (POS) and suffixes because the passage contains a large number of words with suffixes and only a small number of words with prefixes. Also, suffixes are related to the part of speech which is another important aspect of vocabulary comprehension. The handout contained a word list of the target words and six extra words found in the passage in order to minimize intentional noticing. The rationale for adding six extra words was 1) to avoid the participants noticing the target words and 2) because previous studies (e.g., Hu & Nassaji, 2016; Chaharlang and Farvardin, 2018; Gohar et al., 2018; Zou & Xie, 2018) usually included between 10 and 16 words to their experiments as mentioned in the Literature Review (see Table 2.5 in Section 2.3.5). Suffixes were given to all items in the handout. The participants in Group 4 had to identify POS and suffixes of the given items in the handout after finish teaching. Rather than continue focusing on vocabulary or reading, the participants had to practise other skills such as reading as required in the course outline before doing two more vocabulary activities that aim to promote repetition. At the end of the class, I asked the participants to recall these items they have learned by putting a mark on the word they have just encountered from the list that includes both the words they have learned and new words appearing in the passage in an additional exercise sheet to avoid productive use. A list of words were shown quickly (2 seconds, each) on a computer screen. There were both the list of words they have encountered and a new list of words taken from the passage. Lastly, the participants had to recall the items by saying them out loud without looking at the handout. The activity tended to allow spacing between retrieval and support multiple retrievals from several repetitions which may lead to long-term retention.

The participants in **Group 5** had to do an activity on reading, word parts and sentence writing, focusing on *Generative Use*. Although this activity involves word part learning which is similar to the activity used in Group 4, the process of learning was different. Firstly, I taught the participants in Group 5 about the part of speech (POS) and suffixes by using the same word list and a similar supplementary handout to that used with the Group 4. However, after studying the handout they had to look for words with suffixes in the reading passage and write parts of the words in the provided table in the handout. This can be productive if I give the students an opportunity to produce the words by their own using the rules of adding suffixes. The activity may encourage productive generative use in terms of word formation since only one form of the words is presented in the reading passage. After that, they were encouraged to read the reading passage in the unit. Studying word parts could help them guess the meaning of unknown words in the passage. As they had to learn about forming questions in this unit, they were then asked to form questions by using the words from the word list.

The last group, **Group 6**, had to study vocabulary with word cards which receives a high score in all TFA components, except *Generative Use*. For this reason, I also implemented a sentence writing activity with this group in order to increase the potential of receptive and productive generative use because writing using target words involves full (3) scores in Generative Use. As L1 definition can help to avoid negative interference, it was used to explain the words shown in the cards. The participants had to learn the target words from word cards in order to help them understand the reading passage. However, they had to make sentences using the target words before reading. This is to give them a chance to use the word productively. The purpose of using this activity was to explore whether an activity with high scores in all TFA components leads to better vocabulary retention in written production than the other activities receiving lower scores. I might be able to compare the results of learning between this group and the controlled group as well as the other four experimental groups that have lower TFA scores by using these two activities together. Hence, the effectiveness of the TFA together with its effects on *Motivation, Noticing, Retrieval,* and *Generative Use* towards learning and remembering form of words can be explored.

Then, all these activities with supplementary materials were evaluated by using the TFA framework. Self-evaluation could lead to bias. As suggested by several researchers (e.g., Lumley & McNamara, 1995; Johnson et. al., 2008), I decided to invite three teachers (see discussion in Section 3.4.3) who have at least five-year teaching experience in English to do the evaluation in this study. Information regarding this evaluation is provided in Stage 3 (Evaluation) in Section 3.4.3 below.

### 3.4.3 Rating the TFA score of each activity

After I had modified the activities for the treatment and self rated them, to avoid the bias toward my subjective judgements, I then asked three raters to independently rated these activities against the TFA. These three raters had at least five-year teaching experience in the field of language teaching and learning. Following the scoring method suggested by Nation and Webb (2011), one point was given to each TFA criterion (see detail in Section 2.3.3). In the current study, however, three-rater system was used for the purpose of reliability. At least two of them had to say an activity meets the criterion in order to get one point.

The rationale for including three raters was adopted from researchers in the field of measurement (e.g., Tinsley & Weiss, 2000; Stemler & Tsai, 2008) and other fields such as medical diagnosis (e.g., Doktor et al., 2020), memory (e.g., Ratiu & Azuma, 2015) and education (e.g., Caldwell & Moore, 1991; Hall & Sheyholislaami, 2013). Previous research (Lumley & McNamara, 1995; Johnson et. al., 2008) has suggested that multiple raters lead to consistent results. When there is a marked gap or disagreement between three raters, I could rely on the two raters who have the same agreement. Having only two raters can lead to problems in the evaluation and evaluation administration. For instance, if there was a disagreement between two raters, I would need to replace one rater with the third rater. This process could take time. Thus, three raters suggested for high-stakes decisions or testing (Stemler & Tsai, 2008) would give a more accurate result to ensure the reliability of activities used in this study and help to eliminate a disagreement gap that may occur from two raters.

For each activity, if the three raters agree, the TFA criterion gets one point. However, if two among three raters disagree on any criterion, all three raters will be asked to provide reasons and invited to join a training to improve the quality of evaluation. Although agreement among the three raters is expected, this study allows for decisions where only two raters agree. At least two from the three raters need to agree for that criterion to get one point. The results from the raters were then analysed again to interpret whether the components have high or low potential. Any component which is equal to or above the threshold level (two out of three or approximately 66.68% and above) was considered as having high potential in this study. This two out of three threshold is used in test development such as the Index of Item Objective Congruence (or IOC), proposed by Rovinelli and Hambleton (1977). They also applied two out of three as the acceptable norm for evaluation. Thus, any TFA criterion that is below the threshold level was not considered as high in the current study. For instance, Reading Plus Fill-In activity (see Table 3.7 below) received a low (2 out of 4) score in *Retention* because it is below the level of satisfaction (lower than 66.68%). In contrast, it got 2 (out of 3) points, approximately 66.68%, in *Motivation*, implying that this activity may lead to high motivation. To ensure that they raters are understand the criteria, before they completed their tasks, I delivered some training to them by inviting them to practice on the evaluation. I first described them the definition of terms in the TFA framework before demonstrating how to evaluate a sample vocabulary activity. Then, I asked them to select one activity to practice and discuss their points. Finally, I discussed the issues that have arisen from the evaluation to remind them about the aim of the current project. This way, they could have the same understanding on the terms and evaluation purpose. I explain in detail below the process of evaluating the learning materials by the three raters.

### ***Stage 3: Evaluation***

In this stage, I invited three experienced teachers who have been teaching English for more than 5 years to evaluate the selected activities. I started by asking the teachers to read through an explanation of the TFA framework which was taken from Nation and Webb (2011, pp. 7-14) and the description of activities before rating them by using the TFA framework. This aimed to develop a common understanding among raters as mentioned in Stage 2.

Table 3.5 below shows the summary of TFA scores from the first attempt at the evaluation by the three raters.

Based on the first evaluation, the teachers evaluated some TFA criteria differently even though they had reviewed the same provided documents. For example, in the activity for Group 5 (*Generative Use*) activity (reading, word parts and sentence writing) where the criteria are *Does the activity involve retrieval of the word?* and *Is it productive retrieval?* the participants have to work in pairs to form questions and provide answers for each question by using the words and so it involves reductive retrieval of the word. However, one of the teacher rated this at zero on this element. To understand their reasons, I had a discussion with the teachers and found that they did not have a clear picture of each activity from the provided description. They thought that the words will be provided and students can have a look at these words in the reading passage while they are forming questions and providing answers to the questions.

***Table 3.5 Summary result of TFA scores from the first evaluation***

| | | *Motivation* | *Noticing* | *Retrieval* | *Generative* | *Retention* | Total |
|---|---|---|---|---|---|---|---|
| Full scores | | (3) | (3) | (5) | *Use* (3) | (4) | 18 |
| 1 Controlled Group (learning is based on Unit 1 of the coursebook) | | | | | | | |
| No activity | G1 | Low (N/A) | Low (N/A) | Low (N/A) | Low (N/A) | Low (N/A) | (N/A) |
| 5 Intervention Groups with various vocabulary activities | | | | | | | |
| Fill in blank | G2 | **High (2)** | Low (1) | Low (2) | Low (0) | Low (2) | 7 |
| Glosses | G3 | Low (1) | **High (3)** | Low (0) | Low (0) | Low (2) | 6 |
| Identifying Word parts | G4 | Low (1) | Low (1) | **High (4)** | Low (1) | Low (0) | 7 |
| Word parts and writing | G5 | Low (1) | Low (1) | Low (0) | **High (3)** | Low (2) | 7 |
| Wordcards and writing | G6 | **High (2)** | **High (2)** | **High (5)** | **High (3)** | Low (2) | 14 |

*Note: each component must receive at least 66.68% of its total score to be regarded as a high potential*

Because they were not certainly know if this activity can stimulate the process of memory retrieval, they made judgement based on their own understanding and background knowledge about the terms in the framework. They believed that it would be more precise if this framework allowed them to give half a score (or 0.5 point) to some criteria. One teacher also raised up a concerning point about the framework that the given information taken from Nation and Webb' (2011) book does not tend to cover all details of terms presented in the framework. She was afraid that some terms such as *Motivation* (*extrinsic* or *intrinsic*), *receptive* or *productive retrieval*, and *instantiation* might be defined differently in different fields of study. So, she had to spend time looking for more information of those terms from other resources such as books or the Internet in order to comprehend the terms to be able to do the evaluation. As a result, it was a time-consuming task for her.

To address this variation, I decided to provide the two teachers with more contextual information by showing them the lesson plans and a list of the description of terms in the framework that tended to be problematic or difficult to interpret (see also Table 2.7 in Section 2.3.5). Then, I had another meeting with them and explained the activities again in detail. I gave them the documents two weeks before asking them to rate the activities by using the TFA framework again. This was aimed to provide the same concept of understanding to the three teachers and help them feel at ease when doing the evaluation.

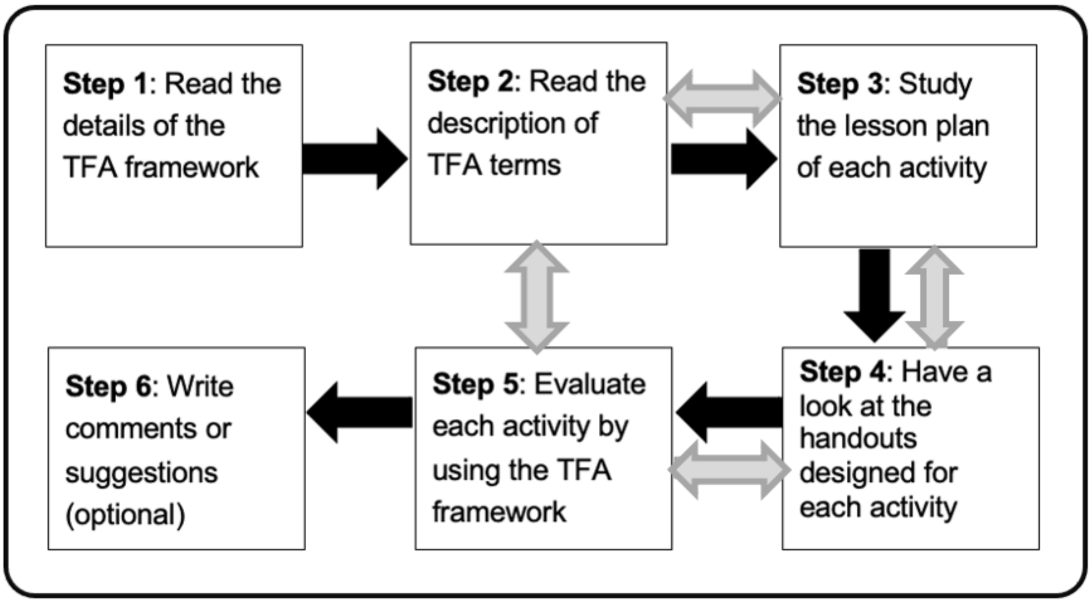

**Step 1**: Read the details of the TFA framework
**Step 2**: Read the description of TFA terms
**Step 3**: Study the lesson plan of each activity
**Step 4**: Have a look at the handouts designed for each activity
**Step 5**: Evaluate each activity by using the TFA framework
**Step 6**: Write comments or suggestions (optional)

**Figure 3.1. Six steps for the activity evaluation**

Also, to control the evaluation process, I asked these teachers to do the evaluation by following the six steps shown in Figure 3.1. Step 6 is optional while steps 1 to 5 are required. They had to follow the black arrows until they reached Step 6 unless they felt unconfident in doing the evaluation or struggled with any terms. If so, they were required to review the description of terms in the TFA framework again starting from Step 2 and follow the gray arrows until they understand all the terms and details of the activities (see description of terms in Table 2.7). The final results of scores from the second evaluation is illustrated in Table 3.6. They tended to be consistent at this time (see each evaluation from Tables 3.7 to 3.11).

***Table 3.6. Six groups with different support on TFA components in the Main Study***

| | | *Motivation* | *Noticing* | *Retrieval* | *Generative* | *Retention* | **Total** |
|---|---|---|---|---|---|---|---|
| **Full score** | | (3) | (3) | (5) | *Use* (3) | (4) | **18** |
| **1 Control Group** *(learning is based on Unit 1 of the coursebook)* | | | | | | | |
| *No vocabulary activity* | **G1** | Low (N/A) | Low (N/A) | Low (N/A) | Low (N/A) | Low (N/A) | (N/A) |
| **5 Experimental Groups** | | | | | | | |
| *Fill in blank* | **G2** | **High (2)** | Low (1) | Low (2) | Low (0) | Low (2) | 7 |
| *Glosses* | **G3** | Low (1) | **High (3)** | Low (2) | Low (0) | Low (1) | 7 |
| *Identifying Word parts* | **G4** | Low (1) | Low (1) | **High (4)** | Low (1) | Low (1) | 8 |
| *Word parts and writing* | **G5** | Low (1) | Low (1) | Low (2) | **High (3)** | Low (2) | 9 |
| *Wordcards and writing* | **G6** | **High (2)** | **High (2)** | **High (5)** | **High (3)** | **High (3)** | 15 |

*Note: The satisfy level of 'high support' of each component is equal to or greater than 66.68%*

Reading plus fill-in activity that will be used with **Group 2 (*Motivation*)** received 2 scores (66.68%) for *Motivation*, 1 score (33.33%) for *Noticing*, 2 scores (40%) for *Retrieval*, 0 score (0%) for *Generative Use*, and 2 scores (50%) for *Retention* (see Table 3.7). *Motivation* is the only component that was rated as high in the Table 3.7 because the satisfy level of 'high support' of each component is equal to or greater than 66.68% as mentioned in Section 3.2.

***Table 3.7. Analysis of a fill-in activity from the raters***

| Component | Criteria | Raters' scores | | | Result |
|---|---|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | 1 | 1 | 1 | **1** |
| | Does the activity motivate learning? | 1 | 1 | 1 | **1** |
| | Do the learners select the words? | 0 | 0 | 0 | **0** |
| ***Noticing*** | Does the activity focus attention on the target words? | 1 | 0 | 1 | **1*** |
| | Does the activity raise awareness of new vocabulary learning? | 0 | 0 | 0 | **0** |
| | Does the activity involve negotiation? | 0 | 0 | 1 | **0*** |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 1 | 1 | 1 | **1** |
| | Is it productive retrieval? | 0 | 0 | 0 | **0** |
| | Is it recall? | 0 | 0 | 0 | **0** |
| | Are there multiple retrievals of each word? | 1 | 1 | 1 | **1** |
| | Is there spacing between retrieval? | 0 | 0 | 0 | **0** |
| ***Generative Use*** | Does the activity involve generative use? | 0 | 0 | 0 | **0** |
| | Is it productive? | 0 | 0 | 0 | **0** |
| | Is there a marked change that involves the use of other words? | 0 | 0 | 0 | **0** |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | 0 | 0 | 0 | **0** |
| | Does the activity involve instantiation? | 1 | 1 | 1 | **1** |
| | Does the activity involve imagination? | 0 | 0 | 0 | **0** |
| | Does the activity avoid interference? | 1 | 1 | 1 | **1** |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*

Next, reading with glosses activity, which will be used with **Group 3 (*Noticing*)**, was rated high in *Noticing*. The activity received 1 score (33%) for

*Motivation*, 3 scores (100%) for *Noticing*, 2 score (40%) for *Retrieval*, 0 score (0%) for *Generative Use*, and 1 score (25%) for *Retention* (see Table 3.8).

***Table 3.8. Analysis of a glosses activity from the raters***

| Component | Criteria | Raters' scores | | | Result |
|---|---|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 1 | 1 | 1 | **1** |
| | Does the activity motivate learning? | 0 | 0 | 0 | **0** |
| | Do the learners select the words? | 0 | 0 | 0 | **0** |
| **Noticing** | Does the activity focus attention on the target words? | 1 | 1 | 1 | **1** |
| | Does the activity raise awareness of new vocabulary learning? | 1 | 1 | 1 | **1** |
| | Does the activity involve negotiation? | 1 | 1 | 1 | **1** |
| **Retrieval** | Does the activity involve retrieval of the word? | 1 | 1 | 0 | **1*** |
| | Is it productive retrieval? | 0 | 0 | 0 | **0** |
| | Is it recall? | 0 | 0 | 0 | **0** |
| | Are there multiple retrievals of each word? | 1 | 1 | 1 | **1** |
| | Is there spacing between retrieval? | 0 | 0 | 0 | **0** |
| **Generative Use** | Does the activity involve generative use? | 0 | 0 | 0 | **0** |
| | Is it productive? | 0 | 0 | 0 | **0** |
| | Is there a marked change that involves the use of other words? | 0 | 0 | 0 | **0** |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 0 | 0 | 0 | **0** |
| | Does the activity involve instantiation? | 1 | 1 | 1 | **1** |
| | Does the activity involve imagination? | 0 | 0 | 0 | **0** |
| | Does the activity avoid interference? | 0 | 0 | 0 | **0** |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*

The reading and identifying word parts activity was designed for the ***Retrieval*** group **(Group 4)**. The activity received 1 score (33%) for *Motivation*, 1 score (33.33%) for *Noticing*, 4 scores (80%) for *Retrieval*, 0 score (0%) from *Generative Use*, and 2 scores (50%) for *Retention* (see Table 3.9).

***Table 3.9. Analysis of reading and identifying word parts from the raters***

| **Component** | **Criteria** | **Raters' scores** | | | **Result** |
|---|---|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 1 | 1 | 1 | **1** |
| | Does the activity motivate learning? | 0 | 0 | 0 | **0** |
| | Do the learners select the words? | 0 | 0 | 0 | **0** |
| **Noticing** | Does the activity focus attention on the target words? | 0 | 0 | 0 | **0** |
| | Does the activity raise awareness of new vocabulary learning? | 1 | 1 | 1 | **1** |
| | Does the activity involve negotiation? | 0 | 0 | 0 | **0** |
| **Retrieval** | Does the activity involve retrieval of the word? | 1 | 1 | 1 | **1** |
| | Is it productive retrieval? | 0 | 0 | 0 | **0** |
| | Is it recall? | 1 | 1 | 1 | **1** |
| | Are there multiple retrievals of each word? | 1 | 1 | 1 | **1** |
| | Is there spacing between retrieval? | 1 | 1 | 1 | **1** |
| **Generative Use** | Does the activity involve generative use? | 0 | 0 | 0 | **0** |
| | Is it productive? | 0 | 0 | 0 | **0** |
| | Is there a marked change that involves the use of other words? | 0 | 0 | 0 | **0** |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 0 | 0 | 0 | **0** |
| | Does the activity involve instantiation? | 1 | 1 | 1 | **1** |
| | Does the activity involve imagination? | 0 | 0 | 0 | **0** |
| | Does the activity avoid interference? | 1 | 1 | 1 | **1** |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*

The activity that gives support to ***Generative Use*** involves reading, word parts and sentence writing. It was designed for the participants in **Group 5**. This activity received 1 score (33.33%) for *Motivation*, 1 score (33.33%) for

*Noticing*, 2 scores (40%) for *Retrieval*, 3 score (100%) from *Generative Use*, and 2 scores (50%) for *Retention* (see Table 3.10).

***Table 3.10. Analysis of word parts and sentence writing from the raters***

| **Component** | **Criteria** | **Raters' scores** | | | **Result** |
|---|---|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 1 | 1 | 1 | **1** |
| | Does the activity motivate learning? | 0 | 0 | 0 | **0** |
| | Do the learners select the words? | 0 | 0 | 0 | **0** |
| **Noticing** | Does the activity focus attention on the target words? | 0 | 0 | 0 | **0** |
| | Does the activity raise awareness of new vocabulary learning? | 1 | 1 | 1 | **1** |
| | Does the activity involve negotiation? | 0 | 0 | 0 | **0** |
| **Retrieval** | Does the activity involve retrieval of the word? | 1 | 1 | 1 | **1** |
| | Is it productive retrieval? | 1 | 1 | 1 | **1** |
| | Is it recall? | 0 | 0 | 0 | **0** |
| | Are there multiple retrievals of each word? | 0 | 0 | 0 | **0** |
| | Is there spacing between retrieval? | 0 | 0 | 0 | **0** |
| **Generative Use** | Does the activity involve generative use? | 1 | 1 | 1 | **1** |
| | Is it productive? | 1 | 1 | 1 | **1** |
| | Is there a marked change that involves the use of other words? | 1 | 1 | 1 | **1** |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 0 | 0 | 0 | **0** |
| | Does the activity involve instantiation? | 0 | 1 | 1 | **1*** |
| | Does the activity involve imagination? | 0 | 0 | 0 | **0** |
| | Does the activity avoid interference? | 1 | 1 | 1 | **1** |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*

Finally, word cards and sentence writing activity with reading that will be used with **Group 6 (*All TFA Components*)** received 2 scores (66.68%) for *Motivation*, 2 scores (66.68%) for *Noticing*, 5 scores (100%) for *Retrieval*, 3 scores (100%) from *Generative Use*, and 3 scores (75%) for *Retention*. All components, including *Retention,* were rated as high (see Table 3.11).

***Table 3.11. Analysis of word cards and sentence writing from the raters***

| Component | Criteria | Raters' scores | | | Result |
|---|---|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 1 | 1 | 1 | **1** |
| | Does the activity motivate learning? | 1 | 1 | 1 | **1** |
| | Do the learners select the words? | 0 | 0 | 0 | **0** |
| **Noticing** | Does the activity focus attention on the target words? | 1 | 1 | 1 | **1** |
| | Does the activity raise awareness of new vocabulary learning? | 1 | 1 | 1 | **1** |
| | Does the activity involve negotiation? | 0 | 0 | 0 | **0** |
| **Retrieval** | Does the activity involve retrieval of the word? | 1 | 1 | 1 | **1** |
| | Is it productive retrieval? | 1 | 1 | 1 | **1** |
| | Is it recall? | 1 | 1 | 1 | **1** |
| | Are there multiple retrievals of each word? | 1 | 1 | 1 | **1** |
| | Is there spacing between retrieval? | 1 | 1 | 1 | **1** |
| **Generative Use** | Does the activity involve generative use? | 1 | 1 | 1 | **1** |
| | Is it productive? | 1 | 1 | 1 | **1** |
| | Is there a marked change that involves the use of other words? | 1 | 1 | 1 | **1** |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 1 | 1 | 1 | **1** |
| | Does the activity involve instantiation? | 1 | 1 | 1 | **1** |
| | Does the activity involve imagination? | 0 | 0 | 0 | **0** |
| | Does the activity avoid interference? | 1 | 1 | 1 | **1** |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*

Nonetheless, I could not be strongly confident that the agreements are not due to chance if I did not do a correlation analysis which is an essential basis for investigating inter-rater reliability analysis of more than two raters.

According to McHugh (2012, p. 282), “If there is likely to be much guessing among the raters, it may make sense to use the kappa statistic, but if raters are well trained and little guessing is likely to exist, the researcher may safely rely on percent agreement to determine interrater reliability.” So, I decided to analyse only the data of activities only for Group 2 (*Motivation*), Group 3 (*Noticing*), and Group 5 (*Generative Use*) by using Spearman’s correlation coefficients because there was no difference in score among the three raters in the result of activities for Group 4: *Retrieval* (see Table 3.9) and Group 6: *All TFA Components* (see Table 3.11). The results from the Spearman’s correlation coefficients of the activity for Group 2 range from 0.791 to 0.892 between the three raters while that of the activities for Group 3 and Group 5 from all pairs was 8.86 and 8.94, respectively at the 0.01 level (2-tailed). If the correlation result is higher than 0.70, it can be interpreted that there was a strong relationship between the raters (Green, 2013). The analysis result of the activity for the *Motivation* group (Group 2) showed a variety of the correlation coefficients when comparing to that of the activities for Group 3 (*Noticing*) and 5 (*Generative Use*). The weakest relationship noticed in between raters 2 and 3 (0.791, $p = 0.01$) can still be interpreted as strong. I ensure the results by using the Fleiss’ Kappa analysis to analyse scores between the three raters. All groups were analyses by this method. Similar to the Spearman’s correlation coefficients, the results of inter-rater reliability by the Fleiss’ Kappa analysis ranged between 0.921 and 1.00 in all groups. This can ensure the strong agreement of scoring between the three raters as the Kappa scores were higher than 8, representing ‘almost perfect’ agreement (Landis and Koch, 1977, p.165).

Despite the results of inter-rater reliability, it seems that there are some criteria under three components of three activities: *Noticing* (See Table 3.8), *Retrieval* (See Table 3.9), and *Generative Use* (See Table 3.10) that were not totally agreed by all raters. Because only one of the three teachers rated those criteria differently, and two-thirds is the threshold level of this study as explained earlier, the differences do not seem to have a major impact on the received scores that were previously calculated. Though, if the raters rely on the same concepts of understanding, it was expected that they should have the same agreements on all eighteen criteria of each activity. Because of this, I decided to invite the three raters to a one-hour online training in aiming to ensure the result of the evaluation. In the training I first asked them to explain the terms in the TFA framework based on their thoughts after reading through the description of terms (as shown in Table 2.7 in Section 2.3.5). After that, I explained the terms that are defined by this study again before giving them the Lesson Plan of Group 2 (*Motivation*) to practise together. I did not show them the result of the previous evaluation, so they did not know the analysis scores of each other. I chose this activity because one of the three teachers gave one point to two criteria: *Does the activity focus attention on the target words?* and *Does the activity involve negotiation?* under the *Noticing* component which can also lead *Noticing* to be considered as high. This could give a different result on the evaluation. During the practice, I worked as a room monitor by asking the teachers to discuss and explain their reasons to support their judgements on each criterion until everyone seems to share the same concepts of understanding on terms and reached a common agreement. I tried to avoid giving my opinions unless the teachers needed more explanation on the activity and terms in the framework. Then, I asked them to also evaluate the activities for Group 3 (*Noticing*) and Group 5 (*Generative Use*) individually because there were one criterion in each activity that one of the teachers rated differently from the others (see Table 3.8 and Table 3.10). Lastly, I asked them to discuss the results with each other. Everyone seemed to have the same agreement based on the discussion. For example, the three raters insisted that *Noticing* in Group 2 (*Motivation*), which was the concerning component mentioned earlier, should receive only one point from the criterion *Does the activity focus attention on the target words?* They also agreed that *Negotiation,* which can occur when students have a chance to discuss language features of vocabulary such as its meaning or spelling, does not tend to be involved in this activity. Based on their experience, it is likely that the students will do everything as directed by the teacher. In this activity, the students will be told not to write down the answers because they will be provided at the end of the activity. Therefore, the raters tended to believe that the students will focus only on asking questions because it seems to be challenging for EFL learners. Because of the training, I can confidently ensure the validity of the results. In the following section, I gave information about the vocabulary test that was used to elicit productive vocabulary knowledge in the present study.

Table 3.12 presents the final TFA scores of each activities based on the three raters' rating. Details of the final materials used in the treatment of the Main study is presented in Appendix 3.

***Table 3.12. Result of TFA scores from the final version of evaluation***

| **Group 2**<br>*Motivation* | **Group 3**<br>*Noticing* | **Group 4**<br>*Retrieval* | **Group 5**<br>*Generative Use* | **Group 6**<br>All four components |
|---|---|---|---|---|
| Reading plus fill in | Reading and using glosses | Reading and identifying word parts | Reading, word parts, and sentences writing | Word cards and writing using target words |
| **TFA = 7** | **TFA = 7** | **TFA = 8** | **TFA = 9** | **TFA = 15** |

*Note: Group 1 is the control group (no vocabulary activity in the Unit of learning)*

## 3.5 Selection of target words

The words for the experiment were selected based on an analysis of a reading passage in the Unit 1 using Cobb's VocabProfiler (2015) from Lextutor. The passage contains 197 word types from the BNC/COCA25000 lists. There are 70%, 20%, and 5% of word types in the most frequent 1,000 (K1), 2,000 (K2), and 3,000 (K3) word-levels, respectively (see Table 3.13). It also contains one- to four-word types from each of the 4,000 (K4), 5,000 (K5), 8,000 (K8), and 12,000 (K12) frequency level which constituted less than or equal to 2%. In order to comprehend 95% of the passage, students need to comprehend words in the three most frequent levels, namely K1, K2, and K3; however, students need to know vocabulary in the K4 and K5 levels at least to reach 98% coverage. This led me to include content words in K1 to K5 frequency levels as what they would need to comprehend the text. Based on the number of vocabulary items covered in each of the other units, Unit 1 should focus on at least twelve items. I first selected the unknown words from K1 to K5 based on the analysis through the Lextutor.

***Table 3.13. Percentage of word types in the analysed reading passage***

| Frequency level | Percentage of word types | Percentage of cumulative coverage |
|---|---|---|
| K1 | 71 | 77.9 |
| K2 | 19 | 90.5 |
| K3 | 5 | 97.1 |
| K4 | 1 | 97.8 |
| K5 | 2 | 99.3 |
| K8 | 1 | 99.5 |
| K12 | 1 | 99.7 |

As students are unlikely to know every word in the high-frequency (K1 to K2) level, words from K3 to lower levels are also unlikely to be known by them. For this reason, I excluded some words such as market (K1), food (K1), dish (K2), rice (K2), and essay (K3) that the students may know. These seem to be the basic words that some of them have often been borrowed for use in the Thai context. For example, Thai people are familiar with the word *market* as it is included in the names of several well-known markets in the country such as Chatuchak Weekend Market. Even though the word *essay* is in the K3 level, I think Thai students know its meaning since it is always included in the instructions of essay writing and we do not usually use the Thai translation for this word. We normally call *essay* when we talk about it. Although I believe that Thai students are quite familiar with these words, the list was rechecked again by two experience teachers in the institution. They also crossed out the same words in the list as they believe that these are the words that the students would know.

Lastly, I looked at the list again to remove such words as influence/influenced and become/becoming that have more than one form presented in the passage. The main purpose of the investigation was on learning related to the stem words, not the ending forms such as -ing or -ed. Also, according to the pilot study results, this is to avoid excessive interaction of the words over the other target words which may affect the results of this study (see Section 3.7 for detail). Then, I had consulted the two invited teachers to eliminate some of the words in the high frequency level as it constituted more than 80% of the initial list of unknown words because there were more than forty unknown words in the list. The remaining was used as the target words for the pilot study (see Table 3.14).

***Table 3.14. The target items for Pilot Study***

| Frequency Level | Target items | Part of Speech | Number of Syllables |
|---|---|---|---|
| **K1** | begin | Verb | 2 |
| | concern | Noun | 2 |
| | government | Noun | 3 |
| | held | Verb | 1 |
| | history | Noun | 3 |
| | important | Adjective | 3 |
| | issued | Verb | 2 |
| | national | Adjective | 2 |
| | protect | Verb | 2 |
| | relationship | Noun | 4 |
| | shortage | Noun | 2 |
| **K2** | creation | Noun | 3 |
| | encouraged | Verb | 3 |
| | identity | Noun | 4 |
| | performances | Noun | 3 |
| | prevent | Verb | 2 |
| | represent | Verb | 3 |
| | replace | Verb | 2 |
| | traditional | Adjective | 4 |
| | strengthen | Verb | 3 |
| **K3** | promote | Verb | 2 |
| | contest | Noun | 2 |
| | colonization | Noun | 5 |
| | responded | Verb | 3 |
| | request | Noun | 2 |
| **K4** | portrayed | Verb | 2 |
| | vendors | Noun | 2 |
| **K5** | decrees | Noun | 2 |

To conclude, there are twenty-eight estimated single words that I attempted to include in the pilot study as presented in Table 3.14. These include eleven words from K1, nine words form K2, five words from K3, two words from K4 , and 1 word from K5 frequency levels. However, I measured the participants' background knowledge of these words again in the Pilot Study (details are given in Section 3.7.1) before identifying the unknown words (see also Section 3.7.2) used as the target learning words in the Main Study. My aim was to introduce and measure words in the forms presented in the reading. For example, if the word in the passage is a plural form, I design the controlled productive test to measure this form of the word (details of the test development is given in Section 3.7.3 and Section 4.4.2).

Based on the Pilot Study (see Section 3.7.1 and Section 3.7.2 for detail of the analysis), ten unknown words were selected as the target words in this study. These words are ***concern*** (K1)*,* ***issued*** (K1)*,* ***shortage*** (K1)*,* ***encouraged*** (K2)*,* ***represent*** (K2)*,* ***strengthen*** (K2)*,* ***request*** (K3)*,* ***portrayed*** (K4)*,* ***vendors*** (K4)*,* and ***decrees*** (K5). Some words in the list involve grammatical features (plural form and past participle), but the result of the pilot study indicated that majority of the participants did not know the meaning of these words. In addition, the Form-Meaning Recall test, or F-MRt from the Pilot Study did not reveal any problems regarding these grammatical features due to clues given to help with the grammars (see discussion in Section 3.7). For these reasons, I decided to include them in this study for further insights into the effects of TFA components on retention of form with various features.

## 3.6 Procedures

This section described how the three tests: the adapted CATSS, adapted VKS, and F-MRt were implemented with the participants in the Pilot Study which was conducted from January 2021 to March 2021. It should be noted that all tests were developed to be taken online due to the pandemic. The potential issues arisen from the online tests may include the temptation to cheat in the online test and the technical problems such as inexperience in using online learning platform. These issues was first tackled by providing more understanding on the main purpose of taking part in the research project to the participants. It was emphasized that there was no extra credit for doing the test so the score will not be considered as a part of the course evaluation. However, I raised the importance of doing the test in aiming to encourage the participants to pay attention to it. I told them that the test will be provided as an additional practice in helping them to prepare for the mid-term exam.

Moreover, I asked for a consent from them to use a camera while learning and taking the test to observe their behaviors and avoid cheating (see ethical approval in Appendices: 4a, 4b, 5a and 5b). During the experiment, they had to turn on their camera throughout the time in classroom. Another issue about the technical problems was managed by checking if there is any obstacles during the distribution of the tests in the Pilot Study. As the data collection was gathered 6 months after the start of the pandemic, all participants were familiar with Microsoft Teams, a tool for online learning. They were encouraged to use this online learning platform in all courses in the university. This was likely to be convenient for them to learn and do the assigned tasks as well as performing tests through the online platform. While I was aware that there might be some limitations, these issues were not regarded as having significant effects to the study.

Detailed steps of the Pilot Study are as follows:

In January 2021, I distributed two online tests: the adapted CATSS and adapted VKS to four classes of the participants who took the same (University English II) course and has the same level of proficiency as the participants in the main study. This group of students were excluded in the main study. Each test was used in the same week, but on a different day according to the participants' class schedule. Because of the COVID-19 situation, on-site instructions at Thammasat University were shifted to online and operated by using online platforms such as Zoom and Microsoft Teams. The instructors in the four randomly selected classes used Microsoft Teams to teach their students. Thus, I asked for permission to visit their online classes. There were three visits for distributing the three tests, one visit for one test.

During the first week of January 2020, the participants received the Information Sheet (see Appendix 5a) giving details of the pilot study. The link of the document was posted in the Microsoft Teams. I asked for help from the instructors of each class to inform their students to read the document before the first visit. They had one week to review it and decide whether they would agree to participate in this research project or not.

After one week, I made a request to visit the students in the four randomly selected classes. The first visit took three hours. I first explained the details of the study to them again in their first language and allowed them to ask questions before making a decision to join in this study. Then, I asked the students to type their name under each section of the Information Sheet and the Consent Form (see Appendix 5a) if they understood the details provided in the documents if they agreed to participate in this study. Then, I explained them the details of the adapted online CATSS. As there are four modalities of the test, I showed them how to do each modality before giving them each test. The sequence of test modalities was the same as that of the new CATSS from the website (see Section 3.7.1 for detail). So, I distributed the first test to the participants by sending them a link of the productive recall modality that contained 20 test items via Microsoft Teams. They had 15 to 20 minutes to do each modality. Although I assumed that the test could be done within an hour, some students could not finish the tests within the provided time. I had to wait for some students to finish each test for about 20 to 30 minutes. The process continued until the four modalities were completed. They had to turn on their camera in Microsoft Teams while they were taking the four tests through Microsoft Form, so that I could observe their behaviours if they try to cheat during the test.

For the second visit, I distributed the adapted VKS to the participants in the four classes via Microsoft Teams to measure their knowledge on the vocabulary in the reading passage of the Unit1. This online test, created by using Microsoft Form, aims at selecting the target unknown words (see detail in Section 3.7.2). Similar to the adapted CATSS, I first explained the participants how to do the test. The participants were required to both write L1 definitions and form a sentence of each word. Because there were twenty-eight words to be tested, they had two hours to complete the test. They also had to turn on their camera until they finished it.

Two weeks later, I recruited 28 participants from one of the four classes to do the Form-Meaning Recall test that was eventually used as Pre-test and Posttests for the main study. It should be noted that later I used the terms Form-Meaning Recall (or F-MRt) with regard to both form and meaning recall, and Form Recall test (or FRt) for further analysis when meaning recall knowledge was not a concern. However, both terms used throughout the report refer to the same test that was implemented at three different times as Pre-test, Immediate Posttest, and Delayed Posttest of the current project. The test was developed from the list of unknown words as a result of the adapted VKS (see details in Section 3.7.2). I allowed two weeks after the implementation of the adapted VKS to construct the F-MRt test and invited three native speakers to validate it. Several factors such as test format and test appropriateness that might affect the test results were carefully observed and developed for the main study.

The observation from the Pilot Study helped to determine the estimate testing time. The participants had one hour to finish the F-MRt with 10 items through a link of Microsoft Form that was distributed via Microsoft Teams. They were also requested to turn on their camera while taking the test. Details about the F-MRt test validation are given in Section 3.7.3.

After the target words were identified by using the adapted VKS, I asked for help from a teacher who is currently teaching the University English II course to form sentences using these ten words. This is to make sure that the sentences are suitable for the students in the course as the teacher knows language proficiency level of these students. Then, I eliminated some letters of the ten target words as suggested by Laufer and Nation (1999) so that fewer than three letters were provided to avoid guessing. After that, three English teachers (two native and one non-native English speakers) who have at least five-year teaching experience were invited to check for item appropriateness. With regard to decontextualisation, the numbers of letters provided in each test item as well as sentence appropriateness were evaluated and improved for the Pilot Study and later for the Main Study (see details in Section 3.7.3 and Section 4.4.2).

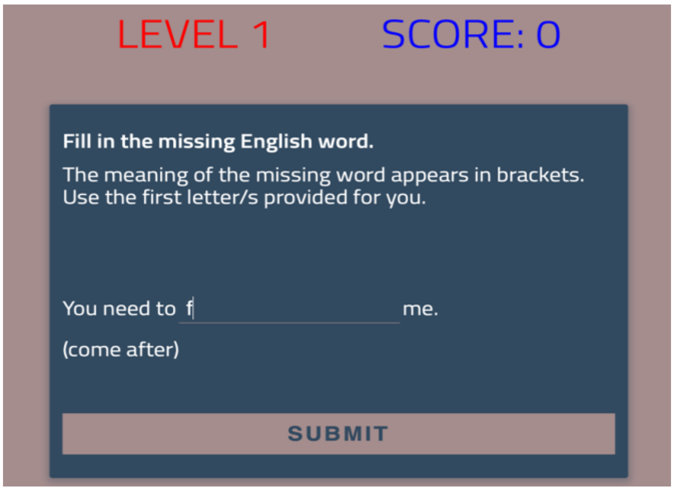
## 3.7 Pilot Study: participants, analyses and results

As mentioned in Section 3.6, three online tests were developed to use in the Pilot Study for different purposes. These tests are 1) the adapted CATSS, 2) the adapted VKS, and 3) the F-MRt for productive vocabulary measurement. Results of these tests would help me to finalise the target words and the tests used in the Main study. The analyses of the adapted CATSS would indicate the background knowledge of vocabulary of the target learners, which can give some general demographic data of the participants before planning the Main Study. The minimum number of samples for a Pilot Study as suggested by Hill (1998) and Issac and Michael (1995) is about 10 to 30 individuals. Because larger sample size would give more power to the calculation of the result, I had an intention to invite 80 students who took the English Foundation Course II at Thammasat University in Thailand from two random sections to participate in the pilot study. As mentioned in Section 3.2, these students have the same age range (between 18 and 20 years old) and English proficiency level (intermediate) as the participants in the main study. So, they were assigned randomly into sections. There are typically 40 students in a class. Due to the pandemic during January 2020 when the Pilot Study was conducted, the class size had been reduced to between 25 to 30 students in a class. I had to include the participants from four classes rather than two classes. As a result, there were 101 Thai undergraduate students (68 females and 33 males) who consented to participate in the pilot study (see the consent form and information sheet in Appendix 5a). While 101 students agreed to participate in doing the adapted CATSS (Computer Adaptive Test of Size and Strength) and the adapted VKS (Vocabulary Knowledge Scale), there were only one class of the participants (28 students) who were invited to take the Form-Meaning Recall test (or F-MRt) in the Pilot Study. This figure is sufficient as suggested by several researchers mentioned earlier. The results of VKS could help me to identify the target words for learning while the validation of the F-MRt could ensure that it is appropriate with the target groups and matches the purpose of the current study. I describe the analyses and results of these tests in detail below in the following sub-sections.

### 3.7.1 The adapted CATSS

The main purpose of using this test was to measure prior vocabulary knowledge of the students who had the same level of English proficiency and took the same English course as the participants in the main study (see also Section 4.4.1). These students were expected to have similar background knowledge of vocabulary in English as the participants in the main study. Exploring students' receptive knowledge could also be useful at this stage although my goal was to find the test that can examine students' general background knowledge on controlled productive vocabulary before inviting them to participate in the present study.

To measure participants' prior vocabulary knowledge, different instruments have been used by previous research on TFA. Hu and Nassaji's (2016) and Chaharlang et al.'s (2018) studies employed the Vocabulary Levels Test, or VLT (Nation, 2001) at 2,000 frequency level while other studies utilised scores from IELTS/TOEFL (Gohar et al., 2018), or Oxford Placement Test, or OPT (Khoshsima & Eskandari, 2017; Chaharlang & Farvardin, 2018). So, I sought for a vocabulary test that could measure general vocabulary knowledge receptively and productively at high frequency level. Unlike studies of Hu and Nassaji (2016) and Chaharlang and Farvardin (2018), I decided to employ the new CATSS (Aviad-Levitzky et al., 2019) at 1,000 and 2,000 frequency levels (see Figure 3.2) instead of Schmitt et al.'s (2001) VLT to ensure the participants' level of proficiency.  The CATSS was chosen because it measures the recall of both word form and word meaning at different 1,000 BNC/COCA word levels. The validity of the test has been confirmed by Laufer and Goldstein (2004) and Laufer and colleagues' (2004) studies. Moreover, this test is also available online on Lextutor (www.lextutor.ca/vp/eng.), a well-known website for vocabulary resources, which makes it convenient for me to deliver the test to the participants.



LEVEL 1 SCORE: 0

Fill in the missing English word.

The meaning of the missing word appears in brackets.
Use the first letter/s provided for you.

You need to f________ me.

(come after)

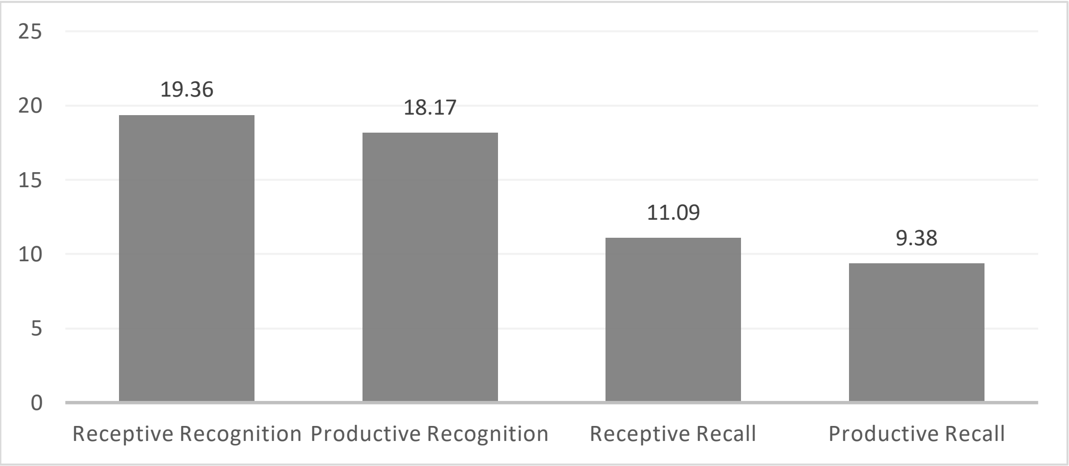
**Figure 3.2. Sample of the new CATSS (productive recall modality)**

As English proficiency of the participants in this study is between lower and upper intermediate levels, testing them on low-frequency words might not be appropriate. Webb, et al. (2017, p. 33) claim that "It is probably only necessary to administer the 2000 word level to beginners since they are unlikely to have mastered any of the subsequent levels." This is similar to Schonell et al.'s (1956) suggestion that knowledge of 2,000 word level, which accounts for about 80% coverage of a text (Schmitt et al., 2001; Nation, 2006), is required for basic communication. Previous studies (e.g., Dang & Webb, 2016a; 2016b; Sudarman & Chinokul, 2018; Dang, 2020; Sun & Dang, 2020) with university EFL students in various contexts such as Vietnam, China, Taiwan, Thailand, Indonesia, Denmark, and Spain have also found that a large number of these students have not yet mastered the most frequent 2,000 words receptively. Research also shows that L2 learners' productive vocabulary knowledge is likely to be smaller than their receptive vocabulary knowledge (Waring, 1997b; Nation, 2013). It is essential to investigate the participants' knowledge of highly frequent words. Therefore, it can be expected that the target participants will be unlikely to master words beyond the most frequent 2,000 words productively. Thus, I only used high-frequency word levels (the 1,000 and 2,000 word levels) of the CATSS for two reasons. First, these levels are suitable for intermediate level students because the test measures not only receptive word knowledge, but also productive use in each frequency level. Although this study included some target learning words in the mid-frequency level (see discussion in Section 3.5), measuring vocabulary proficiency at high frequency level should be sufficient to control for the participants' background knowledge. Moreover, it helps to avoid cognitive fatigue that may affect the performance of the test takers. As they have to take several tests during the pilot study, spending a lot of time taking one test may affect their attention. So, the CATSS at high word-frequency level seems to be appropriate to use for selecting the participants. The findings from the pilot study would help to estimate the vocabulary knowledge of and select the participants to the main study.

Another purpose of using the adapted CATSS was to pilot the test before implementing it in the Main Study. Problems resulting from time on test and test items during the Pilot Study were collected for development. The lack of vocabulary knowledge that was found could also help to justify the selection of levels and modalities of the target vocabulary to be included in the present study. The data from the four test modalities: productive recall, receptive recall, productive recognition, and receptive recognition was gathered in the pilot study and analysed to explore vocabulary knowledge at high frequently level.

In order to identify the data analysis method, I sought for the information from previous research. I mostly relied on the results of Aviad-Levitzky et al.' s (2019) study. Aviad-Levitzky and colleagues, who were the test developers, argued that recall modalities tend to be more difficult than recognition modalities, so that they gave different scores (1 for productive recall, 0.75 for receptive recall, 0.5 for productive recognition, and 0.25 for receptive recognition) to each modality. However, the adapted CATSS used different scoring criteria (see more detail in Section 4.4.1). One point was given to a correct item while zero point was given to an incorrect item. Descriptive statistics such as Mean and Standard Deviation (*SD*) used in the previous study were employed to analyse the findings.

According to the Pilot Study, I noticed consistent patterns in the adapted CATSS scores among the four modalities (see Figure 3.3). The results showed similar scores between receptive recognition (Mean = 19.36, *SD = 0.80*) and productive recognition (Mean = 18.17, *SD = 1.31*), and nearly the same pattern of the results of receptive recall (Mean = 11.09, *SD = 1.95*) and productive recall (Mean = 9.38, *SD = 1.92*), where productive is always lower than receptive. It seems to correspond to the findings of Aviad-Levitzky et al. (2019) in that receptive recognition score ranked the highest (97%, Mean = 19.36), followed by productive recognition (91%, Mean = 18.17), receptive recall (55%, Mean = 11.09), and productive recall (47%, Mean = 9.38), respectively.



*Note: total score of each modality = 20*

**Figure 3.3. The results of the adapted CATSS from four modalities**

The lack of knowledge on receptive and productive recall were found even though the results from the test could be interpreted that the learners have knowledge of vocabulary at high frequency level. It could be expected that learners who know words in low frequency levels would know or understand highly frequent words (Kremmel & Schmitt, 2018). On the other hand, if they do not know highly frequent words regarding some aspects such as recall, we would expect that they might also have problems with both receptive recall and productive recall of words in the mid- and low-frequency levels. As a result, words from 3,000 level onwards in the CATSS would be unnecessary to be measured in this study because knowledge on receptive recall and productive recall at high frequency levels of word knowledge is still needed to be developed to some extent.

By considering each word level, the result illustrates the same pattern as that of the overall scores in the Figure 3.4 in which recognition was higher than recall in all levels. The participants got the highest score (Mean = 9.92, *SD* = 0.27) in receptive recognition of the Level 1, followed by productive recognition (Mean = 9.60, *SD* = 0.63), receptive recall (Mean = 6.49, *SD* = 1.26), and productive recall (Mean = 6.33, *SD* = 1.14), respectively (see Figure 3.4). This means that the participants know 99.2% of the K1 words and 94.4% of the K2 words at the receptive recognition level and 96.0% of the K1 words and 87.5% of the K2 words at the productive recognitions. Ideally, it is expected that students should know 100% of the words at high frequency level. Previous studies using VLT and UVLT (e.g., Hacking et al., 2017; Hu & Nation, 2000; Rodgers, 2013; Xing & Fulcher, 2007; Dang et al., 2020; 2021) set 80% as the acceptable level of mastery.

Despite the level difference, the pattern found from the result of the Level 2 was also similar to that of the Level 1. However, the participants performed better in the Level 1 than the Level 2 in all modalities. In the Level 2, the highest mean score was 9.44 (*SD* = 0.67) from receptive recognition while the results of productive recognition, receptive recall, and productive recall were 8.75 (*SD* = 1.08), 4.67 (*SD* = 1.49), and 3.04 (*SD* = 1.41), respectively.

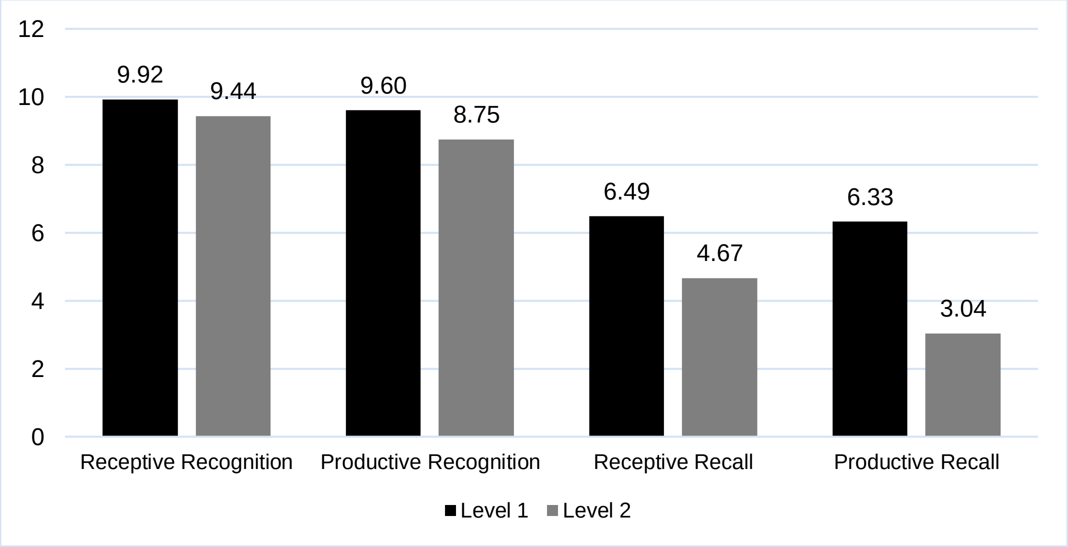


**Figure 3.4. The results of the adapted CATSS by level**

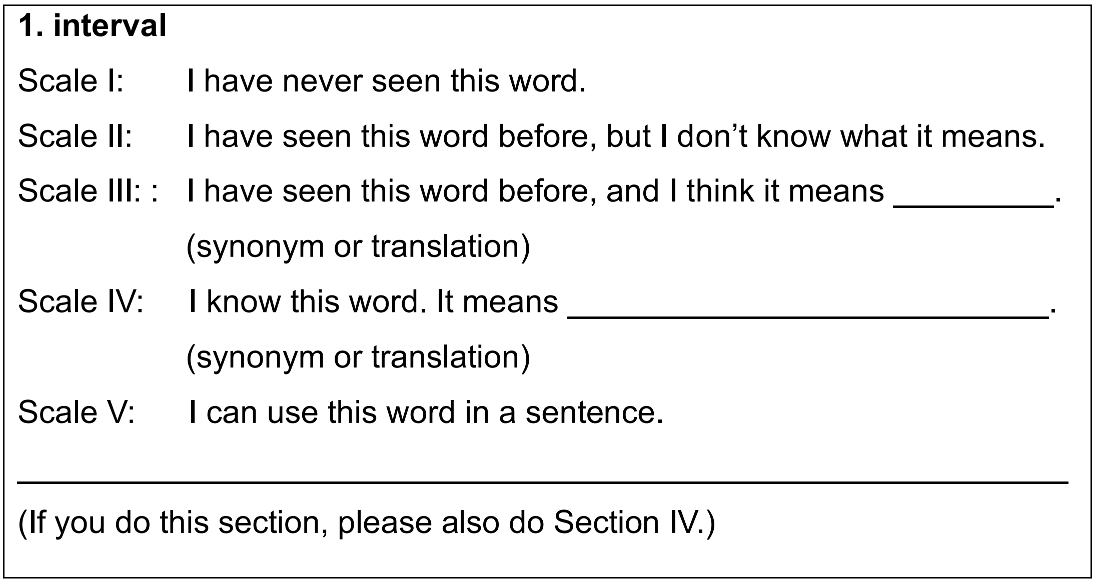
Moreover, both receptive recognition scores (9.92 in Level 1 and 9.44 in Level 2) and productive recognition scores (9.60 in Level 1 and 8.75 in Level 2) were higher than 80%. It is likely that the participants might have no problem with receptive vocab knowledge of K1 and K2 levels. However, it was found that the participants tended to have lower knowledge in terms of recall than recognition in both K1 and K2 levels. The participants only knew 64.9% of the K1 and 46.7% of the K2 words at the receptive recall and 63.3% of the K1 and 30.4% of the K2 at the productive recall. As the participants are likely to have insufficient knowledge of the K1 and K2 words at the recall level, this finding reflects the problems on form and meaning recall which is the main purpose of conducting this study.

In conclusion, results of the adapted CATSS in the Pilot Study indicated that the participants have sufficient knowledge of high frequency words at recognition level but were likely to have inadequate knowledge of receptive and productive recall of both K1 and K2 levels. As the participants in the main study had fairly similar to those in the pilot study, it was expected that they would have adequate knowledge of the most frequent 2,000 word families at the recognition level. However, the test was used with the participants in the Main Study to ensure that each group of the participants has no difference in their prior vocabulary knowledge (see more detail in Section 4.4 in the Chapter 4 and Section 5.1 in the Chapter 5). The results from the Pilot Study also revealed the lack of knowledge on vocabulary recall. This was the reason for including some K1 and K2 items as the target words in the main study.

### 3.7.2 The adapted Vocabulary Knowledge Scale (VKS)

In the Pilot Study, I also measured the participants' knowledge of the target words to identify the words that they do not know by using the adapted VKS. Similar to Gohar et al.' s (2018) study, I used L1 translation to explore the participants' background knowledge of the selected words. As both receptive and productive vocabulary knowledge could be explored through the VKS (Paribakht & Wesche, 1993), and this test has been widely used among research on TFA (e.g., Khoshsima & Eskandari, 2017; Zou, 2017; Zou & Xie, 2018), I therefore adapted Scale IV and Scale V of the test to use in my pilot study. The VKS has five scales (see Figure 3.5) that can measure both receptive and productive knowledge of test takers.

Scale IV of this test can be used to elicit meaning recall as it has been adopted to use widely by previous research (e.g., Kim, 2011; Khoshsima & Eskandari, 2017; Zou, 2017;) to measure receptive vocabulary knowledge on meaning. Because some parts of the VKS may not be necessary to use in the pilot study, the test was adapted to match the purpose of using it.



**1. interval**

Scale I: I have never seen this word.

Scale II: I have seen this word before, but I don't know what it means.

Scale III: : I have seen this word before, and I think it means _________.
(synonym or translation)

Scale IV: I know this word. It means ___________________________.
(synonym or translation)

Scale V: I can use this word in a sentence.

____________________________________________________________

(If you do this section, please also do Section IV.)

**Figure 3.5. Sample of the VKS (Wesche and Paribakht, 1996)**

The test format is shown in Figure 3.6 below. Similar to the adapted CATSS, this adapted VKS was developed by using an online platform which is Microsoft Form. The test includes twenty-eight items of words as mentioned in Section 3.5. Words that majority of those participants could use correctly in terms of meaning were excluded from the study.

**Identity**

means______________________________________________

____________________________________________________

*(English synonym or Thai translation)*

**I can use this word in a sentence:**

________________________________________________________

________________________________________________________

(*please write a sentence in English*)

**Figure 3.6. Sample of the adapted VKS for the Pilot Study**

Below I present the scoring method and analysis of the findings from the VKS.

***Scoring***

As presented in Table 3.15, I adapted the VKS Scoring Scale from Paribakht and Wesche (1993;1997) to grade the participants' responses. I only employed the last three categories of the VKS in the current research since the test for selecting target vocabulary did not aim to measure self-report scores. It attempts to explore unknown words by measuring knowledge of vocabulary in terms of meaning and usage.

***Table 3.15. VKS scoring scale (Paribakht & Wesche, 1997)***

| Categories | Possible scores | Meaning of scores |
|---|---|---|
| I | 1 | The word is not familiar at all. |
| II | 2 | The word is familiar but its meaning is not known. |
| III | 3 | A correct synonym or translation is given. |
| IV | 4 | The word is used with semantic appropriateness in a sentence. |
| V | 5 | The word is used with semantic appropriateness and grammatical accuracy in a sentence |

Similar to Paribakht and Wesche' s (1993; 1997) studies, the target testing word are regarded as unknown by the participants if a synonym in English or definition in the mother language, Thai, was given incorrectly or not provided. This category can be useful in eliciting depth of vocabulary knowledge of the participants. Thus, one point was given when the participants provided an English synonym or Thai translation. However, this study used two scoring systems which are sensitive and strict (See Table 3.16).

***Table 3.16. Scoring criteria for grading definitions of the testing words***

| Rating system | Meaning | Meaning of scores |
|---|---|---|
| (a) sensitive rating | 0 | Incorrect answer |
| | 0.5 | Partially correct answer |
| | 1 | A correct synonym or translation is given. |
| (b) strict rating | 0 | Incorrect answer |
| | 1 | A correct synonym or translation is given. |

The participants received 0.5 point for the word that its meaning is partially correct. For sentence formation, I measured all sentences by relying on two main suggested criteria of Paribakht and Wesche (1993): 'semantic knowledge and grammatical exactness' (p.16). As suggested by the researchers, accuracy in terms of semantic and grammar must relate to the use of the target words.

### *Analysis and interpretation*

With regard to the selection of unknown target words, I initially looked for words that majority of the participants do not know by using the scoring criteria in Table 3.17. It was found that most of them tended to have knowledge of several words such as *relationship*, *begin*, and *prevent* as presented in Table 3.18. About 30% of the participants already knew several words such as *identity*, *performances*, and *promote* while only a few of them (below 30% of the participants) knew words such as *encouraged*, *vendors*, *request*, *strengthen*, *concern*, *decrees*, *represent*, *shortage*, *portrayed*, and *issued*. The number of unknown words in the below 30% list is relatively small, but it is similar to the amount of target vocabulary in previous empirical research (Khoshsima & Eskandari, 2017; Gohar et al., 2018) and that of words to be learned in all units of the course book. I decided to include the words such as *issued* and *venders* that contains some grammatical features in the list of the target unknown words because the participants did not know the meaning of these words, and the result showed that they could provide the correct meanings to the words from the above 30% list.

***Table 3.17. Scoring criteria of the test of unknown words***

| Score | Sensitive Scoring | Strict Scoring |
|---|---|---|
| 1 | • The target word is used in the correct position relative to its part of speech (POS) with semantic appropriateness.<br>• The sentence is grammatically correct or contains few grammatical errors that do not affect the target word, i.e., *She alway **encouraged** her boyfriend.*<br>• The sentence sounds unusual but is grammatically correct, i.e., *I will **represent** my story for you.* | • The target word is used in the correct position relative to its part of speech (POS) with semantic appropriateness.<br>• The sentence is grammatically correct or contains few grammatical errors that do not affect the target word, i.e., *She alway **encouraged** her boyfriend.*<br>• The sentence sounds unusual but is grammatically correct, i.e., *I will **represent** my story for you.* |
| 0.5 | • The word is used in the correct position with semantic appropriateness, but the sentence contains grammatical errors that may affect the meaning of the target word, i.e., *“This book **portrayed** your about beauty.”* | |
| 0 | • The target word is used in an incorrect position relative to its part of speech (POS), i.e., *The **strengthen** is important in daily life.*<br>• The target word is used with semantic inappropriateness (i.e., errors in word choice or spelling that affect meaning).<br>• The sentence contains major grammatical errors that directly affect meaning, i.e., *I wasn’t **represents** project.* | • The target word is used in an incorrect position relative to its part of speech (POS), i.e., *The **strengthen** is important in daily life.*<br>• The target word is used with semantic inappropriateness (i.e., errors in word choice or spelling that affect meaning).<br>• The sentence contains major grammatical errors that directly affect meaning, i.e., *I wasn’t **represents** project.* |

Also, I used the scoring criteria in Table 3.17 to grade sentences that the participants produced from the ten selected words (in below 30% group) to make sure that they could not be able to use these words in written production (see Table 3.18). If they could not give both definitions and produce a sentence using the testing word correctly, it was interpreted to mean that they do not know that word.

***Table 3.18. Results of a vocabulary selection test by a correct synonym or translation***

| Rank | Testing words | *Freq. level | Percentage (%) of the participants knowing the word by definitions *(N =101)* | | Remarks: *percentage of participants knowing the words* |
|---|---|---|---|---|---|
| | | | **Sensitive** | **Strict** | |
| 1 | relationship (n) | K1 | 91.09 | 89.11 | Above 50 % |
| 2 | important (adj) | K1 | 86.63 | 85.15 | |
| 3 | protect (v) | K1 | 90.59 | 89.11 | |
| 4 | begin (v) | K1 | 86.14 | 81.19 | |
| 5 | government (n) | K1 | 82.67 | 81.19 | |
| 6 | history (n) | K1 | 83.66 | 80.20 | |
| 7 | replace (v) | K2 | 64.36 | 62.38 | |
| 8 | creation (n) | K2 | 53.96 | 51.49 | |
| 9 | national (adj) | K1 | 52.97 | 50.50 | |
| 10 | performances (n) | K2 | 52.97 | 50.49 | |
| 11 | identity (n) | K2 | 51.98 | 44.55 | Above 30% but below 50% |
| 12 | traditional (adj) | K2 | 50.99 | 46.53 | |
| 13 | contest (n) | K3 | 47.52 | 46.53 | |
| 14 | responded (v) | K3 | 47.52 | 45.54 | |
| 15 | prevent (v) | K2 | 47.52 | 38.61 | |
| 16 | promote (v) | K3 | 39.11 | 32.67 | |
| 17 | colonization (n) | K3 | 37.13 | 34.65 | |
| 18 | held (v) | K1 | 32.67 | 31.68 | |
| 19 | encouraged (v) | K2 | 29.70 | 27.72 | Below 30% |
| 20 | vendors (n) | K4 | 28.71 | 28.71 | |
| 21 | strengthen (v) | K2 | 20.30 | 15.84 | |
| 22 | shortage (n) | K1 | 13.37 | 10.89 | |
| 23 | request (n) | K3 | 18.81 | 15.84 | |
| 24 | concern (n) | K1 | 17.82 | 15.84 | |
| 25 | decrees (n) | K5 | 16.34 | 10.89 | |
| 26 | represent (v) | K2 | 13.86 | 11.88 | |
| 27 | portrayed (v) | K4 | 8.41 | 6.93 | |
| 28 | issued (v) | K1 | 6.44 | 3.96 | |

*Note: *Freq. = Frequency*

Then, I invited a native English teacher to grade the participants' responses. Among 101 participants, some of them did not form sentences for some target words. There were only 90 sentences that the participants responded. After that, I analysed the received scores from the two raters using Cohen's Kappa Coefficient in order to check inter-rater reliability. The results from both strict and sensitive scoring (see Table 3.19) were 0.91 and 0.88 (n = 90, *p* < .001), respectively. This means that both scoring systems received a strong level of agreement between raters (Landis and Koch, 1977; McHugh, 2012).

***Table 3.19. Analysis of inter-rater reliability from both rating systems***

| Rating System | Number of Valid Cases | Value |
|---|---|---|
| Strict | 90 | 0.905 |
| Sensitive | 90 | 0.875 |

*Note: p < .001*

***Table 3.20. Samples of correct written production***

| Item | Testing words | Sample sentences produced by the participants |
|---|---|---|
| 1 | Encouraged (verb/K2) | *"I will go to **encouraged** to him." (0.5 point)*<br>*"I was **encouraged** to attend the Awards Ceremony." (1 point)* |
| 2 | Vendors (noun/K4) | *"In market has a few **vendor**." (0.5 point)*<br>*"Many **vendors** want to sell their products or services." (1 point)* |
| 3 | Strengthen (verb/K2) | *"He **strengthen** to the body." (0.5 point)*<br>*"Breakfast can **strengthen** my power in the morning." (1 point)* |
| 4 | Shortage (noun/K1) | *"I'm no **shortage** of ideas." (0.5 point)*<br>*"There was a **shortage** of water in summer." (1point)* |
| 5 | Request (noun/K3) | *"I sent a **request** to follow my mom's Instagram." and "His **request** gave her an idea." (1 point, each)* |
| 6 | Concern (noun/K1) | *"You need to have **concern** your test."  (0.5 point)*<br>*"I appreciate your **concern** but I can handle it by myself." (1 point)* |
| 7 | Decrees (noun/K5) | *"Every country has a strong **decrees**." (0.5 point)*<br>*"You are breaking our **decrees**." (1 point)* |
| 8 | Represent (verb/K2) | *"Tom Yum is **represent** Thai's food." (0.5 point)*<br>*"Pad Thai **represents** the uniqueness of Thailand." (1 point)* |
| 9 | Portrayed (verb/K4) | *"This book **portrayed** your about beauty." (0.5 point)*<br>*"The author **portrayed** the life of a policeman." (1 point)* |
| 10 | Issued (verb/K1) | *"Serum cream **issue** last year." (0.5 point)*<br>*"Transport Department **issued** a driving license to me." (1 point)* |

To conclude, although some participants knew the definitions of the target words, only a few students were able to use those words correctly in sentences they produced as presented in Table 3.20. Because of this, I selected the following target unknown words for this study: (K1)*, issued* (K1)*, shortage* (K1)*, encouraged* (K2)*, represent* (K2)*, strengthen* (K2)*, request* (K3)*, portrayed* (K4)*, vendors* (K4)*,* and *decrees* (K5)*.*

### 3.7.3 The Form-Meaning Recall test

As productive knowledge is the aim of this study, I then sought for the test that can elicit the knowledge in terms of form recall of the participants. Previous research related to TFA framework has used similar instruments such as Paribakht and Wesche' s (1993; 1997) VKS or Folse' s (2006) adapted VKS test (e.g., Hu & Nassaji, 2016; Zou & Xie, 2018) and L1 translation (Gohar et al., 2018) as immediate and delayed posttests to explore the vocabulary gains after learning with different activities. Both receptive and productive knowledge can be measured through the five scales of the VKS as presented in Figure 3.5 above. As a result, the test has been widely used in the field of vocabulary. Also, its format is similar to L2→L1 translation test commonly used by researchers (e.g., Hu & Nassaji, 2016; Gohar et al., 2018) as both immediate and delayed Posttests to elicit meaning recall. Since translation method shows a positive effect on meaning recall (Ramachandran and Rahim, 2004), Scale IV of the VKS could be used as a tool to test recall in terms of meaning. Thus, the format seems to be appropriate to measure receptive knowledge mostly. The format of this scale was adopted to use in the current study. This means the participants have to write an appropriate L1 definition for each L2 target item.

Although the VKS could measure recall, it is more suitable for less controlled writings which do not aim to test form recall. I therefore could not use its format to elicit form recall of the participants. I then sought for the other tests that could be able to elicit productive knowledge. These include tests for controlled productive vocabulary such as LEX30 (Meara & Fitzpatrick, 2000) and Productive Vocabulary Levels Test, or PVLT (Laufer & Nation, 1990), and tests for free productive vocabulary such as Lexical Frequency Profile, or LFP (Laufer & Nation, 1995). As discussed earlier in the literature, free and less controlled writing might not be appropriate to use in this study. It seems that only the LEX30 (see Figure 3.7) and PVLT (see Figure 3.8) have been used to measure controlled productive knowledge, but these tests do not either include the target words that this study aims to explore.

30. window _________ _________ _________ _________

**Figure 3.7. Sample of the LEX30 format (Meara & Fitzpatrick, 2000)**

The LEX 30 is more suitable to measure the breadth of productive vocabulary knowledge (Fitzpatrick & Meara, 2004; Walters, 2012) while PVLT seems to be an appropriate test to measure productive vocabulary in terms of knowledge growth (Laufer & Nation, 1999) which this study aims to explore after the participants learn a set of new target words. Also, the LEX30 showed problems and difficulties in interpretation as scores from different frequency levels are awarded similarly (Walters, 2012). This means that the same score is given to all words at a particular level in the test. It might be difficult to identify if test takers have certain knowledge of low frequency words as they can receive high scores if they only produce high frequency words. Also, it seems to be less controlled than the PVLT. For example, if the participants do not write the target words as one of their answers, this might be because they do not remember them or they may be able to retain the words but may happen to write other words that they are more familiar with. Walter (2012) also found that although LEX30 is an appropriate test to measure productive recall, its results seem to be more valid when measuring productive vocabulary use of learners with high proficiency levels. I therefore decided to employ the PVLT test format (see Figure 3.8) to develop a pre-test and two Posttests of the current study because it is more controlled than the other tests and can be used to measure form recall.

The book covers a series of isolated epis_____ from history.

**Figure 3.8. Sample of the PVLT format (Laufer & Nation, 1990, p.37)**

Then, I created the Form-Meaning Recall test (F-MRt) that includes the ten target words mentioned in Section 3.5. The sample of test format is shown in

Figure 3.9 below. Dashes were used as a clue to control number of letters for each test item in order to eliminate chances of other possible answers that might affect the result of this study. The test will be called a Form-Meaning Recall test, or F-MRt throughout the study.

Complete the sentence below with the most appropriate word and write Thai translation or an English synonym of the word.

1. There was a serious **sh**_ _ _ _ _ _ of water last year due to the unusually long hot summer.

   **sh**_ _ _ _ _ _ means ________________________________________
   *(write Thai translation or an English synonym)*

**Figure 3.9. Sample of the Form-Meaning Recall test used as Pre- and Posttest**

This test was first implemented in the Pilot Study to ensure the test validity before implementation. I also reported in this section the analysis and results of the F-MRt used in the Pilot Study. The data analysis of the Form-Meaning Recall test had two stages: 1) the analysis of lexical variation of test items and 2) the analysis of item difficulty.

***Stages 1: Analysing lexical variation***

This is an initial stage in which content and language in the test was checked before the implementation in the pilot study. Firstly, I invited three native English-speaking teachers who have at least five-year teaching experience to check the appropriateness of the items in terms of difficulty (*How well does each sentence match the level of the students/Thai non-English major undergraduates?),* number of letters (*Should the provided letters be reduced or added more to avoid guessing and misleading result?*), and content (*Are there any other possible answers for each item?*). Some of the revised items based on the suggestions are:

Original sentences

1. *"The president of the United State* ***iss*** *_ _ _ a statement to the press yesterday."*

2. *"There are many street **v**_ _ _ _ _ _ selling food and drinks to people at the fun fair."*
3. *"My **c** _ _ _ _ _ _ is that we are not going to get our group project done in time and we are going to fail."*

Suggested sentences

4. *"The president of the United State **i**_ _ _ _ _ a statement to the press yesterday."*
5. *"There are many **v** _ _ _ _ _ _ selling food and drinks to people in Bangkok."*
6. *"My **con** _ _ _ _ is that we are not going to get our group project done in time and we are going to fail."*

They suggested that the sentence *"The president of the United State **iss** _ _ _ a statement to the press yesterday."* is appropriate with English proficiency level of the students, but it is too easy with 3 letters provided as a clue for the target testing word. For the sentence "*There are many street **v**_ _ _ _ _ _ selling food and drinks to people at the fun fair,"* they commented that the word '*vendors'* works ok alone without *"street"* and suggested changing the context to "Bangkok." Moreover, it is likely that the students will write the word '*comment*' rather than '*concern*' in the sentence 3 above, so they suggested adding more letters to avoid other possible answers. However, I only provided one initial letter for each target item in the pilot study in order to check if there is a need for more letters/clues as suggested by the native English speaker teachers. In the pilot study, the test items were alphabetically ordered starting from high frequency bands (K1 and K2) to mid-frequency bands (K3, K4, and K5). The order of the ten items in the pilot study was **Item 1**: concern, **Item 2**: issued, **Item 3**: shortage, **Item 4**: encouraged, **Item 5**: represent, **Item 6**: strengthen, **Item 7**: request, **Item 8**: portrayed, **Item 9**: venders, and **Item 10**: decrees. This is to reduce the possible demotivation if students were unfamiliar with too many words in sequence.

However, the ten target items above were randomly ordered by using an randomisation website (http://www.graphpad.com/quickcalcs/index.cfm) in the main study to avoid the selection bias and accidental bias (Suresh, 2011). This online website helped to randomly shuffle the order of the ten items. The patterns below in Table 3.21 were used to order the items in the Pre-test, Immediate Posttest, and Delayed Posttest.

***Table 3.21. Computer-generated random items across the three tests***

| Sequence of items | Pre-test | Immediate Posttest | Delayed Posttest |
|---|---|---|---|
| 1 | Item 3 | Item 9 | Item 9 |
| 2 | Item 4 | Item 2 | Item 1 |
| 3 | Item 7 | Item 5 | Item 4 |
| 4 | Item 6 | Item 3 | Item 5 |
| 5 | Item 2 | Item 7 | Item 6 |
| 6 | Item 5 | Item 10 | Item 8 |
| 7 | Item 1 | Item 6 | Item 10 |
| 8 | Item 8 | Item 1 | Item 3 |
| 9 | Item 10 | Item 8 | Item 7 |
| 10 | Item 9 | Item 4 | Item 2 |

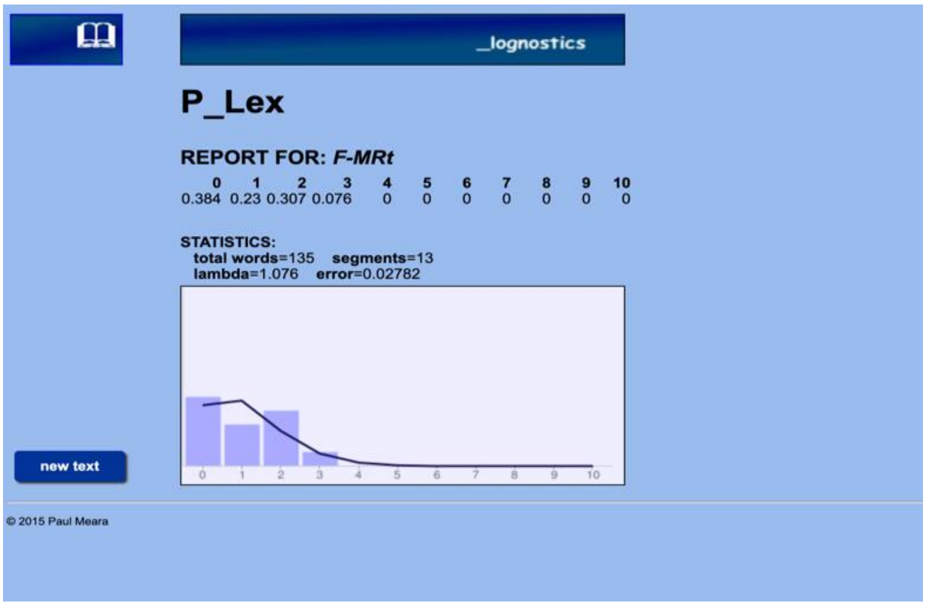
Furthermore, the ten sentences were analysed through two programs which are VocabProfile in Lextutor website and P_Lex, which is a programme developed by Meara and Bell (2001). I used VocabProfile to look at lexical variation in terms of type/token ratio, and P_Lex to measure lexical richness of vocabulary in short texts. P_Lex tells us an index score of proportion of infrequent words occurred in a text which is called ‘lambda’, ranging between 0 and 4.50. The higher the lambda score the greater proportion of infrequent words in the text (Meara & Bell, 2001). The results from VocabProfile revealed that approximately 85.20% of tokens were in K1 level (see Table 3.22).

***Table 3.22. Analysis of F-MRt items by VocabProfile (BNC-COCA1-25K)***

| Freq. Level | Families (%) | Types (%) | Tokens (%) | Cumul.token |
|---|---|---|---|---|
| K1 | 73 (80.20) | 79 (79.80) | 115 (85.20) | 85.20 |
| K2 | 12 (13.20) | 12 (12.12) | 12 (8.90) | 94.10 |
| K3 | 3 (3.30) | 3 (3.03) | 3 (2.20) | 96.30 |
| K4 | 2 (2.20) | 2 (2.02) | 2 (1.50) | 97.80 |
| K5 | 1 (1.10) | 1 (1.01) | 1 (1.07) | 98.50 |
| Off-List | 2 (2.02) | 2 (2.02) | 2 (1.48) | 99.98 |

There were only 3 words in levels 4 and 5, and these words are the target testing words. *Type/token* ratio (TTR) of the text was 0.73, indicating that the text had high lexical richness as the TTR ratio is close to 1.

The analysis of P_Lex showed that the lambda score of each sentence ranged between 0 and 2. When analysing all sentences together, the lambda score was 1.08, meaning that there were only a few infrequent words in all sentences as we can notice from the low value of lambda score of all sentences in Figure 3.10. Because lambda scores seem to be more stable with short length of texts, comparing to the Lexical Frequency Profile (LFP) scores (Meara and Bell, 2001), it can give a higher reliability of the result from a very short text analysis.



P_Lex

REPORT FOR: *F-MRt*

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.384 | 0.23 | 0.307 | 0.076 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

STATISTICS:
total words=135 segments=13
lambda=1.076 error=0.02782

© 2015 Paul Meara

**Figure 3.10. Analysis of all sentences of the F-MRt by using P_Lex**

### *Stages 2: Analysing item difficulty and testing time*

This stage was done after the F-MRt were analysed to check the lexical variation of the language in all items and were implemented with the participants in the Pilot Study. It should be noted that only one class (with 28 persons) of the participants in the Pilot Study were invited to do this test (see details in Section 3.6).

### *Scoring*

After the data collection process, the responses were evaluated by using two rating systems: sensitive and strict (see Table 3.23). One point was given to any correct response on form and meaning for both systems. However, for sensitive rating, 0.5 point was given to answers that were partially correct in terms of minor spellings or comprehensible definitions.

***Table 3.23. Two scoring schemes (sensitive and strict) for F-MRt analysis***

| **Meaning recall** | **Form recall** |
|---|---|
| (a) sensitive rating: 0: incorrect;<br>0.5 partially correct, 1: fully correct | (a) sensitive rating: 0: incorrect;<br>0.5 partially correct, 1: fully correct |
| (b) strict rating: 1: full correct,<br>0: incorrect or partially correct | (b) strict rating: 1: full correct,<br>0: incorrect or partially correct |

### *Analysis and interpretation*

Among the 28 participants, the Mean scores of the overall test from sensitive and strict ratings were 7.94 (*SD = 3.58*) and 7.68 (*SD = 3.63*) out of 20, respectively (see Table 3.24). Since the main aim in doing the pilot study is to analyse the ten constructed items on form, the scores on meaning only helped when the participants did not response to any items of the test on form correctly. The results of the test on meaning were consistent in that when the participants could not produce a correct form of an item the meaning of that item was also not provided. So, at this stage I paid attention to scores of the test on form rather than meaning for developing the10-item test on form. The average scores of the test on form when graded by sensitive and strict rating systems were 4.14 (*SD = 1.70*) and 4.04 (*SD = 1.75*) as shown in Table 3.24 below.

***Table 3.24. Descriptive statistics from the pilot study***

| **Test** | **Scoring System** | **n** | **Total** | **Mean** | ***SD*** |
|---|---|---|---|---|---|
| F-MRt (all 20 sub-items) | sensitive | 28 | 20 | 7.94 | 3.58 |
| F-MRt (all 20 sub-items) | strict | 28 | 20 | 7.68 | 3.63 |
| Form only (10 sub-items) | sensitive | 28 | 10 | 4.14 | 1.70 |
| Form only (10 sub-items) | strict | 28 | 10 | 4.04 | 1.75 |
| Meaning only (10 sub-items) | sensitive | 28 | 10 | 3.80 | 1.92 |
| Meaning only (10 sub-items) | strict | 28 | 10 | 3.64 | 1.95 |

*Note: n = total number of the participants; SD = Standard Deviation*

Normal distribution helps to draw reliable conclusions of the results (Ghasemi & Zahediasl, 2012). Although Kolmogorov-Smirnov (K-S) test has been widely used to test normality, several issues such as on its low power and high sensitivity to extreme values have been criticised by researchers (e.g., Steinskog et. al., 2007). It is suggested that Shapiro-Wilk test is a better choice for normality testing since it can give a better power for a giving significant than the K-S test (Steinskog et. al.,2007; Razali & Wah, 2011). Both tests are suitable for a small sample size less than 50 (Elliott & Woodward, 2007). Hence, it was used to prove normality of the results of the test on form. Test of normality revealed that the data is normally distributed for both sensitive rating (p ≥ 0.53) and strict rating (*p* ≥ .25) at *p* < .05 (see Table 3.25) because the null hypothesis is accepted.

***Table 3.25. Test of normality of the test on form by Shapiro-Wilk***

| | Mean | Mode | SD | Statistic | df | Sig.* |
|---|---|---|---|---|---|---|
| Form (sensitive) | 4.14 | 4.00 | 1.70 | 0.97 | 28 | **0.53*** |
| Form (strict) | 4.04 | 4.00 | 1.75 | 0.95 | 28 | **0.25*** |

**p < .05*

Then, I looked for standard scores (or z-scores) of both rating systems. Standard score can generally be useful to identify possibility of scores that are greater or less than the mean scores within a normal distribution and can be used to compare scores from different measures (i.e., sensitive scoring and strict scoring) that have different normal distributions. Equation 3.1 shows the formular of z-score calculation.

$$\text{Standard (z) score} = \frac{X-\mu}{\sigma}$$

$x = score$
$\mu = Mean$
$\sigma = Standard\ Deviation$

**Equation 3.1. Formular for the standard score calculation**

As I might not be able to compare the raw scores received from different normal distributions and measures, the standard (z) score was used to compare the scores of each participant (N = 28) from different rating system:

sensitive and strict. Both scores were ranked from smallest to largest in Table 3.26 below.

***Table 3.26. Comparison of raw scores and z-scores of the test on form***

| Sensitive | | | Strict | | |
|---|---|---|---|---|---|
| participant | raw scores | z-scores | participant | raw scores | z-scores |
| 14 | 1 | -1.84404 | **10*** | **1** | **-1.7318** |
| **10*** | **1.5** | **-1.55067** | 14 | 1 | -1.7318 |
| 12 | 2 | -1.2573 | 12 | 2 | -1.16132 |
| 18 | 2 | -1.2573 | 18 | 2 | -1.16132 |
| 28 | 2 | -1.2573 | 28 | 2 | -1.16132 |
| 17 | 3 | -0.67056 | 3 | 3 | -0.59085 |
| 20 | 3 | -0.67056 | **5*** | **3** | **-0.59085** |
| 22 | 3 | -0.67056 | 17 | 3 | -0.59085 |
| 26 | 3 | -0.67056 | 20 | 3 | -0.59085 |
| 3 | 3.5 | -0.37719 | 22 | 3 | -0.59085 |
| 2 | 4 | -0.08382 | 26 | 3 | -0.59085 |
| **5*** | **4** | **-0.08382** | 2 | 4 | -0.02037 |
| 9 | 4 | -0.08382 | **7*** | **4** | **-0.02037** |
| 13 | 4 | -0.08382 | 9 | 4 | -0.02037 |
| 16 | 4 | -0.08382 | 13 | 4 | -0.02037 |
| 24 | 4 | -0.08382 | 16 | 4 | -0.02037 |
| 27 | 4 | -0.08382 | **21*** | **4** | **-0.02037** |
| **7*** | **4.5** | **0.20955** | 24 | 4 | -0.02037 |
| **21*** | **4.5** | **0.20955** | 27 | 4 | -0.02037 |
| 15 | 5 | 0.50292 | 15 | 5 | 0.5501 |
| 25 | 5 | 0.50292 | 25 | 5 | 0.5501 |
| 1 | 6 | 1.08966 | 1 | 6 | 1.12058 |
| 4 | 6 | 1.08966 | 4 | 6 | 1.12058 |
| 6 | 6 | 1.08966 | 6 | 6 | 1.12058 |
| 8 | 6 | 1.08966 | 8 | 6 | 1.12058 |
| 11 | 6 | 1.08966 | 11 | 6 | 1.12058 |
| 19 | 7 | 1.6764 | 19 | 7 | 1.69105 |
| 23 | 8 | 2.26313 | 23 | 8 | 2.26152 |

*Note: a star sign (*) = different ranking and scores due to different rating systems*

Although there might be some differences of scores rated by different rating system in the Table 3.26, both raw score and standard score analysed by Paired-Samples T test showed no statistically significant difference between sensitive rating and strict rating (see Table 3.27). It represented a 98 percent shared variance (0.99 x 0.99) in the performance of the two systems. This means that there was no significant difference in scores graded by different systems.

***Table 3.27. Results of paired-samples T-test between the two scoring schemes***

| | | n | Mean | *SD* | Correlation | Sig.* |
|---|---|---|---|---|---|---|
| **Raw scores** | Sensitive | 28 | 4.14 | 1.70 | 0.99 | **< .01*** |
| | Strict | 28 | 4.04 | 1.75 | | |
| **Z-scores** | Sensitive | 28 | .00 | 1.00 | 0.99 | |
| | Strict | 28 | .00 | 1.00 | | **< .01*** |

*Note: *P < .05*

Because sensitive scoring tended to give more precise findings than strict rating, I used z-scores of the sensitive rating to find possible scores of the high group (top 25%) and low group (lowest 25%) by reversing the formula in Equation 3.2. Before calculation, the z-score table was utilised to identify z-scores of the low and high group at the values 0.25 (low) and 0.75 (high) in the table. The table showed approximate z-scores of -0.67 and 0.67 at the values 0.25 and 0.75, respectively.

$$X = (z \times \sigma) + \mu$$

$x = score$
$\mu = Mean$
$\sigma = Standard\ Deviation$
$z = standard\ score$

**Equation 3.2. A reversed formula of standard score**

When using the formula in Equation 3.2, the possibility of score that the 25% low group received was 3. This means that the possible scores the participants in this group received was equal to or lower than 3, which seems to be normal for the low proficiency group.

On the other hand, a score of 5.5 was the result from the calculation of the 25% high group, meaning that the lowest possible score of the high group was equal to or higher than 5.5. Because the maximum score of this test was 8, the possible scores of the high-performance group tended to range between 5.5 and 8 (out of 10 points). It could be interpreted that the form recall test might be difficult as the top 25% group also got low scores (from 5.5 to 6). The concern led me to pay attention to each individual item to ensure that the test items matched the level of the participants in this study.

Item difficulty, or *p*-value reflects internal consistency of test items. It can be evaluated by calculating the proportion of participants who get a correct answer of an item (see Equation 3.3). The higher the *p*-value of an item, the lower the level of difficulty and vice versa (Green, 2013).

$$P = \frac{R}{T}$$

*P = Item difficulty value*
*R = Number of students who answered an item correctly*
*T = Total number of students*

**Equation 3.3. A formula for item difficulty calculation (Crocker and Algina, 1986)**

In general, item analysis has been widely used to calculate difficulty values of a multiple-choice format. Even though test items used in this study are gap-filling format, the difficulty indexes can give me a direction of how to develop each test item of the test on form. The *p*-value of an item typically varies between 0 and 1. For multiple-choice format, the ideal difficulty index for an appropriate item can be identified by calculating percentage of halfway between pure guess (chance of choosing a correct choice) and total number of test item. For instance, if it is a four multiple-choice test item (25% of chance), the ideal difficulty value for a test of 100 items would be as follow:

$$\text{Ideal } p - \text{value} = 25 + \left(\frac{100 - 25}{2}\right) = 62.5$$

**Equation 3.4. Formula of the ideal difficulty index for an appropriate item**

The optimal value above shows that *p*-value of a test item should not be higher than approximately 0.63 otherwise that item would be too easy (Thompson & Levitov, 1985). However, empirical studies (e.g., Tollefson, 1987; Sabri, 2013; Imorde et al., 2020) have had different criteria to determine optimal level of item difficulty due to different test formats. They suggested that the optimal difficulty values of easy items should range between 0.75 and 0.85. In 1999, McCowan and McCowan suggested levels of optimal difficulty with various number of options for a 100-item test (see Table 3.28).

***Table 3.28. Optimal difficulty values (McCowan and McCowan, 1999, p.19)***

| Number of options | Optimal Difficulty Level |
|---|---|
| 2 | 0.75 |
| 3 | 0.67 |
| 4 | 0.63 |
| 5 | 0.60 |

Furthermore, Oosterhof and Coats's (1984) study found that completion test formats tended to be difficult than multiple-choice formats when the difficulty values were compared. Thus, the appropriate value of a test should be altered based on the test purpose (Green, 2013) and format. In this study, I determined top 30% (0.70) and low 30% (0.30) as criteria for checking item difficulty. Bachman (2004) notes that the selected criteria has been used by many test developers in the field of language assessment. If the difficulty value of an item is equal to or lower than 0.30, that item tends to be too difficult because less than 30% of the participants answer the item correctly. Contrarily, an item might be too easy when the value is equal to or higher than 0.70. The *p*-values of the ten items with interpretations for improvement were presented in Table 3.29.

According to the analysis, some items including Item 6 (*p* = 0.21/0.18), Item 8 (*p* = 0.18/0.18), and Item 10 (*p* = 0.18/0.14) were too difficult and need a revision for improvement while only one item, Item 2 (*p* = 0.68/0.68) tended to be easy. Because lexical variation of all ten items has been checked (see Table 3.22 and Figure 3.4), the degree of difficulty found from individual item analysis should depend largely on the target vocabulary itself. This reflects the word bands (high- and mid-frequency levels) and the extent of the clue (one initial letter) that was provided for each target item in the Pilot Study.

***Table 3.29. Difficulty values of the F-MRt items with interpretation***

| | | **Sensitive**<br>(*p*-value) | **Strict**<br>(*p*-value) | **Interpretation** |
|---|---|---|---|---|
| **Item 1** | concern | 0.64 | 0.64 | *good* |
| **Item 2** | issued | 0.68 | 0.68 | *Somewhat high/quite easy* |
| **Item 3** | shortage | 0.43 | 0.39 | *good* |
| **Item 4** | encouraged | 0.57 | 0.50 | *good* |
| **Item 5** | represent | 0.46 | 0.46 | *good* |
| **Item 6** | strengthen | 0.21 | 0.18 | *too difficult/need revision* |
| **Item 7** | request | 0.39 | 0.36 | *good* |
| **Item 8** | portrayed | 0.18 | 0.18 | *too difficult/need revision* |
| **Item 9** | vendors | 0.50 | 0.50 | *good* |
| **Item 10** | decrees | 0.18 | 0.14 | *too difficult/need revision* |

For example, Item 2: '*The president of the United State i _ _ _ _ _ a statement to the press yesterday.'* contains a target word (issued) that belongs to high frequency level (K1). Because of this, the vocabulary item should not be revised even though its difficulty value is close to 0.70. However, it seems that this item content was irrelevant to the participants' context which is Thai. I decided to change a few words to make it are relevant to the Thai context. The word 'president' was changed to 'Prime Minister' which is a term that also appears in the reading passage of the unit. When developing difficult items (6, 8, and 10) I relied largely on comments from the three native English speaker teachers who gave me some advice on the test earlier. The items were developed by adding more clues (number of letters) and using Thai context (for the Item 8) rather than Europe. Since Item 10 contained a target word of Item 2 (issued), it was possible to replace it with another high frequency word (announced, K2). This might be the reason why the participants got a high score for Item 2. Also, the sentence structure of items 9 and 10 were revised as they had a reduced relative clause. Moreover, it was found from the participants' responses that Item 1 tended to have more than one possible answer as identified by one of the Native English teachers. Some participants filled the word ‘comment’ instead of ‘concern’ in the answer of this item. Besides, the length of words in each item was considered in the development of the test items. I reduced the numbers of words in Item 1 because the original sentence was long.  Hence, there were 6 items developed to ease the level of difficulty and to avoid misleading results as follow:

Original sentences

**Item 1:** *“My **c**_ _ _ _ _ _ is that we are not going to get our group project done in time and we are going to fail.”*

**Item 2:** *“The president of the United State **i**_ _ _ _ _ a statement to the press yesterday.”*

**Item 6:** *“Body builders **s**_ _ _ _ _ _ _ _ _ _ their muscles by lifting weights.”*

**Item 8:** *“Last year, he **p**_ _ _ _ _ _ _ _ Napoleon at the school play about the French Revolution.”*

**Item 9:** *“There are many **v**_ _ _ _ _ _ selling food and drinks to people in Bangkok.”*

**Item 10:** *“The government has issued some **d**_ _ _ _ _ _ setting strict rules on property rights.*

Revised sentences

**Item 1:** *“My **con***_ _ _ _ I am going to fail this class because I missed an exam*.”*

**Item2:** *“The Prime Minister of the Thailand **i**_ _ _ _ _ a statement to the press yesterday.”*

**Item 6:** *“Body builders **st**_ _ _ _ _ _ _ _ _ their muscles by lifting weights.”*

**Item 8:** *“Last year, he **po**_ _ _ _ _ _ _ Naresuan the Great at the school play about the Ayutthaya Kingdom.”*

**Item 9:** *“There are many **v**_ _ _ _ _ _ who sell food and drinks to people in Bangkok.”*

**Item 10:** *“The government has approved some **de**_ _ _ _ _ that set strict rules on property rights.*

Therefore, the target testing words that are contained in the test on form of the F-MRt comprise 6 words (*encouraged, strengthen, shortage, concern, represent*, and *issued*) from high frequency level and 4 words (*vendors, request, decrees,* and *portrayed*) from mid frequency level. I added the two extra words from the high frequency (K1) level to make the participants feel at ease and to encourage them to finish the test because of two reasons. First, the target items were randomly ordered as showed in Table 3.21, and some infrequent words, which seem to be less familiar and more difficult than the high frequent words, appeared in the very first items. This could lead to the loss of attention to complete the whole test. Moreover, the target words used in the previous studies (e.g., Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018; Gohar et al, 2018; Zou & Xie, 2018) usually range between 10 and 16 words (see Section 2.3.5). Because the two extra words were selected from the known-word list resulted from the test of the target word selection (see Table 3.18 in Section 3.7.2), adding two additional known words to the ten-item list does not seem to require a lot of demand to complete the task. However, these two additional items will be excluded from the analysis of this study.

Finally, the new twelve items were analysed again in terms of lexical variation by using VocabProfile to check the levels of the words included in all items, and P_Lex to measure lexical richness of vocabulary. In the new test version, most words are in K1 (80.67%) level, followed by K2 (12.61%), K3 (2.52%), K4 (1.68%), and K5 (0.84%) levels, respectively (see Table 3.30).

***Table 3.30. Word levels of the developed items of F-MRt***

| Freq. Level | Families (%) | Types (%) | Tokens (%) | Cumul.token |
|---|---|---|---|---|
| K1 | 87 (80.60) | 96 (80.67) | 134 (85.40) | 85.40 |
| K2 | 15 (13.90) | 15 (12.61) | 15 (9.60) | 95.00 |
| K3 | 3 (2.80) | 3 (2.52) | 3 (1.90) | 96.90 |
| K4 | 2 (1.90) | 2 (1.68) | 2 (1.30) | 98.20 |
| K5 | 1 (0.90) | 1 (0.84) | 1 (0.60) | 98.80 |
| Off-List | - | 2 (1.68) | 2 (1.27) | 100 |

The off-list words in the text are *Bangkok* and *Supermodel*. All words from mid-frequency level (K3-K5) presented in the result of the analysis are the target words to be tested. The low lambda score (0.93), which usually range between 0 and 4.50, from the P_Lex analysis of the revised version showed that the test consists of low proportion of infrequent words (Meara and Bell, 2001). The Pilot Study also showed that the participants needed at least one hour to complete the test of ten items. For this reason, I decided to provide one and a half hours for the participants in the Main Study to do this twelve-item test.

***Table 3.31. Ten target words and test items for the Main Study***

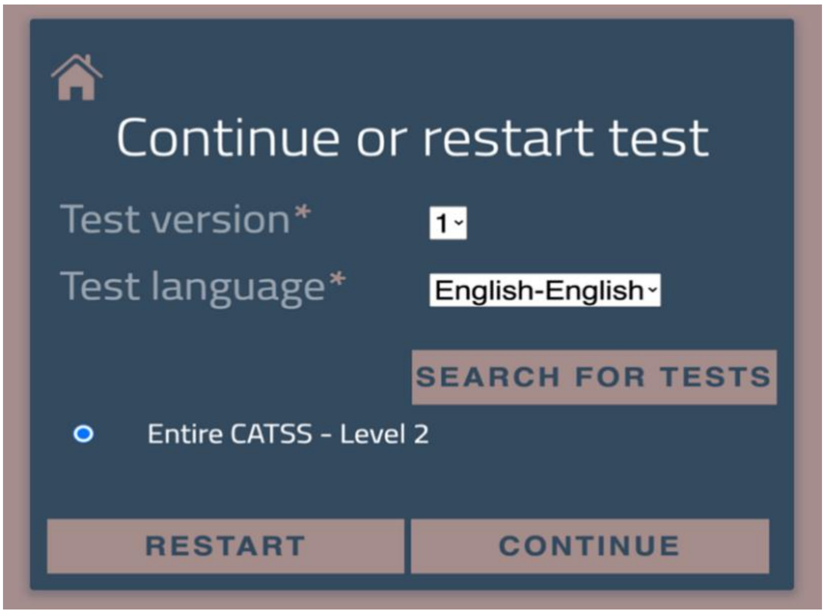
| Target words | | Test item |
|---|---|---|
| 1 | Concern (noun/K1) | My **con**_ _ _ _ I am going to fail this class because I missed an exam. |
| 2 | Issued (verb/K1) | The Prime Minister of the Thailand **i**_ _ _ _ _ a statement to the press yesterday. |
| 3 | Shortage (noun/K1) | There was a serious **sh**_ _ _ _ _ _ of water last year due to the unusually long hot summer. |
| 4 | Encouraged (verb/K2) | My mother has always **e**_ _ _ _ _ _ _ _ _ me to follow my dream of becoming a supermodel. |
| 5 | Represent (verb/K2) | These five colours on the map **re**_ _ _ _ _ _ _ five different countries. |
| 6 | Strengthen (verb/K2) | Body builders **st**_ _ _ _ _ _ _ _ their muscles by lifting weights. |
| 7 | Request (noun/K3) | My sister is making a **r**_ _ _ _ _ _ for the band to play her favorite song. |
| 8 | Portrayed (verb/K4) | Last year, he **po**_ _ _ _ _ _ _ Naresuan the Great at the school play about the Ayutthaya Kingdom. |
| 9 | Vendors (noun/K4) | There are many **v**_ _ _ _ _ _ who sell food and drinks to people in Bangkok. |
| 10 | Decrees (noun/K5) | The government has announced some **de**_ _ _ _ _ that set strict rules on property rights. |
| **Extra words** | | **Test item** |
| 11 | Begin (verb/K1) | All of my classes this term **b**_ _ _ _ at 9 o'clock in the morning. |
| 12 | Protect (verb/K1) | Wearing face masks and washing hands can **pro** _ _ _ _ viruses from spreading. |

The items in Table 3.31 were later used in the Main Study, but the order of each item was different depending on the randomisation patterns showed in Table 3.21 above. To conclude, the Pilot Study found that the participants did not have background knowledge of the selected ten words aiming to use as the target words in the Main Study as presented in Table 3.18. These ten words include words from high frequency bands (K1 and K2) and mid-frequency bands (K3, K4, and K5). The data analysis and findings of the Main Study are presented in Chapter 4 and Chapter 5, respectively to answer the research questions.

## 3.8 Lessons from the Pilot Study

The Pilot Study helps to validate the designed instruments for the Main Study so that the main analysis results can be reliable for interpretation. It also give information as a guideline for the preparation and distribution of tests which help to inform the design of the Main Study. During the process of preparation, I realised several issues arisen from the new CATSS and the F-MRt.

First, the new CATSS should be adapted to match the research design. In the Pilot Study, I tried out the online CATSS to check if I could do only Level 1 and Level 2 of the provided online test from the Lextutor website. There were several problems that I found from the online test. First, the website required new users to provide their information such as email address, first language, grade, and gender in order to create a new account before logging into the online test. Using this process for the online CATSS could take time and might result in loss of interest of the participants to join in this study. Also, the users' personal information can be seen by the web developers. According to the ethics approval, participants' personal information must be protected. Furthermore, even though the website allows test takers to continue or restart a previous test they did, they could not select other modalities or tests if they choose the 'Continue' option (see Figure 3.11).



Continue or restart test

Test version* 1

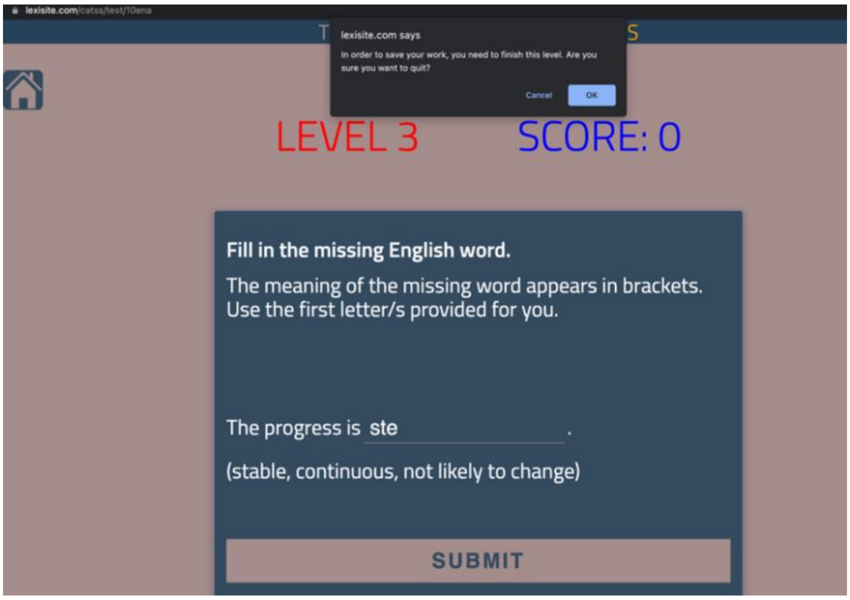
Test language* English-English

● Entire CATSS - Level 2

**Figure 3.11. A problem on continuing the previous test of the CATSS**

Based on the trials, there was a bar presented on the right hand side of the screen (see Figure 3.11) that I could search for tests, but there was no choice for me to choose after I tried to click on it. I could only select the Entire CATSS-LEVEL 2 as shown in the Figure 3.11 in order to continue doing the test because I had finished the Level 1 of the productive recall test which was the first modality presented on the Entire CATSS option.

Importantly, it is not convenient for test takers to do only the first two levels of the four modalities (receptive recognition, productive recognition, receptive recall, and productive recall). Based on my experience, I was requested to log in again every time I tried to change the modality. After I finished Level 1 and Level 2 of the productive recall test which was the first modality presenting in the Entire CATSS option and tried to change the modality by clicking at the 'Home' icon on the left top corner of the page (see Figure 3.12), it showed on the top of the screen that I had to continue doing the next level.



lexisite.com/catss/test/10ena

lexisite.com says

In order to save your work, you need to finish this level. Are you sure you want to quit?

LEVEL 3 SCORE: 0

Fill in the missing English word.

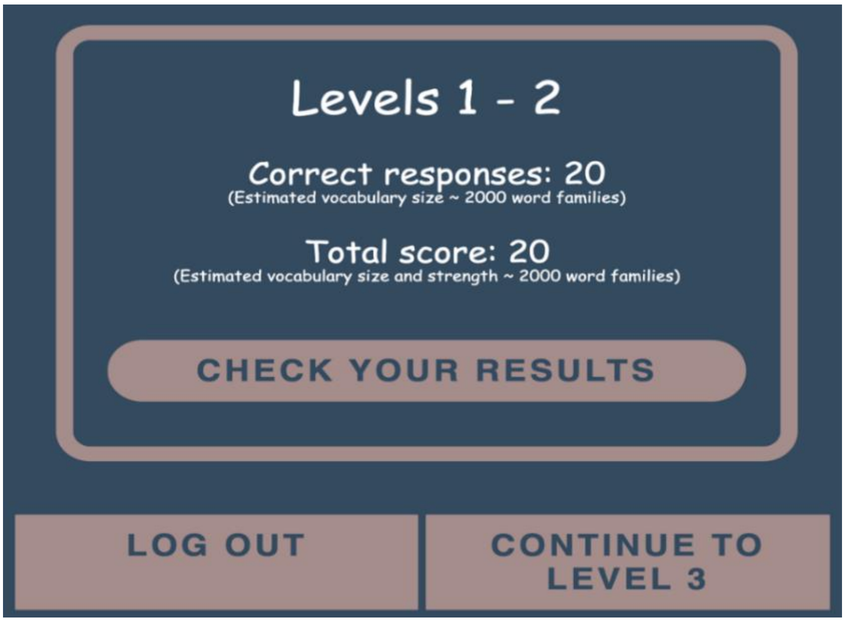
The meaning of the missing word appears in brackets. Use the first letter/s provided for you.

The progress is ste .

(stable, continuous, not likely to change)

**Figure 3.12. A problem on an attempt to change the modality of the CATSS**

After I clicked 'ok,' it brought me to the window shown in Figure 3.13 to continue doing the next level. Because I had an intention to do only the first two levels of the four modalities, I then clicked at the 'Home' icon, and it brought me to the log in window again. Although the website allows users to select each modality of the available tests, it tends to be inconvenient for the participants of this study to put several attempts to logging into the website.



Levels 1 - 2

Correct responses: 20
(Estimated vocabulary size ~ 2000 word families)

Total score: 20
(Estimated vocabulary size and strength ~ 2000 word families)

**Figure 3.13. A problem on changing the modality of the CATSS after clicking at 'ok'**

It seems that the participants have to finish all 14 levels of each modality to avoid the process of multiple logging in, which is time-consuming and unfeasible due to time limitation during class. Also, this can cause frustration among the participants and affect the test results. Finally, I found another limitation to collect scores of the four modalities from the website. If a test user gets a correct answer of the word in the receptive recall modality, which was offered as the first modality in the entire CATSS test, that word will not be presented again in the other tests as receptive recall is claimed by the test developers to be the most difficult modality (Aviad-Levitzky et al., 2019). Accordingly, I would not be able to collect the score from the receptive recall, productive recognition, and receptive recognition of all items unless the participants did not give correct answers to every item in the productive recall test.

Due to the above reasons, it seems unfeasible to use the CATSS that is available online on the Lextutor (https://www.lextutor.ca/tests/). I had to adapt the test by eliciting only test items from Level 1 and Level 2 of the four modalities of the CATSS to develop an online test. Microsoft Form was used to create this online test. The adapted test has the same formats and sequences of test modalities as the original CATSS from the website, except for the words in each level. This is because the words in the original CATSS were shuffled every time I tried to do the test of each modality. So, in the present study the words in each level of the adapted CATSS were randomly ordered. The test was designed based on the original new CATSS, so started from productive recall, followed by receptive recall, productive recognition and receptive recognition, respectively. There are 10 items for one level which means the adapted CATSS contains 20 items for each modality or 80 items in total.

Moreover, with regard to the F-MRt, some concerning issues have been arisen when constructing the test prior to the implementation in the Pilot Study. This was related to the format of the test. Because the PVLT (see also Figure 3.8) can be used as a diagnostic test (Laufer and Nation, 1990), it has been used widely among scholars (e.g., Laufer; 1998; Stæhr, 2008; Abdullah et al., 2013; Henriksen & Danelund, 2015). Test takers are asked to complete the missing letters in order to form a correct word such as the word ‘episodes’ in Figure

3.8 above, to match the provided context. The number of initial letters used as a cue varies depending on the length of each word and the possibility of alternative words that can be used in each sentence. According to Laufer and Nation (1990, p. 37), “If two letters could start two possible words in the given sentence, an additional letter was added to eliminate this possibility.” However, to lower any possibility to guess the target word without understanding the context, about half of the letters in a word tends to be removed. However, the format can be problematic because some possible words from different frequency levels might suit the context of some items. Jonathan (2010, p.32) illustrated an example of a test item *“Her beauty and cha______ had a powerful effect on men”* (Laufer and Nation, 1999, p 46) which has several possible answers. While the suggested answer is the word *charm* from high frequency word level*,* high English proficiency learners might fill-in other possible words in lower frequency levels such as *chatter* or *charisma* to make the sentence meaningful, resulting in failure in doing this item. Since there is no established test that can perfectly measure productive vocabulary knowledge in all aspects, I drew on the test formats of the VKS and PVLT since Scale IV of the VKS, which requires L1 translation, could measure meaning recall while the PVLT could elicit form recall. Due to this issue, the controlled productive vocabulary test or Form-Meaning Recall test was designed to include both initial letter(s) and dashes as a clue for each item. This is to control for the other possible answers and to match the aim of the current study.

## 3.9 Main analytic method and tool

Analyses of variance (ANOVAs) have been commonly used in second language research (Cunning, 2012), and studies related to the TFA and word recall (e.g., Keating, 2008; Hu & Nassaji, 2016; Khoshsima & Eskandari, 2017; Chaharlang et al., 2018) to compare the results of Pre-test and posttests among three or more study groups. A one-way ANOVA can be used to report the analysis of the collected data to compare the differences among the six groups of the participants within each test in the current study. However, the main aim of this study is to investigate the effects of the TFA framework on word gain and retention from the Pre-test, immediate-posttest, and delayed-posttest. Other inferential statistics such as Mixed Design ANOVA, General Linear Models (GLM) and Mixed-Effects Models (MEM) seem to be more useful to explore the significant differences from the repeated measurement (Pre-test, Immediate Posttest, and Delayed Posttest) collected from the six groups. This is said to be a '*hierarchical data structure*' (Field et. al., 2012, p.857). Figure 3.14 provides the example of the hierarchical data structure of the *Control* group ($n$ = 38) in which variables (i.e., Pre-test and posttests) are clustered, or nested within other variables (i.e., participants in the group). Test scores from the repeated measure of the Form-Meaning Recall test (testing time variable) depends on each participant, ranging from id 1 to id 38 (individual participant variable). These participants are clustered within the same group (Control) to compare with the experimental groups. As the participants were treated with different treatments, they are clustered or nested within six groups: *Control, Motivation, Noticing, Retrieval, Generative Use* and *All TFA Components* (group variable or learning factors). Each group had the same structure as Group 1 (Control group) presented in Figure 3.14, but difference in number of the participants within group. This led to the need of multilevel models to investigate whether vocabulary activities (treatments) used with the participants in six groups leads to different test results.

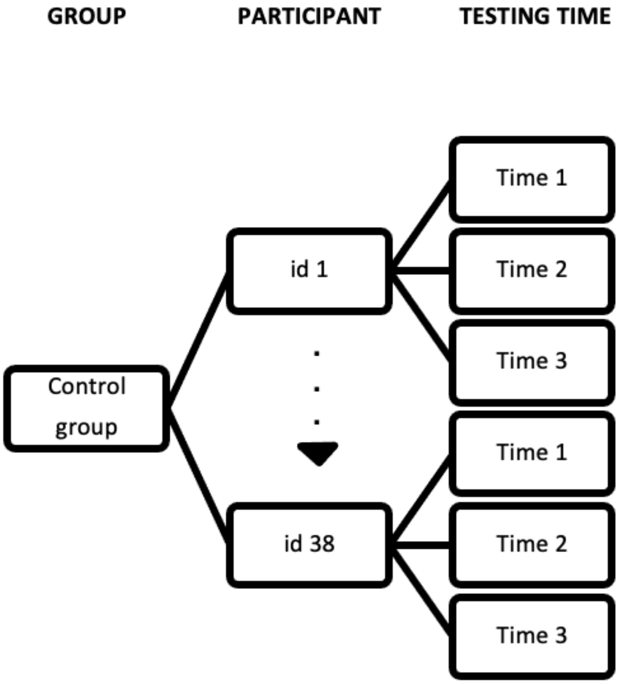


**Figure 3.14. The hierarchical data structure of Group 1 of the current study**

The decision regarding the selection of method and model for data analysis of the current study relied heavily on two main reasons. First, it should help to analyse hierarchical data due to the design of the study (see Figure 3.14). The suitable method that matches the research design helps to answer the research questions accurately. Second, the analytic method should meet the assumptions. For each method, there are assumptions to be checked. If the data of the current study violates the assumptions, the interpretation can be biased (Field et. al., 2012; Field & Wilcox, 2017). Basic assumptions such as normality distribution of data, homogeneity of variance, multicollinearities (for linear models), etc. were checked before selecting the analytic method. If the data violates the assumptions, non-parametric statistics such as Generalized Linear Model and robust tests should be used for more accurate interpretation. However, many researchers (e.g., Field, et. al., 2012; Winter, 2020) suggested that robust tests, centering variables and data transformation processes such as log transformation, square root transformation, and reverse score transformation are only necessary when the data violates the assumptions.

As mentioned earlier, there are three statistical analysis methods that seem to be useful for the current study: Mixed Design ANOVA, General Linear Models and Linear Mixed-effects Models, which was developed from the General Linear Models (Field et al., 2012). Because the main data violated some assumptions such as normal distribution, several solutions were compared (see Section 4.5.3.2 and Section 4.5.3.3 below for details). The problem was mainly because the data contained the Form-Meaning Recall test scores of all six groups, consisting of both control and experimental groups. Based on the hypotheses of the current study, each group were expected to have similar result within groups but different results between groups across the three testing time periods due to the effects of treatments. Therefore, it was likely that the data taken from all scores of the six groups would affect a distribution of the data and violate the homogeneity of variance. While Generalized Linear Model, which is a non-parametric statistic, can deal with this problem, it is less good at addressing the hierarchical research design of this study, as presented in Figure 3.14, than other statistical methods such as Mixed Design ANOVA and Mixed-effects Models. According to Field and colleagues (2012, p. 193), “If a statistical model is still accurate even when its assumptions are broken it is said to be a robust test.” Because there is no non-parametric alternative for Mixed Design ANOVA and Mixed-effects Models (Pallant, 2001), other techniques including robust tests from several packages such as “WRS2” (Mair & Wilcox, 2020) and “robustlmm” (Koller, 2016) in R should be used (see details in Section 4.5.3) when assumptions are violated (Field & Wilcox, 2017; Mair & Wilcox, 2020) in order to get more valid results. So, a comparison between the original or classic Linear Mixed-effects model and the robust estimation model was compared in the current project (see discussion in Section 4.5.3.3). As argued by several researchers (Glass et al., 1972; Games, 1983; Field et al., 2012), data transformation might not always be the best choice as it changes the original means and skewed distributions that can lead to negative consequences of analysis. The reason for using transformations therefore depended on how well it can improve the data and the selected model to perform better. However, the original model showed a better fit than the robust model and so the current study relied on the original Linear Mixed-effects Model (see discussion in Section 4.5.3.2 and Section 4.5.3.3).

When compared to other inferential statistics for multilevel structure design, Mixed-effects Models, or linear mixed-effects regression modelling allows researchers to explore the ‘relationship of interest’ (Winter, 2013, p. 2) or ‘independent variable of interest’ (Cunning, 2012, p. 370) between variables that can affect the test results of the experiment. Brown (2021, p. 2) stated that “Mixed-effects modeling allows a researcher to examine the condition of interest while also taking into account variability within and across participants and items simultaneously.” The models have been widely used in psycholinguistics and linguistics research (Gries, 2015). Because a Mixed-effects Model pays attention to not only fixed effects (i.e., teaching methods, treatments, testing time: Pre-test and posttests), but also random effects (i.e., age, sex or educational background) that may arise from the experiment, it offered a useful analytic method for the present study (see more arguments in Section 4.5.3). As mentioned earlier, the “lme4” package (Bates et al., 2015) for multilevel models in R allows researchers to apply the function lmer() for the classic mixed-effects model. So, it was utilised as a main analytical tool in the present study. Where there is a statistically significant difference among them, a post-hoc test which is a Tukey’s Honest Significant Difference (HSD) was employed to identify the differences.

ANOVAs and Mixed-effects Models can be implemented in various analytic tools such as SPSS and R programme. The main analysis of the current study only used R, which is a free downloadable statistical package from http://www.r-project.org to analyse the collected data because of two main reasons. First, SPSS requires users to pay for the license in order to use the programme. Apart from its expensive cost, scholars (e.g., Milovanović and Perišić, 2020; Winter, 2020) asserted that SPSS might not be appropriate to use with complex statistics that required programable coding with functions for data analysis. It is a ‘menu driven’ tool (Cunning, 2012, p. 372) which is not as flexible in terms of generating statistical commands as other packages such as R. This means that R allows users to modify codes for data analysis. Because a single linear mixed-effects model was appropriate to use as the main analytic tool to investigate the findings of the present study, R tends to be more useful than SPSS. This way, I could observe the variables that might have effects on the test results and be able to interpret the findings more accurately.

# Chapter 4
# Main Study

The analysis of the Pilot Study in the previous chapter helps to shape the main research methods and prepare instruments used in the Main Study. This chapter gives information of the main study regarding the sample size in Section 4.1, target learning words in Section 4.2 and vocabulary learning activities in Section 4.3. In Section 4.4, it covers research instruments which are a test for the participants' prior vocabulary knowledge (CATSS), the F-MRt for measuring productive knowledge and post-treatment questionnaire. This chapter also explains procedures of the Main Study in Section 4.5 before describing the analysis of the adapted CATSS and F-MRt. Data from the experiment include (a) the scores from the CATSS to measure the participant's prior knowledge of general vocabulary before the treatment and (b) the scores on the Pre-test, Immediate Posttest, and Delayed Posttest of the Form-Meaning Recall test (F-MRt) to measure the participants' productive knowledge of the target words before and after the treatment. These data sets would serve as the main data set to answer the research questions. Data from the post-treatment questionnaire was the supplementary data to explore the in-depth information and compare the results with the F-MRt data. This chapter will describe the analysis of each data set in detail. The test analyses start with the analysis of the CATSS in Section 4.6, followed by the statistical analysis of the F-MRt and questionnaire in Section 4.7. Data analysis results are not included in this chapter. They are explained in the Chapter 5: Main Study Findings.

## 4.1 Participants

The main study was conducted in the same context as the Pilot Study. The participants in the main study were 247 Thai undergraduate students from six sections of the University English II course. There were 58 males and 172 females aged between 18-20 years old. There were 17 students that preferred not to say their gender. In my research context, students are typically grouped into classes with about 40-50 students in each section at the beginning of each term. By using the results of my pilot study, the lowest acceptable sample size was calculated through G-power, a programme for power analysis, at 111. If there are 40 students in each class, the total number of the participants from the six interventions will be approximately 240. This figure ensures the study had adequate power to answer the research questions. Although the number of participants changed during data collection, it was sufficient to interpret the results given the 111 figure from the power calculation.

The language proficiency of the participants in the main study is similar to that of the participants in the Pilot Study (see also Section 3.2) because they were taking the same course offered by the Language Institute of Thammasat University. These participants are regarded as intermediate level students because the participants' level of English proficiency is equal to the immediate level as mentioned earlier in section 3.2. Based on my experience in teaching students in this course, I have reasonable knowledge about language proficiency and learning obstacles of the students. However, I checked their English proficiency scores and requested the students in this class from two different semesters to take a vocabulary test again to explore the level of vocabulary knowledge. In terms of vocabulary knowledge, the Pilot Study indicated that the students tend to have sufficient receptive knowledge of words at high frequency levels based on the CATSS results (see details in Section 3.7.1). This is similar to the result obtained from the participants in the main study (see more details in Section 5.1 below). Because the CATSS results from both the Pilot Study and Main Study showed that the students in this class have similar background knowledge in vocabulary that suits the purpose of my study, I used the purposive sampling method to select six groups of participants to be included in this study for the time-management purpose. For logistical reasons, I could not collect data from two groups at the same time. Therefore, I had to select six groups that were scheduled on a different time in the same week to be able to do the experiments.

## 4.2 Target words

Based on the Pilot Study (details are given in Section 3.5 of Chapter 4), ten unknown words were selected as the target words in this study. These words are ***concern*** (K1)*,* ***issued*** (K1)*,* ***shortage*** (K1)*,* ***encouraged*** (K2)*,* ***represent*** (K2)*,* ***strengthen*** (K2)*,* ***request*** (K3)*,* ***portrayed*** (K4)*,* ***vendors*** (K4)*,* and

***decrees*** (K5). Some words in the list involve grammatical features (plural form and past participle), but the result of the Pilot Study indicated that majority of the participants did not know the meaning of these words (see also Section 3.7.2). In addition, the Form-Meaning Recall test from the Pilot Study did not reveal any problems regarding these grammatical features due to clues given to help with the grammars (see discussion in Section 3.7.2). For these reasons, I decided to include them in this study for further insights into the effects of TFA components on retention of form with various features.

## 4.3 Vocabulary learning activities for the treatment

Detailed descriptions of the development of these activities have been provided in Section 3.4. This section will briefly describe them. There are five experimental groups and one control group in the Main Study. So, five vocabulary activities were designed to control for *Motivation, Noticing, Retrieval* and *Generative Use* as presented in Table 4.1. Group 2 was rated as high for *Motivation* while Group 3, 4 and 5 were rated as high for *Noticing, Retrieval* and *Generative Use*, respectively (see Section 3.4 for details). This way, the predictive power of each TFA component can be compared to address the Research Question 2.

***Table 4.1. Five experimental groups and designed activities with TFA scores***

| Groups | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| **TFA Focus** | *Motivation* | *Noticing* | *Retrieval* | *Generative Use* | *All TFA Components* |
| **Designed Activities** | Reading plus fill in | Reading and using glosses | Reading and identifying word parts | Reading, word parts, and sentences writing | Word cards and writing using target words |
| **TFA Score** | **TFA = 7** | **TFA = 7** | **TFA = 8** | **TFA = 9** | **TFA = 15** |

*Note: Group 1 is the Control group*

For Group 6, word cards and writing using target words activities are combined to give high support to all TFA components: *Motivation, Noticing,* *Retrieval* and *Generative Use.* It was designed to compare different TFA scores from low (Control group) to moderate (Group 2 – Group 5) and high (Group 6: TFA = 15) with the F-MRt results in order to answer the Research Question 1. This will help to verify the overall effectiveness of the TFA framework as argued by previous studies (see Section 2.3 for detail). Six lesson plans for all groups and supplementary materials for the five designed activities were presented in Appendix 2 and Appendix 3, respectively.

## 4.4 Main research instruments

This section explains how I administered the research instruments to discover the participants' vocabulary knowledge in the Main Study. The four research instruments comprise 1) the adapted CATSS, 2) controlled productive vocabulary test, called Form-Meaning Recall test (or F-MRt) that was developed in the Pilot Study, and 3) a questionnaire.

***Table 4.2. Research questions, instruments, and groups of participants***

| RESEARCH QUESTIONS | INSTRUMENTS | GROUPS OF PARTICIPANTS |
|---|---|---|
| **RQ1.** Do activities with high TFA scores result in better retention of single words in productive form recall? | 1.) a Pre-test<br>2.) treatments: activities with low (or zero) and high TFA scores<br>3.) an Immediate Posttest<br>4.) a Delayed Posttest | 1 control group and 5 experimental groups |
| **RQ2.** What is the individual role of the different TFA components on promoting vocabulary retention through vocabulary tasks? | 1.) a Pre-test<br>2.) treatments: activities with high scores in four different TFA components<br>3.) an Immediate Posttest<br>4.) a Delayed Posttest | 4 experimental groups: *Motivation, Noting, Retrieval, Generative Use* |

Table 4.2 shows the research tools which include five vocabulary activities for the five intervention groups and the different versions of the F-MRt test used with all six groups to help answer the research questions of this study.  I first

provided rationales for using the adapted CATSS in the Main Study and gave details of the developed in Section 4.4.1 and 4.4.2, respectively. Then, I present details of a questionnaire in Section 4.4.3. The information about three versions of the F-MRt used as Pre-test, Immediate Posttest, and Delayed Posttest was given in Section 3.7.3 (see also Section 4.5) while the rationales for developing the F-MRt were covered in Section 3.8 above.

### 4.4.1 The adapted CATSS

As explained in Section 3.7 (Pilot Study), the adapted CATSS was used with the purpose to identify the participants' knowledge of vocabulary in terms of productive recall, productive recognition, receptive recall, and receptive recognition at high frequency word level. In the Main Study, the test was implemented one week after the Delayed Posttest with the six groups of participants.

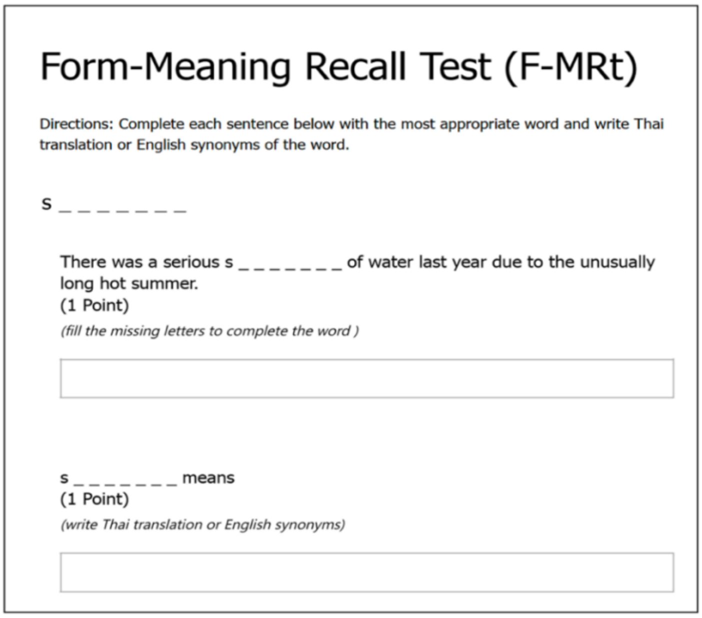
The adapted CATSS was administered after the F-MRt for two main reasons. The first issue related to the time required for the participants to do the CATSS before the experiment. The experiment used teaching material of the Unit 1 and this unit was scheduled during the first two weeks in the course outline for all sections in this course (see details in Section 3.3 and Section 3.4).

During the first week I also had to cover some ethical issues. I had sent the Information Sheet and Consent Form (see Appendix 5b) to the participants one week before the course starts. However, in the first class I needed to spend one hour describing details of the course to the participants before giving them the Consent Form as well as the Pre-test (F-MRt) to do in class. This took about 2 hours.

In the following week, the participants in the six groups had to study the Unit 1 material based on the lesson plan designed for each group and take a one-and-a-half-hour Immediate Posttest. This class lasted for three hours while the CATSS requires at least an hour to complete. This would impose a great burden on the participants so the test could not be used during the first two weeks. The second reason was the potential impact of the adapted CATSS on the Delayed Posttest. Words that appeared in the CATSS could lead to misleading results of the Delayed posttest on Week 4. Lastly, the participants might lose attention in doing the Delayed Posttest if they were required to take a long test every week from the beginning of the course. There needed to be a space between the Immediate Posttest and Delayed Posttest for the participants. As the network building of vocabulary knowledge is slow (Haastrup & Henriksen, 2000) and productive vocabulary size tends to grow slower than receptive vocabulary size (Laufer & Paribakht, 1998; Laufer, 1998; Laufer & Goldstein 2004; Webb 2008; Zhong & Hirsh 2009), within three weeks of the interventions it can be predicted that the use of the adapted CATSS after implementing the F-MRt would not have a significant impact on its result. For these reasons, the CATSS was distributed to the participants during Week 5, one week after implementing the Delayed Posttest.

### 4.4.2 A controlled poductive vocabulary test (or F-MRt)

A Form-Meaning Recall test, or F-MRt is a controlled productive vocabulary test that was used as a Pre-test, Immediate Posttest, and Delayed Posttest in the Main Study. It was adapted from two existing tests: Productive Vocabulary Level Test, or PVLT (Laufer & Nation, 1990) and Vocabulary Knowledge Scale, or VKS (Paribakht & Wesche, 1993; 1997) due to a lack of productive vocabulary test that matches the design of the present study. As the design of the test was explained in Section 3.7.3 and Section 3.8, I only provided the format of the test in this section. A sample item of the online F-MRt is presented in Figure 4.1 below (see the full version of the F-MRt in Appendix 6).



Form-Meaning Recall Test (F-MRt)

Directions: Complete each sentence below with the most appropriate word and write Thai translation or English synonyms of the word.

s _ _ _ _ _ _ _

There was a serious s _ _ _ _ _ _ _ of water last year due to the unusually long hot summer.
(1 Point)
*(fill the missing letters to complete the word )*

s _ _ _ _ _ _ _ means
(1 Point)
*(write Thai translation or English synonyms)*

**Figure 4.1. A sample item of the online F-MRt**

The test contains 12 items which include ten target words and 2 extra words (see also discussion on the test format in Section 3.7.3). Each item consists of two sub-items. The first sub-item measures form recall while the second sub-item aims at eliciting vocabulary knowledge in terms of meaning recall.

### 4.4.3 A questionnaire

Due to the effect of small class size and a concern in the evaluation of support for motivation, I decided to include a questionnaire containing general questions about the participants' biodata and learning experience. Similar to the experiment, it has to be an online questionnaire because all classes were discouraged from meeting face-to-face until November 2021. The questionnaire (see Appendix 7) contains 7 to 8 items in both Thai (L1) and English (L2). The items in the questionnaire include:

1. How old are you? ____________
2. What gender do you identify as?
   - ❑ Female
   - ❑ Male
   - ❑ Prefer not to say
   - ❑ Other________________
3. How long have you been studying English? _____________
4. Please enter one of your English proficiency scores such as O-NET score for English or TU-GET/IELTS/TOEFL score that you have got recently <u>with the name of the test</u> (*For example, O-NET=55*)

   ____________________
5. Have you taken TU105 course before? **<u>*Note: if 'no' Question 6 will be skipped.*</u>*
   - ❑ Yes
   - ❑ No
6. When did you take this course?
   - ❑ Summer course 20/21
   - ❑ Semester 2-20/21
   - ❑ Semester 1-20/21
   - ❑ Before semester 1-20/21

7. How did you feel with learning today?
   - ❑ enjoyed
   - ❑ challenged
   - ❑ enjoyed and challenged
   - ❑ neither enjoyed or challenged
8. How did you feel with learning by using activities or supplementary handouts in the class today?
   - ❑ 1 = not at all interested
   - ❑ 2 = not very interested
   - ❑ 3 = neutral
   - ❑ 4 = somewhat interested
   - ❑ 5 = very interested

In conclusion, there were four research instruments in the Main Study. Prior to the experiment, the participants' background knowledge of vocabulary was checked through the adapted CATSS. Then, the proposed vocabulary activities explained in Section 3.4 and Section 4.3 were used with the five intervention groups: *Motivation, Noticing, Retrieval, Generative Use*, and *All TFA Components*. There was no supplementary activity and handout provided for the *Control* group (Group 1). However, all six groups were required to take the same test which is the F-MRt before and after the learning. At the end of each class, the biodata questionnaire, followed by the developed F-MRt were implemented with the participants. The F-MRt was used three times during the experiment to measure short-term and long-term retention on word form and its meaning. Section 4.5 below will explain procedures of the implementation.

## 4.5 Procedures

There were four visits for each group, three hours for a visit. All visits were held through an online platform, Microsoft Teams. Unlike on-site learning, I am aware that online learning might not be easy to manage and control. Therefore, I required the participants to turn on their camera and encouraged them to pay attention to the activities in class as the learning was designed based on the content that would eventually be tested in the mid-term or final exam. This should be able to raise awareness of the participants in learning.

After the participants agreed to participate in this research, they were required to meet via Microsoft Teams because they had the university email that was registered for them to be able to access to the programme. To seek for answers of the two research questions (see Section 2.5), the Form-Meaning Recall test (F-MRt) that has been validated (see Section 3.7.3 for detail) was distributed with the participants in the Main Study one time before and two times after the interventions. The administration of all research instruments of the main study is shown in Table 4.3.

***Table 4.3. The implementation of the main research instruments***

| Week | Plan/Activity | Time/Duration |
|---|---|---|
| 1 | **Pre-test:** a twelve-item F-MRt | One and a half hours before the lesson |
| 2 | **Experiment** (learning with different activities) | One and a half hours at the beginning of class |
| | **Biodata questionnaire** | 15 minutes before taking the Immediate Posttest |
| | **Immediate Posttest:** a twelve-item F-MRt<br>(shuffled items) | One and a half hours after the experiment |
| 3 | **No visit** | |
| 4 | **Delayed Posttest:** a twelve-item F-MRt (shuffled items) | One and a half hours before the lesson |

In the beginning of August 2021, the participants at Thammasat University in Thailand had to take a F-MRt as a pre-test one week before the experiment. A week later, they had to learn different vocabulary activities based on the purpose and design of this study as mentioned in Section 3.4 and Section 4.3. To control for variation of the teaching and learning in this current study, I taught all the participants myself. After the learning, the participants in each class received a link of the questionnaire at the end of the lesson via Microsoft Teams. It took approximately 15 minutes to complete 8 questionnaire items (see Appendix 7). The participants might not be able to recall their actual feelings when learning with different activities if they have to do it after taking a long (one-and-a-half-hour) test. Therefore, this questionnaire was administered 15 minutes before taking an immediate F-MRt. After it was done, the link led the participants to the test. They were asked to take an immediate F-MRt for one and a half hours (see the discussion of test duration in Section 3.7.3). This is to measure vocabulary gains and short-term retention of the ten target words. However, the test was reordered to clear the participants' memory.

The Delayed Posttest can be designed for days, weeks, or months after the experiment (Nation & Webb, 2011). Although the duration of a delayed Posttest seems to vary, it usually ranges between one or two weeks (Avila & Sadoski, 1996; Waring, 1997a; Keating, 2008) after the experiment. A study (De Vos et. al., 2017) found significant differences of decay knowledge between twenty-minute Delayed Posttest and six-months Delayed Posttest. However, previous recent research (i.e., Johanna et al., 2019) suggested that a longer delayed posttest could reveal decay of knowledge better than few minutes or a few days delayed posttest. It helps to ensure long-term retention of word knowledge. Because memory seems to decay over time, after two weeks of the interventions, the participants had to sit the F-MRt test with different order of items again in order to assess long-term vocabulary retention from the delayed Posttest.

## 4.6 Analysis of the adapted CATSS

According to Aviad-Levitzky, et. al., (2019, p. 348), "[The CATSS] reflects the different difficulty associated with the degrees of strength of knowledge…" Scoring depends on levels of difficulty of each test type: recognition and recall. The highest score (one point) is given to form recall items while lower scores: 0.75, 0.5, and 0.25 are given to meaning recall, form recognition and meaning recognition items, respectively. In the current study, it seems more appropriate to give one point, which was used in the validation of the new CATSS by Aviad-Levitzky and colleagues (2019), to each test item regardless of the modality. This way, the score from each modality can be compared. Besides, this test was adapted to use in this study for the purpose of participant recruitment of the main study. Strict scoring can help me to identify the actual problems that the participants tend to be facing with.

In the current study, the adapted CATSS was used to check the participants' prior knowledge of vocabulary in terms of meaning recall, form recall, meaning recognition, form recognition at high frequency word level. This section will provide details of the scoring and interpretation process together with the results from the descriptive statistics analysis (mean, standard division and mode) and the inferential statistical analysis (One-way between groups ANOVA) to compare the CATSS scores of the six groups of participants.

In term of scoring, once the participants had completed the adapted CATSS, I first marked each modality of the CATSS by giving one point for correct answers and zero point to any incorrect answers. I used the same scoring criteria used in the Pilot Study (see discussion in Section 3.7.7.1). As suggested by scholars (e.g., Nation, 2008; Van Zeeland & Schmitt, 2013; Sudarman & Chinokul, 2018), the mastery levels for spoken text and written text are set at 90% and 80%, respectively. Since I focused on the production of written words rather than spoken words, the acceptable level of mastery used in the current study is 80%, which is the cut-off point for mastery adopted used in previous studies (e.g., Dang et al., 2000, Hun & Nation, 2000; Xing & Fulcher, 2007; Rogers, 2013; Hacking et al., 2017; Dang et al., 2021). After I had completed the first marking, the data that passed the mastery level were second marked by another experienced teacher to check reliability of the result. Once the scoring had been finalised, the data were input in Excel spreadsheet for statistical analyses. The R programme (R Core Team, 2021) was used for the analysis for consistency with the analysis of the Form-Meaning Recall test scores (see also Section 3.9 and Section 4.7.3 for the reasons of using R for the analysis). First, the 'psych' package (Revelle, 2022) was used to explore descriptive statistics of the prior general vocabulary knowledge of each group of participants. Then, the R base packages from R version 4.1.2 (R Core Team, 2021) were used to perform one-way between groups ANOVA to identify whether there were any significant differences among these groups. Analysis was done for CATSS overall scores as well as the scores on each modality of the test. Before running one-way between groups ANOVA, I also used the Shapiro-Wilk Normality through the packages ‘irr’ (Gamer et al., 2019) and ‘lpSolve’ (Berkelaar, et al., 2020) to check the normality of the data. This section has described the analysis of the CATSS for measuring the participants’ background knowledge of vocabulary. The results of the test is explained in Chapter 5 to show the final number of the participants taking part in the main study and to compare background knowledge between six groups of the participants. The next two sections in this chapter will present the process of the descriptive and inferential statistical analyses of the form-recall test in turn.

## 4.7 Analysis of the Form-Meaning Recall test (F-MRt)

The Form-Meaning Recall test was designed and delivered to the participants as a Pre-test, Immediate Posttest, and Delayed Posttest in the experiment. This test helped to measure the vocabulary learning gains and retention of each group from the treatment at the form and meaning recall. The Form-Meaning Recall test (F-MRt) consisted of two parts: form (English spelling) and meaning (using L1 equivalents or L2 synonyms). Each part included 10 target test items and two additional items as mentioned in the Research Methodology (see Chapter 3). Initial letters of each word were also given to control the answer (see also Section 3.7.3, Section 3.8, Section 4.4.2 and the full F-MRt test in Appendix 6). This section starts with describing how the data from the Form-Meaning Recall test were input and scores. Then, it reports the descriptive statistics analysis of scores (sensitive and strict) on the F-MRt of the six groups of participants in the three testing times. Finally, it summarises the results of descriptive statistics of the test before investigating the significant differences among the six groups by using inferential statistics.

### 4.7.1 Scoring the Form-Meaning Recall test

The criteria used in the current study were adapted from the classification of errors on form which was investigated by previous studies (Arnaud, 1984; Engber, 1995), together with the criteria used by related research in controlled productive vocabulary (Laufer & Nation, 1995; 1999) and suggestions from Read (2000), the expert in the field of vocabulary assessment (see Table 4.4).

***Table 4.4. Common lexical errors and sensitive rating system from previous studies***

| | **Classification of lexical (form only) errors adapted from previous studies** | | **Criteria used for the PVLT** | **Suggestions from the expert in the field of vocabulary assessment** |
|---|---|---|---|---|
| **Author(s), year** | *Arnaud, 1984 (p.19)* | *Engber, 1995 (p.146)* | *Laufer & Nation, 1999 (pp.38-39)* | *Read, 2000 (p. 174)* |
| Interpretation: **Acceptable** | Minor spelling mistakes | Phonetically similar-semantically unrelated | Minor spelling mistakes and grammatical mistakes | Minor spelling in terms of structural errors (e.g., -s, -ed) |
| Interpretation: **Unacceptable** | Major spelling mistakes (uninterpretable) | Word distorted-major spelling error | | |

According to the suggestions from Read (2000, p. 174) and Laufer and Nation (1999, pp. 38-39), any structural or grammatical errors that were found from incorrect target items such as *encourage* (instead of *encouraged*) should be ignored for sensitive rating. That means one point is given if there are any structural or grammatical errors presented in the answers. However, this scoring system can be problematic when raters have to justify between minor and major errors. This could bring about some questions as "*What are considered to be minor errors?*" and "*Should the words strenghten, sthrengthen, and stregthen be given one score equally*?" if Read (2000) and Laufer and Nation's (1999) scoring system was employed. Due to this concern, the current study used both strict and sensitive scoring scheme to evaluate the test items of the F-MRt. Table 4.5 shows two scoring scheme: sensitive and strict used in this study.

***Table 4.5. Sensitive and strict scoring schemes***

| score | Strict rating category | Sensitive rating category | Examples of the (in)correct word forms for each category |
|---|---|---|---|
| **1** | Fully correct | Fully correct | *issued, encouraged, decrees, vendors, strengthen, portrayed,* etc. |
| **0.5** | | Mostly correct:<br>-FORM: spelling errors due to phonetically similarity) and/or missing one letter<br>-MEANING: can provide an answer but not entirely correct | *issue, encourages, decess, venders/vandors, strenghten, potrayed/portayed,* etc. |
| **0.25** | | Partially correct:<br>- FORM: spelling errors in terms of structural/grammatical errors (e.g., -s, -ed) and/or missing letters<br>-MEANING: fail to provide a complete answer but can be interpreted | *encorage, decres, vender, strenhten/ stregthten/stregthen, potray/portay,* etc. |
| **0** | Incorrect | Incorrect | Uninterpretable or no answer |

The sensitive scoring is more precise and gives 0.5 of a point to minor spelling errors due to phonetically similarity and 0.25 of a point to any structural or grammatical errors that could be interpreted. On the other hand, the criteria 'mostly correct' and 'partially correct' were considered as incorrect for the strict scoring scheme. This way the interpretation of the results could be more accurate (see discussion in Section 2.3.5). The reason that 0.75 point was not added to the sensitive scoring system is because the F-MRt was controlled by using dashes for each testing item. If it happens that at least one letter is missing from an item, it is quite clear that the item is considered as a mostly correct or partially correct item and could not be given 1 point. However, if there are some misspelling errors within that item, the 0.5 and 0.25 categories could be used to justify the errors in order to interpret the results. So, the score 0.75 was not applicable in the current study.

Even though the criteria was used, I am aware that self-grading can cause bias. As mentioned in Section 3.4.3, using more raters could help to ensure the reliability of the grading results (e.g., Lumley & McNamara, 1995; Johnson et al., 2008). Therefore, in the present study, one experienced non-native English teacher with a Ph.D. degree was also invited to grade the Form and Meaning Recall test (F-MRt). The two raters have the same L1 as the participants in the study. After the participants had completed the Form recall test at each time, I used both strict scoring and sensitive scoring schemes in Table 4.5 to grade the Form-Meaning Recall Test. The grading was done by two raters (I and another six-year experienced teacher in the institution). For strict scoring, each correct answer worth one mark. Minor spelling errors were ignored. This is different from the sensitive scoring system in which minor errors were given 0.5 points while major errors that could be interpreted were given 0.25 points (see the grading criteria in Table 4.5). Then, the dataset of both sensitive and strict scores were analysed by using Cohen's Kappa reliability through R programme. This was done to ascertain if the two raters have high agreement on rating the test items to avoid bias that may occur from self-evaluation that can result in misleading interpretation.

The results from inter-rater (Cohen's Kappa) reliability showed substantial to very almost perfect levels of agreement (Landis and Koch, 1977) in all tests (Pre-test, Immediate Posttest, and Delayed Posttest) and both rating systems (sensitive and strict) as presented in Table 4.6.

***Table 4.6. Cohen's Kappa reliability of the test between two raters***

| | | **Sensitive Scoring** | | | **Strict Scoring** | | |
|---|---|---|---|---|---|---|---|
| Modality | n | Pre-test | Immediate Posttest | Delayed Posttest | Pre-test | Immediate Posttest | Delayed Posttest |
| Form and meaning | 247 | 0.69 | 0.96 | 0.97 | 0.86 | 0.99 | 0.99 |
| Form only | 247 | 0.90 | 0.95 | 0.98 | 0.99 | 0.99 | 0.99 |

*p* < .001*; n = number of the participants*

Even though the Pre-test score of Form and Meaning recall test rated by the sensitive rating system showed lower level of agreement (Kappa = 0.69, *p* < .001) than the other modalities, this level is considered as substantial level of agreement (Landis and Koch, 1977). Also, the Form only revealed high degree of agreement between the two raters (Kappa = 0.90, *p* < .001). When taking a look at the meaning test alone, it was found that the problematic items that were graded differently are due to unclear part of speech of the Thai (L1) language. Typically, Thai people use some nouns such as 'concern' and 'shortage' as verbs or adjectives and usually omit the prefixes such as ความ (Kwam) or การ (Gaan) that identify those words as a noun. It happens quite often that Thai people use verbs and nouns interchangeably because they have the same meaning. This caused the differences in sensitive scores of the meaning test items that contains L1 translation as one rater gave 0.5 point while another rater gave 0.25 point to those items. Although this grading problem could be solved by regarding the problematic items, it seems unnecessary to spend a lot of time rechecking those items again as meaning recall and grammars (part of speech) are not the aim of the investigation. As clarified in the previous chapters (Chapter 1 and Chapter 3), meaning recall was only added to the test to ensure that the participants could partially/fully write the target word form as a result of gaining the knowledge of that word after the learning. Rechecking those meaning recall items would bring more drawbacks than benefits to the present study because it can delay the expected plan which tends to have a limited timeline. Importantly, this problem was not found from strict rating of form and meaning test that only zero and one point were given to incorrect and fully correct items, respectively. As this study mainly focuses on form recall, and the test of meaning was only used to ensure the results of the form recall test, this could be assumed that both sensitive and strict scores of the Form-Meaning Recall Test graded by the two raters are reliable. However, the different graded scores were checked again by the two raters to reach the same agreement. This is to ensure that the grading results were accurate and reliable. Once the scores of the test had been finalised, they were input in the Excel spreadsheets for statistical analysis with R. The following sections explain the analysis of descriptive statistics of the Form-Meaning Recall test with a provision of data prepared for further inferential statistics analysis.

### 4.7.2 Descriptive statistics analysis of the Form-Meaning Recall test (FRt)

The descriptive statistics information of the tests were obtained from the repeated measurement dataset of each group by using Mean (M) scores and Standard Deviations (SD). The descriptive statistics information helps us to understand the dataset and the potential relationships between variables. The test scores were analysed by using the ‘Psych’ package (Revelle, 2022) in R to explain descriptive statistics obtained from the six groups of participants. I first reported the descriptive statistics obtained from the repeated measurement dataset of all groups by using Mean (M) scores and Standard Deviations (SD) to show form and meaning scores graded by the two scoring systems (sensitive and strict). Table 4.7 below shows that the overall Pre-test scores were lower than the overall scores of both posttests.

Also, the Form Recall test (FRt) scores were higher than half of the total scores of the Form-Meaning Recall test (F-MRt) in both posttests regardless of the scoring schemes. This could mean that the participants were able to remember the form of words better than their meanings. However, when taking a look into the meaning answers alone again, those wrong answers were mainly because of the omit of the prefixes such as ความ (Kwam) or การ (Gaan), similar to the issue mentioned in Section 4.7.1, meaning that the participants could remember the core meaning of the correct word forms but tended to have problems with grammatical knowledge.

***Table 4.7. Mean scores (and Standard Deviations) between the Form-Meaning Recall test (F-MRt) and Form Recall test (FRt)***

| | | Pre-test | | Immediate Posttest | | Delayed Posttest | |
|---|---|---|---|---|---|---|---|
| Scoring | n | F-MRt | FRt | F-MRt | FRt | F-MRt | FRt |
| *Sensitive* | 247 | 2.07<br>(1.33) | 0.9<br>(0.74) | 11.80<br>(4.41) | 6.59<br>(2.56) | 11.76<br>(4.84) | 6.67<br>(2.91) |
| *Strict* | 247 | 1.55<br>(1.13) | 0.74<br>(0.77) | 10.32<br>(4.35) | 6.26<br>(2.78) | 10.26<br>(4.61) | 6.43<br>(3.00) |

*Note: total score of Form-Meaning Recall test (F-MRt) is 20; total score of Form Recall test (FRt) is 10; n =total number of participants; (SD) = Standard Deviation*

Since the main aim of the main analysis was to explore the participants’ productive knowledge on form and grammatical aspect is beyond the scope of this study, I will focus on the descriptive statistics analysis of Form Recall test, which is a part of the Form-Meaning Recall (F-MRt) test used as a main test with all groups of the participants to explore the knowledge gains and retention of form recall in controlled written production. In the current study, I also explained the analysis of data from both sensitive and strict scores of the Form Recall test (FRt) to compare the results of the two scoring schemes. The analysis results is presented in Chapter 5: Main Study Findings. The following section described the analysis of the Form Recall test by using Mixed-effects Model which was selected due to its effectiveness and appropriateness to the present study (also see discussion in Section 3.9).

### 4.7.3 Inferential statistics analysis of Form Recall test (FRt)

The descriptive statistics allows us to see if there are any difference between the learning gains of the six groups. Meanwhile, the inferential statistics enables us to determine whether these differences are significant or not, which provides insights into the effects of TFA-related activities on the short-term and long-term retention of the target words. Mixed-effects Model and pairwise comparison of the Form Recall test (FRt) scores at the three testing times (Pre-test, Immediate Posttest, and Delayed Posttest) were used in the inferential statistics analysis of the current study. Previous studies related to the TFA and word recall (e.g., Keating, 2008; Hu & Nassaji, 2016; Khoshsima & Eskandari, 2017; Chaharlang et al., 2018) only used ANOVAs through SPSS programme for the analysis. This method of SPSS analysis cannot perform data wrangling (Winter, 2020) which can be seen as a limitation of using the programme. Also, the ANOVAs cannot fit a model to the data as they treats all factors as fixed factors. Mixed-effects Models, on the other hand, take data dependence into account, resulting in higher validity of the findings (Yu et. al., 2022) and better estimates to draw accurate interpretations (Winter, 2022). As explained in Section 3.9, Mixed-effects Models take both fixed and random effects into account, so that the non-independent clusters of data (also called random effects) such as subject (or participant) and other variables that are not of primary interest can be analysed into one' s analysis for appropriate inferences. By using the Mixed-effects Models, the random effects (i.e., sex, background knowledge of English, motivation in learning) taken from the questionnaire data (see details in Section 4.4.3) can also be analysed together with the Form Recall test scores to explore their potential influences on retention of the target word forms. Given the strengths of Mixed-effect Models over traditional ANOVAs, a series of linear Mixed-effects Models were used to answer the two research questions in the Main Study. This section will first describe how to construct the appropriate models for multilevel data analysis in R programme. Then, it will provide information related to assumptions checking and solutions in dealing with non-parametric data of the current study. Lastly, it will give details on using Mixed-effects Models for inferential statistics analysis of the main study.

#### 4.7.3.1 Constructing the models for multilevel data analysis

In this section, I explain the analysis of the main data by using inferential statistics which are Mixed-effects Models (MEMs) and pairwise comparison. Most research data have variables that are not clustered or nested, so they are grouped at a single level, and can be analysed by simpler statistics such as t-test or One-way between groups ANOVA. In this experimental study, however, the dataset contains some variables (also called predictors) such as groups and time (Pre-test and Posttests) that are nested and considered as hierarchical data (see also Section 3.9). For example, the students in Group 2 (*Motivation*) and Group 3 (*Noticing*) received different treatments before taking the two posttests, so it is possible that the treatments can affect the test results of each group. Groups and classes can be regarded as a ‘contextual variable’ (Field et, al., 2012, p. 856) because the participants in different groups had dissimilar classroom experiences during the experiment. So, the participants in each group are nested within their class. This nested data may require a multilevel model for data analysis.

To assess the need for a multilevel model, I used the ‘nmle’ package (Pinheiro et, al., 2022) in R to explore whether there is the evidence of variation across contexts (groups). The package allows me to use the lme() function which can also be used for non-linear models. The model in Figure 4.2 shows the syntax that helps to fit the baseline model that only the intercept, representing by ‘1’ in the function, was predicted. The sensitive score, representing by ‘SSscore’ in the model was used to explore the variation.

```
interceptOnly <- gls(SSscore ~ 1, data = FRt.MEM, method = "ML")
```

```
summary(interceptOnly)
```

**Figure 4.2. The model of only intercepts**

Then, the result was compared with that of another model (see Figure 4.3) with intercepts varying across groups because groups are regarded as a contextual variable. The syntax ‘random = ~1 | group’ was added to the model since ‘group’ was taken into account as a random effect. ANOVA was used to test the change of log-likelihood (-2LL) to see if the fit of the model is significantly improved. If the *p* value is less than 0.05, the change is highly significant (Field et al., 2012)

```
randomInterceptOnly <- lme(SSscore ~ 1, data = FRt.MEM, random = ~1 |
group, method = "ML")

summary(randomInterceptOnly)
```

**Figure 4.3. The model of random intercepts across groups**

```
randomInterceptOnly <- lme(SSscore ~ group + time, data = FRt.MEM,
random = ~1 | group, method = "ML")

summary(randomInterceptOnly)
```

**Figure 4.4. The model of random intercept and predictors**

The comparison result showed that there was a significant difference between the two models, $\chi^2(1) = 45.25$, *p* = < .001. This means that I should pay attention to the variability in intercepts as the intercepts vary significantly among the participants across groups. The main aim of this study, however, is to investigate the effects of each TFA component that was employed with different group of the participants on the Form Recall test score across three time periods. More predictors were added to the syntax (see Figure 4.4) to find out if the results can help to answer the research questions. The intercept ‘1’ was changed by predictors (group and time) that may affect the results of the test score.

Several single linear Mixed-effects Models (MEMs) were then constructed and computed through R. The full model is presented in Table 4.8. As mentioned earlier, variables from the questionnaire data was also taken into account as random effects (predictors) for analysis. By adding and removing the predictors, it was found that only two variables: group and time had significant difference results. The inclusion of age, sex, background knowledge of English, motivation in learning and motivation towards the implemented activity did not significantly improve the model fit.

***Table 4.8. The full MEM model***

| | |
|---|---|
| **Predictors** | 1) group, 2) time, 3) age, 4) sex, 5) background knowledge of English, 6) motivation in learning, 7) motivation towards the implemented activity |
| **Model syntax** | FRtMod0 <- lmer(SSscore ~ group*time + Age + Sex + BackgroundEN + Motiv1 + Motiv2 + (1 \| participant), data = FRt.MEM, REML = FALSE, control = lmerControl(optimizer = "bobyqa")) |

Thus, the best fit model as a baseline model for data interpretation of this study included only the predictors (group and time) that significantly improved model fit as shown in Table 4.9 below. Then, I tried another way which helps with a repeated measurement design or ‘time series data’ as suggested by Field and his colleagues (Field, et al, 2012, p.895) to recheck the results. At the first stage, the model was restructuring based on the baseline model that includes only the intercept (see Model 1 in Table 4.10). This is similar to the method presented in Figure 4.3.

***Table 4.9. The fit model for MEM analysis***

| | |
|---|---|
| **Predictors** | 1) group, 2) time |
| **Model syntax** | FRtMod1 <- lmer(SSscore ~ group*time + (1 \| participant), data = FRt.MEM, REML = FALSE, control = lmerControl(optimizer = "bobyqa")) |

However, the option ‘na.action = na.exclude’ was added to this model for avoiding errors that might happen from the missing, or N/A data. I also used the maximum-likelihood estimation by adding the method = “ML” to the syntax.

Then, Model 2 (see Table 4.10) was constructed to allow the intercept varying across the participants by adding a new option ‘random = ~1 | participant’.

***Table 4.10. Models for rechecking the significance of time for repeated measurement***

| | |
|---|---|
| **Model 1** | intercept <- gls(SSscore ~ 1, data = FRt.MEM, method = "ML", na.action = na.exclude) |
| **Model 2** | randomIntercept <- lme(SSscore ~ 1, data = FRt.MEM, random = ~1 \| participant, method = "ML", na.action = na.exclude) |
| **Model 3** | timeRI <- update(randomIntercept, .~. + time) |
| **Model 4** | timeRS <- update(timeRI, random = ~ time \| participant) |

After that, the *update* function in R was employed to add ‘time’ as a fixed effect to the baseline model (see Model 3 in Table 4.10). This helps to avoid typing a long syntax again as the function helps to retain the syntax from the baseline model (Field et al, 2012). As for Model 3, the model showed that ‘time’ which is a new fixed effect was added to the Model 2 syntax (randomIntercept) by applying the *update* function. Later, Model 4 in Table 4.10 was created by including the random part of the model, ‘random = ~ time | participant’ to the Model 3. This random slope was added to allow the intercepts to vary across the participants. Finally, the *Anova* function was used to compare the statistically significant difference among the 4 models. The results showed that ‘time’ had an effect on the score as adding the fixed effect of time to the model significantly improved the baseline model ($\chi^2(1) = 677.61$, *p* = < .0001). Also, there was a significantly different between Model 3 and Model 4, meaning that adding a slope for the effect of time across the participant had significantly improved the model ($\chi^2(1) = 391.59$, *p* = < .0001). However, there were no statistically different between Model 1 and Model 2 ($\chi^2(1) = 0.00$, *p* = .99). This could confirm that the effect of ‘time’ is significant and should be added as a predictor for a fixed effect. This is because it tended to improve the base line model when it was added to Model 3 and Model 4.

The process mentioned earlier helps to ensure the significance of the selected predictors to be included to the fit model. After that, the data was analysed by using ‘lme4’ package (Bates et al., 20015) in R programme. The dependent variable in all models were Form Recall test scores from Pre-test, Immediate

Posttest, and Delayed Posttest. The model puts an emphasis only on the interaction of groups (different treatments) with time in the repeated measurements (Pre- and Posttests) as well as other dependence variables (predictors) that might affects the test results to estimate varying intercepts across groups.

Similar to other inferential statistics, nevertheless, the data should not violate basic assumptions such as normality, homogeneity, or multicollinearity for linear models. The section below explained the process of checking assumptions before using the selected model to analyse the main research data of the current study.

#### 4.7.3.2 Checking assumptions

As mentioned in Section 3.9, data interpretation can be misleading if assumptions are violated. Yet, this process was ignored by many studies (Field et. al., 2012; Hu & Plonsky, 2019). Mixed-effects Models are based on linear models. Several assumptions for linear models were applied to check if the data meets the requirements such as normality of distribution, linearity, heteroskedasticity, and multicollinearity of variance. Generally, the assumptions can be assessed by visualisations: histograms, boxplots or QQ plots. Figure 4.5 below presents the histogram from the main data of the current study.

It should be noted that the data included the scores of all groups across the three testing time periods (Time1: Pre-test; Time 2: Immediate Posttest; Time 3: Delayed Posttest) as presented in the selected statistical model. The histogram, which does not visually appear to be poor, illustrated that the data tended to be normally distributed with a skewness of 0.13 and a kurtosis of -1.51 (SE 0.13). Figure 4.5 shows that the data which was the Form Recall test (FRt) scores (total = 10 points) of all groups across the three tests tended to be normally distributed. Although the kurtosis result showed a negative value (-1.51, SE = 0.13), the violation of normal distribution is often not the problem with a large number (n = 247) of the sample size (Lumley et al., 2002) according to the Central Limit Theorem (Fischer, 2021).

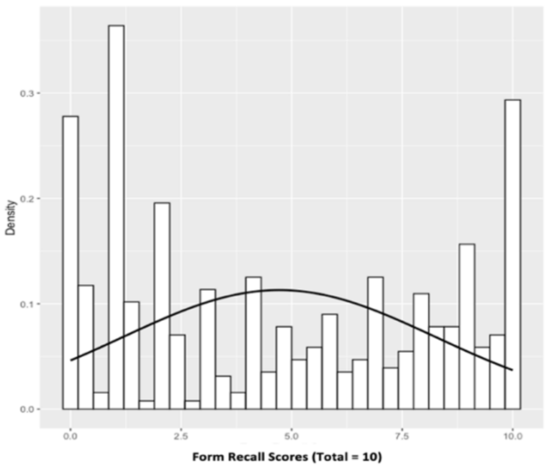


**Figure 4.5. A histogram of the collected data of the main study**

Besides, non-normality tended to be the least concern for linear regression models (Gelman & Hill, 2007; Lumley et al., 2002) and is not a problem when using Mixed-effects Models as suggested by Winter (2020) and Field et al. (2012). To check if the outliers affect the normality of the data, I then removed the outliers that were presented by the boxplot (see also Figure 5.2 and details in Section 5.2.1). The 'rstatix' package (Kassambara, 2021) in R was also used to identify these outliers with the function 'identify_outliers()'. After the outliers were removed, the result was not different from the analysis of the original data. Therefore, I decided to keep these outliers owing to the concerns on number of participants and reduction of the estimation power. For non-normal distribution data, transformations and robust tests, however, can be the alternative solutions (Field et. al., 2012). The information regarding these alternative ways was provided in the Section 4.7.3.3 below.

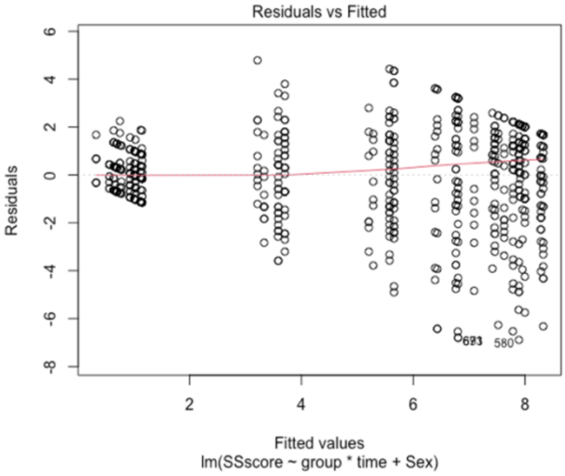


**Figure 4.6. Q-Q plot showing residuals and fitted values**

Another assumption is linearity which can be checked by using the Q-Q plots illustrated in Figure 4.6. It presented residuals and fitted values of the data among the six groups across three time periods. Generalisation of the findings can be limited if the plot shows non-linearity relationship of the data (Field et al., 2012). Although it was not a perfect straight line, the data was likely to be linear as represented by the red line in the graph.

Then, multicollinearity was checked by using the ‘MuMIn’ package (Bartoń, 2022) in R. The data met the collinearity assumption as indicated by the Variance Inflation Factor (VIF). The VIF values for the predictors *group* (VIF = 1.04), *time* (VIF = 1.00), and *sex* (VIF = 1.04) were not greater than 10, meaning that the multicollinearity was not a concern (Everitt & Skrondal, 2010; Field et. al., 2012).

Therefore, it can be assumed that the selected statistical test is appropriate to use as the assumptions were checked and only the issues regarding the outliers and normality were identified. Some solutions to these problems are using log or square root for data transformation and/or applying robust tests (Field et. al., 2012). The section below gave information about these alternative solutions.

#### 4.7.3.3 Data transformation

A number of studies (e.g., Öksüz et al., 2020; West, 2022) used data transformations such as log transformation (natural logarithms or common logarithms) to solve the problem on non-normal distribution of data. It can help with highly skewed distribution (Feng et al., 2012; West, 2022). However, some researchers in the field of Psychology (e.g., Games, 1984) and Linguistics (e.g., Schütze & Sprouse, 2014) suggested that this method is not always the best practice as the relationship between the data can be distorted from eliminating right-tail outliers. In the current study, the results from the analysis of original data and the common log (base 10) transformed data were compared.

***Table 4.11. Comparison between the original and transformed data models***

| **Models** | **AICs** |
|---|---|
| **modelF.null** (original data) | 3066.0 |
| **modelF.log** (log transformed data) | 4097.3 |

The Akaike' s information criterion, or AIC which is 'a goodness-of-fit measure' helped to identify the best fit model by estimating the amount of the parameters to be estimated (Field et al., 2012, pp. 868 & 913). The smaller number indicates the better-fitting measure. The result revealed that the model with the original data had a smaller value of AIC than model containing the log transformed data (see Table 4.11). This indicates that the original model can estimate more predictor variables than the data that was transformed by using logarithm. Corresponded to Games (1984) and Schütze and Sprouse' s (2014) suggestions, the untransformed scores were likely to be more useful than the unfitting transformed data. Although a robust test as explained in Section 3.9, "robustlmm" (Koller, 2016) in R can be used if data transformation does not seem to be the best solution, it seemed to be unnecessary as the assumptions were checked and did not appear to be excessively violated. Thus, the original data was used for analysis of the current study. The Section 4.7.3.4 below presents the analysis of the non-robust test using Group 1 (Control group) and other experimental groups (Groups 2-6) as a reference level, respectively.

#### 4.7.3.4 Analysis of Form Recall test by Mixed-effects Models

The package lmerTest (Kuznetsova, et al., 2013) was used to analyse the fit model (see the model fit in Table 4.9). To explore the marginal and conditional $R^2$ for Mixed Models, the functions "r2()" and "model_performance()" from the package performance (Lüdecke et al., 2021) were also implemented. The Form Recall test (sensitive scoring) scores on the Pre-test (Time 1), Immediate Posttest (Time 2), and Delayed Posttest (Time 3) were compared by using the linear Mixed-effects Model to answer the two research questions. To answer Research Question 1: "*Do activities with high TFA scores result in better retention of single words in productive form recall?",* I constructed a linear Mixed-effects Model (see Table 4.9). In this model, score on the Form

Recall test was the dependent variable. The fixed effects were Time (Pre-test, Immediate Posttest, Delayed Posttest), Group (Control*, Motivation, Noticing, Retrieval, Generative Use,* and *All TFA Components*), and the interaction between Time and Group. The Control group (Group 1) and Pre-test (Time 1) were set as a reference level for Group and Time, respectively. While sex, age, participants' background knowledge of English and motivation were taken into consideration as random effects, the analysis of models explained in Section 4.7.3.1 showed no impact of these variables on the test scores. As a result, the only random effect of interest of the main analysis was the subject (individual participant). As the Control group received no treatment, the comparison between this group and the experimental groups could reveal the fact that whether or not the learning occurs due to the influence of the TFA variables. While the analysis result from the Pre-test (Time 1) helped to explore the participants' prior knowledge on form recall that might affect the research results, the analysis of the Posttests (Time 2 and Time 3) could reveal the retention of word forms resulted from different treatments. The estimate values along with $p$-values and $R^2$ for the fixed and random effects were explained to answer the research questions in Section 5.1 and Section 5.2 of the Chapter 6: Main Study Findings. Another analysis was to examine the interaction of group and time on the test score. As the significant differences were found (also see details in Section 5.1 and Section 5.2), the pairwise comparison using the "*emmeans*" package in R (Lenth, 2021) was employed to identify which group, as treated by no or different treatment, can perform better than the others. The result was later compared with the TFA scores that were evaluated in Section 3.4.3 (see also Table 3.5) to answer the Research Question 1 in Section 5.2.2.1.

To answer Research question 2: "*What is the individual role of the different TFA components on promoting vocabulary retention through vocabulary tasks?",* I compared the results between the three tests (Time 1 vs. Time 2, Time 1 vs. Time 3, and Time 2 vs. Time 3) among the four experimental groups (*Motivation, Noticing, Retrieval,* and *Generative Use*). This was done by relevelling the model; that is, altering the reference level from Control (Group 1) to *Motivation* (Group 2), *Noticing* (Group 3), *Retrieval* (Group 4), *Generative Use* (Group 5), and *All TFA Components* (Group 6) in turn. The reference level was changed because the second question focused on the effect of *Motivation, Noticing, Retrieval,* and *Generative Use* on retention of form recall. This helps to recheck the analysis findings whether there is a consistency of the results. The *mutate()* function from the *dplyr* package (Wickham, 2022) was used to store the vectors in a new column for relevelling. As the data was stored by the *read_csv()*, the character vectors must be converted to factor vectors in order to relevel the reference. Later, I changed a reference level of testing time in the model from Pre-test (Time 2) to Immediate Posttest so that the comparison results between the Immediate Posttest (Time 2) and Delayed Posttest (Time 3) could be compared. Similar to the analysis of the Research Question 1, the linear Mixed-effects Model and pairwise comparison with Bonferroni adjustments for multiple comparisons by using the package "*emmeans*" were implemented to examine the extent to which TFA encourages *Motivation Noticing, Retrieval,* and *Generative Use* to support retention of form recall. The analysis result is reported Section 5.2.2.2 in the Chapter 5.

# Chapter 5
# Main Study Findings

The previous chapter has described the steps taken to analyse the main data sets collected from the experiment in the main study: (a) the scores from the adapted CATSS to measure the participant's prior knowledge of general vocabulary before the treatment and (b) the scores on the Pre-test, Immediate Posttest, and Delayed Posttest of the Form-Meaning Recall test (F-MRt) to measure the participants' productive knowledge of the target words before and after the treatment. It also described the analysis of the supplementary data from the post-treatment questionnaire. This chapter will report the findings of each data set in turns. It begins with the participants' prior knowledge of general vocabulary as measured by the CATSS (see Section 5.1). Next, it reports learning gains and retention of the target words of the six groups of participants as measured by the Form Recall test (FRt) before and after the treatment (see Sections 5.2).

## 5.1 The participants' prior knowledge of general vocabulary as measured by the adapted CATSS

***Table 5.1. CATSS overall scores from six groups of participants***

| Group | *n* | Mean (*SD*) | Shapiro-Wilk Normality Test (*p*-value) |
|---|---|---|---|
| 1 (Control) | 38 | 66.37 (3.24) | 0.31 |
| 2 (*Motivation*) | 40 | 66.23 (3.46) | 0.23 |
| 3 (*Noticing*) | 41 | 66.63 (4.00) | 0.37 |
| 4 (*Retrieval*) | 40 | 66.88 (3.57) | 0.15 |
| 5 (*Generative Use*) | 41 | 66.80 (3.47) | 0.37 |
| 6 (*All components*) | 46 | 66.28 (3.54) | 0.69 |

*Note: n = number of participants; SD = Standard Deviation*

Table 5.1 presents the CATSS overall scores of each group of participants. After the results of normality was checked in the preliminary analysis, the inferential statistics was used to explore the significant difference between six groups of the participants with regard to their prior knowledge of general vocabulary. The result of the one-way between group ANOVA analysis indicated that the participants from all groups had similar background knowledge of vocabulary because there was no statistically significant difference among the six groups, *F* (1, 245) = 0.07, $p$ = .79. The effect size analysed by using eta squared ($n^2$= 0.00029) also confirmed that the difference between these six groups was less than 1% ($n^2$ < .01) which is considered as a very small difference (Cohen, 1988). This means that the participants in the six groups had similar prior knowledge of general vocabulary before the treatment.

Table 5.2 shows the scores of each group in each CATSS modality. It shows that all groups had meaning recognition the highest, followed by form recognition, meaning recall, and form recall, respectively.

***Table 5.2. CATSS scores of each modality from six groups of participants***

| | | Form Recall | Meaning Recall | Form Recognition | Meaning Recognition |
|---|---|---|---|---|---|
| Group | *n* | Mean (*SD*) | | | |
| 1<br>(Control) | 38 | 12.84<br>(1.39) | 14.05<br>(2.12) | 19.74*<br>(0.60) | 19.74*<br>(0.69) |
| 2<br>(Motivation) | 40 | 13.43<br>(1.57) | 14.00<br>(1.83) | 19.23<br>(1.05) | 19.57<br>(0.71) |
| 3<br>(Noticing) | 41 | 13.15<br>(2.42) | 14.41<br>(1.64) | 19.49<br>(0.87) | 19.59<br>(0.71) |
| 4<br>(Retrieval) | 40 | 13.30<br>(2.02) | 14.38<br>(1.85) | 19.50<br>(0.72) | 19.70<br>(0.69) |
| 5<br>(Generative Use) | 41 | 13.56<br>(1.82) | 14.24<br>(1.92) | 19.32<br>(0.99) | 19.68<br>(0.65) |
| 6<br>(All TFA Components) | 46 | 13.55*<br>(1.86) | 13.51*<br>(1.86) | 19.06<br>(0.77) | 19.62<br>(0.74) |

*Note: n = number of participants; SD = Standard Deviation; *the noticeable different pattern*

Results of the one-way between group ANOVA analysis showed that the participants in all groups had similar prior vocabulary knowledge regardless of the aspects of vocabulary knowledge being measured (see the *p*-value column in Table 5.3).

***Table 5.3. One-way ANOVA results by each modality***

| | | Sum of Squares | *df* | Mean Square | *F* | *p** |
|---|---|---|---|---|---|---|
| Form Recall | Between Groups | 9.8 | 1 | 9.779 | 2.794 | 0.0959 |
| | Within Groups | 857.6 | 245 | 3.500 | | |
| Meaning Recall | Between Groups | 3.5 | 1 | 3.486 | 0.987 | 0.322 |
| | Within Groups | 865.7 | 245 | 3.534 | | |
| Form Recognition | Between Groups | 0.04 | 1 | 0.0373 | 0.051 | 0.822 |
| | Within Groups | 179.59 | 245 | 0.7330 | | |
| Meaning Recognition | Between Groups | 0.01 | 1 | 0.0142 | 0.029 | 0.864 |
| | Within Groups | 118.34 | 245 | 0.4830 | | |

*Note: *p < .05*

To conclude, the CATSS results confirm that the participants in six groups had similar prior knowledge on general vocabulary in all modalities (form recall, meaning recall, form recognition, and meaning recognition) before the treatment. This indicates that prior knowledge was likely to have minimal effects on the six groups' acquisition of the target words from the treatment Data analysis of the main instrument, Form Recall test used in the current study to explore the effects of *Motivation*, *Noticing*, *Retrieval*, and *Generative Use* on vocabulary learning are explained in the below section.

## 5.2 The participants' knowledge of the target words at three testing times as measured by the Form Recall test (FRt)

The Form Recall test, or FRt is a part of the Form-Meaning Recall (F-MRt) test used as a main research instrument with all groups of the participants. The results of the FRt analysis help to prepare data for interpretation that addresses the research questions in Section 2.5. In this section, I will first report the descriptive statistics obtained from the repeated measurement dataset of each group by using Mean (M) scores and Standard Deviations

(SD) in Section 5.2.1. Then, I will present the analysis results from inferential statistics in Section 5.2.2. The findings were reported in line with the research questions of the current study.

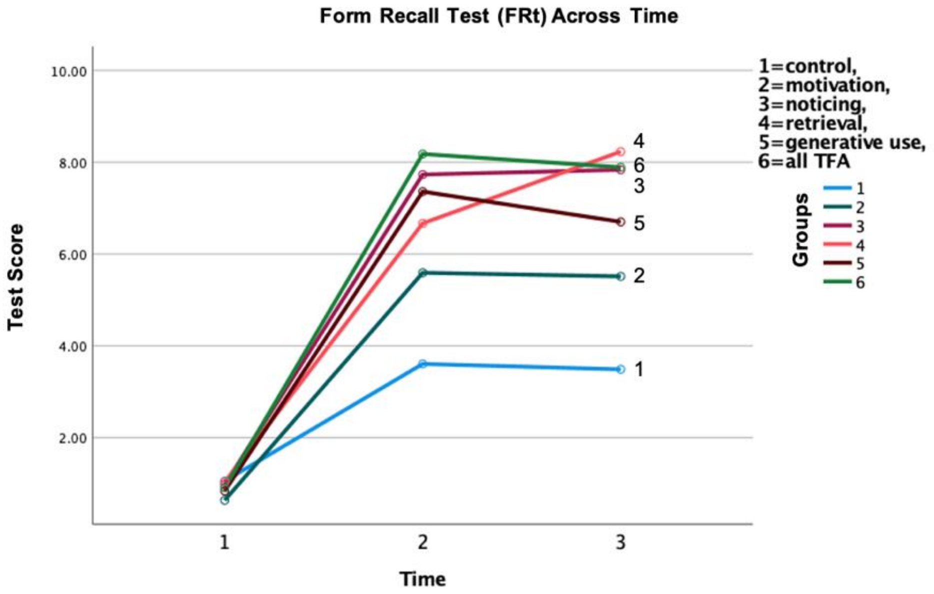
### 5.2.1 Results of the descriptive statistics analysis of the FRt

The information was reported in terms of *time* (Time 1: Pre-test, Time 2: Immediate Posttest, and Time 3: Delayed Posttest) and *group* (Group 1: Control, Group 2: *Motivation*, Group 3: *Noticing*, Group 4: *Retrieval*, Group 5: *Generative Use*, Group 6: *All TFA components*). This section includes the overall descriptive statistics data, details of descriptive statistics data of each test (Pre-test, Immediate Posttest and Delayed Posttest) and the summary of the descriptive statistics analysis results.

#### 5.2.1.1 The overall descriptive statistics data

In this section, descriptive statistics obtained from the Pre-test (Time 1), Immediate Posttest (Time 2), and Delayed Posttest (Time 3) were compared to explore the effects of different treatments on form recall knowledge. Ideally, we would expect learners to be able to use the language perfectly without mistakes in use. For this reason, I will report the overall FRt results by focusing on strict scores although the sensitive scores was eventually selected for several reasons in the main inferential statistics analysis using Mixed-effects Models. The main reason is that sensitive scores could help to reveal the almost-perfect stage of learning and memory retention, and how each modality helps students with memory recall and so were taken into account for later analysis (see more rationales and details in Section 5.2.1.4 and Section 5.2.1.5, respectively).

The result from the Control group were also compared with the intervention groups as presented in Figure 5.1 and Table 5.4. It showed that the participants in the intervention groups (Groups 2 - 6) had similar scores on Pre-test as those of the *Control* group (Group 1), implying that they were likely to have similar prior knowledge on the testing words. However, all intervention groups had higher scores on the posttests (both Time 2 and Time 3) than the control group (see Figure 5.1).



*Total score is 10; Time 1 = Pre-test, Time 2 = Immediate Posttest, Time 3 = Delayed Posttest*

**Figure 5.1. Line graph of FRt mean strict scores by group and time**

All the experimental groups received greater mean scores than the *Control* group in all posttests, indicating that the treatments might have a positive influence on the development of form recall knowledge. Knowledge gain was then calculated by taking the scores from the Pre-test and Posttests to calculate the percentage of gain or loss on form recall knowledge. The score of Time 2 (Immediate Posttest) was subtracted from the Time 1 (Pre-test) score to explore the percentage of knowledge gain. This formula was also used to compare the gain or loss of form recall knowledge between the immediate and Delayed Posttests. According to Table 5.4, Group 1 (Control) gained only 23% of knowledge at the form recall level from Time 1 to Time 2 (short-term retention) and Time 1 to Time 3 (long-term retention) while Group 6, which had a high degree of all TFA components, had the highest percentage of gain, showing 71% from Time 1 to Time 2. Although Group 2 (*Motivation*) received lower scores than the other intervention groups from both Posttests, its percentage of knowledge gain (from Time 1 to Time 2) was still double that of the Control group. This indicated that the treatments were likely to contribute to the learning.

***Table 5.4. Mean strict scores (and Standard Deviations) of the FRt***

| | | Pre-test | Immediate Posttest | Delayed Posttest | Approximate percentage (%) of knowledge gain | |
|---|---|---|---|---|---|---|
| Group | *n* | | Mean (*SD*) | | from Time 1 to Time 2 | from Time 1 to Time 3 |
| Control | 38 | 0.84<br>(0.79) | 3.16<br>(2.09) | 3.13<br>(2.06) | 23% | 23% |
| Motivation | 40 | 0.63<br>(0.73) | 4.85<br>(2.89) | 5.29<br>(2.03) | 42% | 47% |
| Noticing | 41 | 0.61<br>(0.77) | 7.61<br>(2.29) | 7.63<br>(2.74) | 68% | 70% |
| Retrieval | 40 | 0.90<br>(0.84) | 6.50<br>(2.67) | 8.07<br>(2.43) | 56% | 72% |
| Generative Use | 41 | 0.61<br>(0.70) | 7.05<br>(1.83) | 6.44<br>(3.39) | 64% | 58% |
| All TFA components | 46 | 0.85<br>(0.79) | 7.93<br>(1.48) | 7.65<br>(2.11) | 71% | 68% |

*Note: total score is 10; n = number of participants; SD = Standard Deviation*

The sections below explore the results from sensitive and strict scores of each testing time to identify the effects of the treatments from the two different scoring systems. It should be noted that the one-way between groups ANOVA reported in the following sections was only used for the purposes of comparing the difference among the two scoring systems. It was not used as the main inferential statistic in the current study. As mentioned earlier, the Mixed-effects Model was used as the main analysis method and explained in Section 5.2.2.

#### 5.2.1.2 The Pre-test (Time 1) graded by two scoring schemes

Table 5.5 represents the descriptive statistics from the Pre-test of the form recall investigation. While the maximum scores of form ranged between 2 and 3 (out of 10) points, the minimum score graded by both sensitive and strict rating systems was the same (Minimum = 0) in all groups. Then, the Pre-test scores from six groups were compared through R programme to explore whether or not the participants in each group have similar prior knowledge of the words in the test (see Table 5.6).

***Table 5.5. Descriptive statistics of FRt as a Pre-test***

| | | M | | *SD* | | Min | | Max | |
|---|---|---|---|---|---|---|---|---|---|
| GROUP | n | sensitive | strict | sensitive | strict | sensitive | strict | sensitive | strict |
| Control | 38 | 1.05 | 0.84 | 0.75 | 0.79 | 0 | 0 | 3 | 3 |
| Motivation | 41 | 0.63 | 0.63 | 0.73 | 0.73 | 0 | 0 | 2 | 2 |
| Noticing | 41 | 0.98 | 0.61 | 0.67 | 0.77 | 0 | 0 | 2.5 | 2 |
| Retrieval | 40 | 1.03 | 0.90 | 0.83 | 0.84 | 0 | 0 | 3 | 3 |
| Generative Use | 41 | 0.84 | 0.61 | 0.63 | 0.70 | 0 | 0 | 2.5 | 2 |
| All TFA | 46 | 0.91 | 0.85 | 0.79 | 0.79 | 0 | 0 | 2.5 | 2 |

*Note: n = number of participants;* M*=Mean; SD=Standard Deviation; Min=Minimum score; Max=Maximum score*

The results showed that there is no statistically significant difference among the six groups of the participants on both sensitive scoring, *F* (5, 241) = 1.794, *p* = .115 and strict scoring, *F* (5, 241) = 1.277, *p* = .274 (see Table 5.6).

***Table 5.6. One-way ANOVA of the Pre-test (sensitive and strict scoring)***

| Scoring | | Sum of Squares | *df* | Mean Square | *F* | *Sig** |
|---|---|---|---|---|---|---|
| sensitive | Between Groups | 4.88 | 5 | 0.9754 | 1.794 | .115 |
| | Within Groups | 131.03 | 241 | 0.5437 | | |
| strict | Between Groups | 3.81 | 5 | 0.7610 | 1.277 | .274 |
| | Within Groups | 143.61 | 241 | 0.5959 | | |

**p < .05*

Both the results from the CATSS as mentioned in Section 5.1 and the Pre-test show that the participants in each group have similar background knowledge of vocabulary in terms of form recall. This suggests the results from Immediate Posttest and Delayed Posttest were unlikely to have been influenced by the participants' prior knowledge.

#### 5.2.1.3 The Immediate Posttest (Time 2) graded by two scoring schemes

As with the Pre-test, scoring was evaluated by two raters using two scoring scheme (sensitive and strict). The data was also computed by using R to identify the Means (M), Standard Deviations (SD), and significant differences among the six groups.

***Table 5.7. Descriptive statistics of FRt as an Immediate Posttest***

| GROUP | n | M | | *SD* | | Min | | Max | |
|---|---|---|---|---|---|---|---|---|---|
| | | sensitive | strict | sensitive | strict | sensitive | strict | sensitive | strict |
| Control | 38 | 3.61 | 3.16 | 1.83 | 2.09 | 0 | 0 | 7.5 | 7 |
| Motivation | 41 | 5.59 | 4.85 | 2.38 | 2.89 | 0.75 | 0 | 10 | 10 |
| Noticing | 41 | 7.73 | 7.61 | 2.24 | 2.29 | 1.25 | 1 | 10 | 10 |
| Retrieval | 40 | 6.67 | 6.50 | 2.62 | 2.67 | 2 | 2 | 10 | 10 |
| Generative Use | 41 | 7.36 | 7.05 | 1.85 | 1.83 | 2.25 | 2 | 9.5 | 9 |
| All TFA Components | 46 | 8.18 | 7.93 | 1.36 | 1.48 | 4.25 | 4 | 10 | 10 |

Note: n = number of participants; *M*=Mean; *SD*=Standard Deviation; Min=Minimum score; Max=Maximum score

Irrespective of the scoring schemes, it was found that Group 6 (All TFA Components), the group that was treated with the highly rated activity in all four TFA modalities: *Motivation, Noticing, Retrieval,* and *Generative Use*, received the highest Mean score (M = 8.18 for sensitive scoring, *SD* = 1.36; Mean = 7.93, *SD* = 1.48 for strict scoring) than the other groups (see Table 5.7).

The participants in Group 1 (Control group) got the lowest scores (Mean = 3.61, *SD* = 1.83 for sensitive scoring; Mean = 3.16, *SD* = 2.09 for strict scoring), followed by Group 2 *Motivation* (Mean = 5.59, *SD* = 2.38 for sensitive scoring; Mean = 4.85, *SD* = 2.89 for strict scoring), Group 4: *Retrieval* (Mean = 6.67, *SD* = 2.62 for sensitive scoring; Mean = 6.50, *SD* = 2.67 for strict scoring), Group 5: *Generative Use* (Mean = 7.36, *SD* =1.85 for sensitive scoring; Mean = 7.05, *SD* =1.83 for strict scoring), and Group 3: *Noticing*

(Mean = 7.73, *SD* = 2.24 for sensitive scoring; Mean = 7.61, *SD* = 2.29 for strict scoring), respectively as presented in Table 5.8.

***Table 5.8. FRt scores of the Immediate Posttest by ranking***

| Group by ranking (lowest to highest score) | n | M | | *SD* | |
|---|---|---|---|---|---|
| | | sensitive | strict | sensitive | strict |
| Group 1: Control | 38 | 3.61 | 3.16 | 1.83 | 2.09 |
| Group 2: *Motivation* | 41 | 5.59 | 4.85 | 2.38 | 2.89 |
| Group 4: *Retrieval* | 40 | 6.67 | 6.50 | 2.62 | 2.67 |
| Group 5: *Generative Use* | 41 | 7.36 | 7.05 | 1.85 | 1.83 |
| Group 3: *Noticing* | 41 | 7.73 | 7.61 | 2.24 | 2.29 |
| Group 6: *All TFA Components* | 46 | **8.18** | **7.93** | 1.36 | 1.48 |

*Note: n = number of participants; M = Mean; SD = Standard Deviation; total score is 10; the highest score is in boldface*

Also, it can be noticed by the lower *SD* scores of the participants in the *Generative Use* group (*SD* = 1.85 for sensitive scoring; *SD* = 1.83 for strict scoring) and the *All TFA Components* group (*SD* = 1.36; *SD* = 1.48 for strict scoring) that the participants in these groups tended to have similar amount of knowledge gain while those in the *Motivation* group (*SD* = 2.38 for sensitive scoring; *SD* = 2.89 for strict scoring), the *Noticing* group (*SD* = 2.24 for sensitive scoring; *SD* = 2.29 for strict scoring), and the *Retrieval* group (*SD* = 2.67 for sensitive scoring; *SD* = 2.62 for strict scoring) seemed to have higher standard deviation due to variation of score. The participants from the *All TFA Components* group had the highest mean scores from both scoring systems with the lowest standard deviations when compared to the other groups with no or less amount of TFA support. Moreover, their minimum score (sensitive = 4.25; strict = 4 out of 10 points) was more than double of that of the other groups. This means that after receiving the treatment the majority of the participants in this group tended to perform well in the Immediate Posttest. The treatment seemed to facilitate most students in this group to gain some certain vocabulary knowledge if the Pre-test score (sensitive: Mean = 0.91, *SD* = 0.79; strict: Mean = 0.85, *SD* = 0.79) and Immediate Posttest score (sensitive: Mean = 8.18, *SD* = 1.36; strict: Mean = 7.93, *SD* = 1.48) were compared.

Then, inferential statistics were used to explore the significant differences between the mean scores of the six groups of the participants to compare the results from both sensitive and strict scoring methods. The data was computed by using the R programme. The analysis results of both sensitive scoring and strict scoring (see Table 5.9) showed that there were statistically significant differences among the six groups, sensitive scoring: *F* (5, 241) = 26.6, *p* = < .001; strict scoring: *F* (5, 241) = 26.87, *p* = < .001.

***Table 5.9. One-way between group ANOVA of the Immediate Posttest (sensitive and strict scoring)***

| Scoring | | Sum of Squares | *df* | Mean Square | *F* | *Sig** |
|---|---|---|---|---|---|---|
| **sensitive** | Between Groups | 573.6 | 5 | 114.72 | 26.60 | **<.001*** |
| | Within Groups | 1039.5 | 241 | 4.31 | | |
| **strict** | Between Groups | 678.3 | 5 | 135.66 | 26.87 | **<.001*** |
| | Within Groups | 1216.6 | 241 | 5.05 | | |

**p < .05*

After that, the data was analysed by using a *post hoc* test and also computed through R programme. The main purpose was to explore the difference between the results of sensitive scoring and strict scoring methods. This helped to ensure the results from inter-rater (Cohen's Kappa) reliability (see also Table 4.6 in Section 4.7.1) that showed a strong level of agreement between raters in both rating systems (sensitive and strict). The data from sensitive and strict scoring methods were analysed by using multiple comparison through Tukey's Honest Significant Difference (HSD).

The overall results showed similarities among the comparison of six groups by using both sensitive and strict rating systems. The mean sensitive score of the Control group (Group1: Mean = 3.61, *SD* = 1.83) was significantly lower than that of all experimental groups (Group 2: Mean = 5.59, *SD* = 2.38; Group 3: Mean = 7.73, *SD* = 2.24; Group 4: Mean = 6.67, *SD* = 2.62; Group 5: Mean = 7.36, *SD* = 1.85; Group 6: Mean = 8.18, *SD* = 1.36). The same trend was found from strict rating system scores as can be noticed from Table 5.8 above.

However, there was a slightly difference between the *post hoc* results from the sensitive rating and strict rating systems. It was found that the mean value of strict score was significantly different between the *Retrieval* and the *Motivation* groups (*p* < .01, 95% C.I. = [3.08, 0.21]) while there was no statistically significant difference in mean sensitive scores between these two groups (*p* = 0.18, 95% C.I. = [2.40, -0.25]). The *All TFA Components* group received the highest mean score than the other groups in both sensitive (Mean = 8.18, *SD* = 1.36) and strict rating (Mean = 7.93, *SD* = 1.48), but the score was close to that of the *Noticing* group (Sensitive: Mean = 7.73, *SD* = 2.24; Strict: Mean = 7.61, *SD* = 2.29) and the *Generative Use* group (Sensitive: Mean = 7.36, *SD* = 1.85; Strict: Mean = 7.05, *SD* = 1.83) in the Immediate Posttest. The *post hoc* test confirmed the similarity among them as there was no statistically significant difference in mean scores between the *All TFA Components* group and the *Noticing* group (*p* = 0.45, 95% C.I. = [1.77, -0.88]) and the *All TFA Components* group and the *Generative Use* group (*p* = 0.82, 95% C.I. = [2.14, -0.50]).

All in all, according to the analysis of the Immediate Posttest presented in this section, both sensitive and strict scores of the *Control* group were statistically significantly lower than all the other groups that received some certain support from the TFA framework. Furthermore, the *All TFA Components*, which received the highest support from the framework, not only received the highest mean score on both sensitive and strict scoring systems, but also had significantly difference from the Control (Group 1), *Motivation* (Group 2), and *Retrieval* (Group 4) groups. However, the *All TFA components* group was not significantly different from the *Noticing* (Group 3) and *Generative Use* (Group 5) groups in both rating systems (see more details in Section 5.2.2.1 and Section 5.2.2.2).

#### 5.2.1.4 The Delayed Posttest (Time 3) graded by two scoring schemes

This section reported the data analysis of the Delayed Posttest by using two scoring schemes (sensitive and strict). Similar to the analysis of Pre-test and Immediate Posttest, the collected data of the Delayed Posttest was initially analysed by using both descriptive statistics and One-way between groups

ANOVA to investigate the differences of two different rating scores among the six groups. The inferential statistics analysis explained in this section was only implemented to explore the difference between the two scoring scheme. The main analysis by using Mixed-effects Model and pairwise comparison as inferential statistics was reported in Section 5.2.2 and not included in this section.

***Table 5.10. Descriptive statistics of FRt as a Delayed Posttest***

| GROUP | N | M |  | *SD* |  | Min |  | Max |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  | sensitive | strict | sensitive | strict | sensitive | strict | sensitive | strict |
| 1 *Control* | 38 | 3.49 | 3.13 | 1.85 | 2.06 | 0 | 0 | 8 | 8 |
| 2 *Motivation* | 41 | 5.51 | 5.29 | 1.99 | 2.03 | 2 | 2 | 10 | 10 |
| 3 *Noticing* | 41 | 7.84 | 7.63 | 2.67 | 2.74 | 1 | 1 | 10 | 10 |
| 4 *Retrieval* | 40 | 8.23 | 8.07 | 2.34 | 2.43 | 2 | 2 | 10 | 10 |
| 5 *Generative Use* | 41 | 6.70 | 6.44 | 3.38 | 3.39 | 0 | 0 | 10 | 10 |
| 6 *All TFA* | 46 | 7.89 | 7.65 | 2.00 | 2.11 | 2.25 | 2 | 10 | 10 |

*Note: Total score is 10; N=number of participants; M=Mean; SD=Standard Deviation; Min=Minimum score; Max=Maximum score*

Table 5.10 represented the two-week Delayed Posttest scores graded by both sensitive and strict scoring systems. The Control group (Group 1) received the lowest Mean scores from both sensitive (Mean = 3.49; *SD* = 1.85) and strict scoring (Mean = 3.13; *SD* = 2.06). The Mean scores of the experimental groups were almost double that of the *Control* group in both grading methods. This is similar to the results obtained from the Immediate Posttest. When compared with the Immediate Posttest results, the Mean scores for both types of the grading of all groups dropped slightly, except that of the *Retrieval* group (Group 4).

While the *All TFA Components* group (Group 6) received the highest Mean score in the earlier test (Immediate Posttest), the *Retrieval* group (Group 4) got the highest Mean scores in both scoring types (sensitive: Mean = 8.23; *SD* = 2.34; strict: Mean = 8.07; *SD* = 2.43) in the Delayed Posttest. Then, one-way between groups ANOVA and Tukey's *post hoc* test were employed to explore the differences of the Delayed Posttest results among the six groups of participants to compare the scores graded by the two different systems. If the results from both grading methods are similar, there should bring very little or no difference on the analysis results when one grading method are selected for further analysis. The same result as the Immediate Posttest was found from the Analysis of Variance (see Table 5.11). There was a statistically significant difference among the six groups in both sensitive scoring: $F$ (5, 241) = 22.46, $p$ = < .001 and strict rating: $F$ (5, 241) = 22.41, $p$ = < .001.

***Table 5.11. One-way between group ANOVA of the Delayed Posttest (sensitive and strict scoring)***

| Scoring | | Sum of Squares | df | Mean Square | F | Sig* |
|---|---|---|---|---|---|---|
| **sensitive** | Between Groups | 661.8 | 5 | 132.37 | 22.46 | < .001* |
| | Within Groups | 1420.2 | 241 | 5.89 | | |
| **strict** | Between Groups | 702.9 | 5 | 140.57 | 22.41 | < .001* |
| | Within Groups | 1511.6 | 241 | 6.27 | | |

*Note: *p < .05*

After that, the *post hoc* test was used to discover the differences among these groups. This way, I could compare the results from both rating types which to some extent brought benefits to the main inferential analysis. The results from both sensitive scoring and strict scoring were the same, except the difference between the *Retrieval* and the *Generative Use* groups. While the sensitive result showed that there was no statistically significant difference between these two groups ($p$ = 0.06, 95% C.I. = [0.02, -3.08]), the difference was found from the strict scoring ($p$ = 0.04, 95% C.I. = [-0.04, -3.23]) in which a partial correction is unacceptable. However, the $p$ value of the strict scoring was just a little greater than that of the sensitive scoring. Also, it can be noticed that the Standard Deviation (*SD*) of the strict scores (2.11) was higher although not much different than that of the sensitive scores (2.00). This tended to show that the scores graded by strict scoring were more spread out than graded by sensitive scoring.

Overall, the Delayed Posttest yielded similar results to the Immediate Posttest results in most aspects. The *Control* group (Sensitive: Mean = 3.49; *SD* = 1.85; Strict: Mean = 3.13; *SD* = 2.06) was significantly different from all experimental groups again showing that the treatments used with the experimental groups influenced word gain and retention (see more details in Section 5.2.2.1 and Section 5.2.2.2). Similar to the Immediate Posttest, the *post hoc* test revealed that the *All TFA Components* group (Sensitive: Mean = 7.89; *SD* = 2.00; Strict: Mean = 7.65; *SD* = 2.11) was not significantly different from the *Noticing* (Sensitive: Mean = 7.84; *SD* = 2.67; Strict: Mean = 7.63; *SD* = 2.74) and *Generative Use* (Sensitive: Mean = 6.70; *SD* = 3.38; Strict: Mean = 6.44; *SD* = 3.39) groups, and that the *Motivation* group had significantly lower mean score (Sensitive: Mean = 5.51; *SD* = 1.99; Strict: Mean = 5.29; *SD* = 2.03) than other experimental groups in both scoring scheme. While the *All TFA Components* group was not different from the *Retrieval* group in the Immediate Posttest, it had significantly lower score (Sensitive: Mean = 7.89; *SD* = 2.00; Strict: Mean = 7.65; *SD* = 2.11) than the *Retrieval* group (Sensitive: Mean = 8.23; *SD* = 2.34; Strict: Mean = 8.07; *SD* = 2.43) in the Delayed Posttest in both scoring systems. This is because the *Retrieval* group got the highest Mean score in the Delayed Posttest (Mean = 8.23, *SD* = 2.34 for sensitive rating; Mean = 8.07, *SD* = 2.43 for strict rating) which was higher than its Immediate Posttest score (Mean = 6.67, *SD* = 2.62 for sensitive rating; Mean = 6.50, *SD* = 2.67 for strict rating). This unusual pattern was discussed in the Chapter 6: Discussion. Furthermore, in both scoring systems it was found that the *Motivation* group (Sensitive: Mean = 5.51; *SD* = 1.99; Strict: Mean = 5.29; *SD* = 2.03) was not significantly different from the *Generative Use* group (Sensitive: Mean = 6.70; *SD* = 3.38; Strict: Mean = 6.44; *SD* = 3.39) which is different from the Immediate Posttest result. Apart from that, a similar pattern from the Immediate Posttest was found in the Delayed Posttest as mentioned earlier (see also Figure 5.1 above).

To conclude, the results of both scoring schemes being employed were almost the same in both the immediate post-test and delayed post-test. In the following sections (Section 5.2.2.1 and Section 5.2.2.2), I will report results from both scoring schemes. However, I would focus more on to the sensitive scoring system for two reasons. First, the Standard Deviation (*SD*) of sensitive rating scores in all tests was obviously lower than that of the strict rating scores. This means that the strict scores tended to be more spread out and less reliable than the sensitive scoring data. Moreover, the sensitive scoring criteria are more precise than the strict scoring in which some mistakes such as minor misspellings were accepted and rewarded with either 0.25 or 0.5 point instead of zero. Also, some items in the test contained irregular forms of words. Accepting minor or some major mistakes could bring more accurate findings and interpretation to the current study. For example, the word '*encouraged*' in the Form Recall Test required the past (*-ed*) form that tended to involve grammatical knowledge in order to get one point for the strict rating. Yet, if the participants wrote the word '*encourage*' without the past (*-ed*) form in the answer, it could be interpreted that learning occurs to some extent and a score of 0.5 instead of zero point should be rewarded since they could write the present form of the word. For these reasons, the data from sensitive rating was selected for the main analysis. This could also help with the obstacle regarding time limitation of the present study. The results from this analysis compared with the results analysed by the Mixed-effects Model were explained in more details and reported in line with the research questions of the current study in Section 5.2.2 below.

#### 5.2.1.5 Summary of the descriptive statistics from sensitive scores

Due to the provided reasons, sensitive scoring data was focused in the summary of the descriptive statistics. For a clearer understanding, this section reported the summary of sensitive scores analysis by showing Mean and *Standard Deviation* (*SD*) in Table 5.12 as well as visualising the analysed data through a boxplot (see Figure 5.2). Although all groups had similar pre-test score, the *Control* group had lower scores on the two Posttests than the experimental groups. Among the experimental groups, the Mean score of the *Motivation* group were the lowest in both posttests. The Delayed Posttest scores of the *Control*, *Motivation*, *Noticing*, and *All TFA Components* groups dropped from the Immediate Posttest while the scores of the *Noticing* and *Retrieval* groups increased. Due to the increase, the *Retrieval* group (TFA score = 8) got the highest Mean scores among all groups, which was higher than the *All TFA Components* group that had the highest TFA score (TFA score = 15). Yet, the Mean scores cannot indicate the significant difference among these groups. The *post hoc* test was used and the results were reported in Section 5.2.2.1 and Section 5.2.2.2.

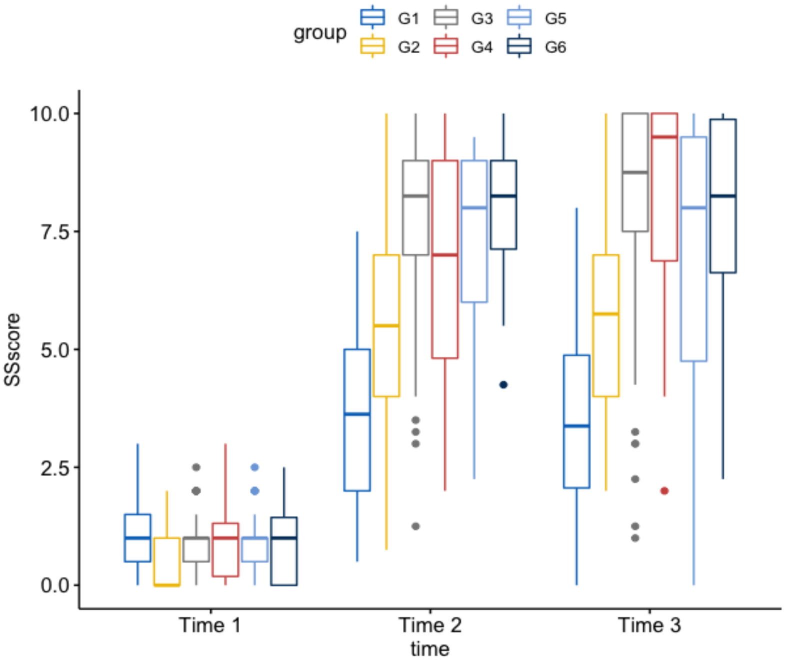
***Table 5.12. Mean sensitive scores (and Standard Deviations) on the Form Recall test (sensitive scores)***

| Group | Pre-test | Immediate Posttest | Delayed Posttest |
|---|---|---|---|
| Control *(n=38)* | 1.05 (0.75) | 3.61 (1.83) | 3.49 (1.85) |
| Motivation *(n=41)* | 0.63 (0.73) | 5.59 (2.38) | 5.51 (1.99) |
| Noticing *(n=41)* | 0.98 (0.67) | 7.73 (2.24) | 7.84 (2.67) |
| Retrieval *(n=40)* | 1.03 (0.83) | 6.67 (2.62) | 8.23 (2.34) |
| Generative Use *(n=41)* | 0.84 (0.63) | 7.36 (1.85) | 6.70 (3.38) |
| All TFA Components *(n=46)* | 0.91 (0.79) | 8.18 (1.36) | 7.89 (2.00) |

*Note: Total score is 10; n = number of participants*

To visualise the data in R, the package “ggpubr” (Kassambara, 2020) was used. This package allows me to use the ‘ggboxplot()’ function that helps to create easily publication ready plots and identify outliers from the dataset. The Figure 5.2 below illustrates the boxplot of the participants’ sensitive score across the three time periods. According to the results mentioned in Section 5.2.1.2, the Pre-test (Time 1) showed no significant difference among the six groups. Figure 5.2 shows the similar pattern of scores among these groups on the Pre-test. Also, it was found that the treatments led to higher mean scores in both Posttests (Time 2 and Time 3). All groups, except the Control group got considerably higher Posttest scores than the Pre-test scores even though some groups such as the *Motivation* (Mean = 0.63; *SD* = 0.73) and the *Noticing* (Mean = 0.98; *SD* = 0.67) groups got lower mean scores than the *Control* group (Mean = 1.05; *SD* = 0.75) in the Pre-test (see also Figure 5.2). Among the five experimental groups, the *Motivation* group tended to improve less than the others. While the Immediate (Time 2) and Delayed (Time 3) posttest scores of the *Motivation* group (Time 2: Mean = 5.59; *SD* = 2.38; Time 3: Mean = 5.51; *SD* = 1.99) increased considerably after receiving the treatment, its scores were significantly lower than those of the *Noticing* group

(Time 2: Mean = 7.73; *SD* = 2.24; Time 3: Mean = 7.84; *SD* = 2.67) and *All TFA Components* group (Time 2: Mean = 8.18; *SD* = 1.36; Time 3: Mean = 7.89; *SD* = 2.00) in both tests. Unlike the other groups, the *Retrieval* group (Time 2: Mean = 6.67; *SD* = 2.67; Time 3: Mean = 8.23; *SD* = 2.34) and *Noticing* group (Time 2: Mean = 7.73; *SD* = 2.24; Time 3: Mean = 7.84; *SD* = 2.67) got higher mean scores in the Delayed Posttest than the Immediate Posttest. This unusual upward trend was explained in the discussion chapter (Chapter 6).



*Note: G1 = Control group, G2 = Motivation group, G3 = Noticing Group, G4 = Retrieval group, G5 = Generative Use group, G6 = All TFA components group*

**Figure 5.2. A boxplot of sensitive score across three time periods**

The boxplot showed that there were outliers that can affect the interpretation of the results. I then removed all the outliers to see whether this would affect the result of the analysis. However, the modified data set that resulted from excluding was no different from the original dataset. Due to the concern on the number of the participants, I decided to keep the outliers as the removal of the outliers seemed to bring more drawbacks than benefits to the current study. Although the mean scores of the experimental groups in both Posttests were considerably higher than the control groups after receiving treatments, this tended to be normal among experimental studies that collect data at different points in time. The line graph (Figure 5.1) and boxplot (Figure 5.2) showed that the data analysis was robust even though a few assumptions seemed to be violated (see more details in Section 4.7.3.2).

### 5.2.2 Results of the inferential statistics analysis of the FRt

This section presents the results of the inferential statistics with the linear Mixed-effects Models. Results of the analysis enables us to address the two research questions. This section will present the findings related to each question in turn.

#### 5.2.2.1 Research question one: the effectiveness of Technique Feature Analysis (TFA) Framework on retention of word form

**Research Question 1:** *Do activities with high TFA scores result in better retention of single words in productive form recall?*

As explained in Section 3.3 and Section 3.4, the recent study was designed to compare one Control group (no TFA score) with five experimental groups with different TFA scores. There was no vocabulary activity in the selected unit of learning so that the Control group was not treated by any vocabulary components in the TFA framework. Thus, the TFA framework could not be applied to the Control group. The *All TFA Component* group received the highest TFA score (TFA = 15), followed by *Generative Use* (TFA = 9) and *Retrieval* (TFA = 8), respectively. *Motivation* and *Noticing* had the same lowest TFA score of 7 when compared to the other groups (see also Table 5.1).
To answer the first research question, I compared the Form Recall test (FRt) scores with the TFA scores of all groups. As the two scoring schemes were used, results related to the sensitive scheme is presented first, followed by the strict scheme.

**Sensitive scoring scheme**

Table 5.13 presents the mean scores of the six groups on three testing times using the sensitive scores. Regardless of the group, there was always an increase in the testing scores from Pre-test to Immediate Posttest, and from Pre-test to Delayed Posttest. The mean Pre-test scores among the six groups showed a similar result in that the participants did not have adequate knowledge of the target single words in the test. The increase of both posttests scores could represent the gain and retention of vocabulary knowledge after the learning. However, there seem to be different in the amount of word gain according to the posttest mean scores.

***Table 5.13. Mean Sensitive Scores (and Standard Deviations) of the Pre-test, Immediate Posttest, and Delayed Posttest on the Form Recall test***

| Group | n | TFA score | Pre-test | Immediate Posttest | Delayed Posttest |
|---|---|---|---|---|---|
| Control | 38 | n/a | 1.05 (0.75) | 3.61 (1.83) | 3.49 (1.85) |
| Motivation | 41 | 7 | 0.63 (0.73) | 5.59 (2.38) | 5.51 (1.99) |
| Noticing | 41 | 7 | 0.98 (0.67) | 7.73 (2.24) | 7.84 (2.67) |
| Retrieval | 40 | 8 | 1.03 (0.83) | 6.67 (2.62) | 8.23 (2.34) |
| Generative Use | 41 | 9 | 0.84 (0.63) | 7.36 (1.85) | 6.70 (3.38) |
| All TFA Components | 46 | 15 | 0.91 (0.79) | 8.18 (1.36) | 7.89 (2.00) |

*Note: n = number of participants; Total TFA score is 18; Total test score is 10; n/a = not applicable as there is no vocabulary activity used in this group*

Results of the linear mixed effects model analysis provides further insights into the significance of the differences. Table 5.14 presents the results of the model with the Control group and Pre-test as the reference for group and time, respectively. The total model which concerns both the fixed and random effects explained 74% of the variance (conditional $R^2$ = .74) while the fixed effects (group, time, and group by time interaction) explained 72% of the variance in the scores (marginal $R^2$ = .72). From the second to sixth rows of Table 5.14, group does not show as a main effect ($p$ > .05) as the scores of each group comprise all three tests together. However, the seventh and eighth rows of Table 5.14 shows that time has significant main effects on the participants' test score ($p$<.001). Their mean scores on the Immediate Posttest were significantly higher than those on the Pre-test (b = 2.55, SE = 0.41, $p$ < .001).

***Table 5.14. Linear Mixed-effects Model comparing the sensitive scores of FRt over the three testing times (Pre-test as Reference Level)***

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 1.05 | 0.30 | [0.46, 1.65] | 733.06 | 3.47 | **0.001** |
| Group (Control vs. Motivation) | -0.42 | 0.42 | [-1.24, 0.41] | 733.06 | -0.99 | 0.32 |
| Group (Control vs. Noticing) | -0.08 | 0.42 | [-0.90, 0.75] | 733.06 | -0.18 | 0.85 |
| Group (Control vs. Retrieval) | -0.02 | 0.42 | [-0.85, 0.81] | 733.06 | -0.05 | 0.96 |
| Group (Control vs. Generative Use) | -0.22 | 0.42 | [-1.04, 0.61] | 733.06 | -0.52 | 0.61 |
| Group (Control vs. All TFA Components) | -0.15 | 0.40 | [-0.95, 0.66] | 733.06 | -0.35 | 0.72 |
| Time (Pre-test vs. Immediate Posttest) | 2.55 | 0.41 | [1.74, 3.36] | 493.99 | 6.18 | **< .001** |
| Time (Pre-test vs. Delayed Posttest) | 2.43 | 0.41 | [1.62, 3.25] | 493.99 | 5.90 | **< .001** |
| Group (Control vs. Motivation): Time (Pre-test vs. Immediate Posttest) | 2.40 | 0.57 | [1.28, 3.53] | 493.99 | 4.20 | **< .001** |
| Group (Control vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 4.20 | 0.57 | [3.08, 5.33] | 493.99 | 7.33 | **< .001** |
| Group (Control vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | 3.08 | 0.58 | [1.95, 4.22] | 493.99 | 5.35 | **< .001** |
| Group (Control vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | 3.97 | 0.57 | [2.85, 5.10] | 493.99 | 6.93 | **< .001** |
| Group (Control vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 4.72 | 0.56 | [3.62, 5.81] | 493.99 | 8.46 | **< .001** |
| Group (Control vs. Motivation): Time (Pre-test vs. Delayed Posttest) | 2.44 | 0.57 | [1.32, 3.57] | 493.99 | 4.26 | **< .001** |
| Group (Control vs. Noticing): Time (Pre-test vs. Delayed Posttest) | 4.43 | 0.57 | [3.30, 5.55] | 493.99 | 7.72 | **< .001** |
| Group (Control vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 4.77 | 0.58 | [3.63, 5.90] | 493.99 | 8.27 | **< .001** |
| Group (Control vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | 3.43 | 0.57 | [2.31, 4.56] | 493.99 | 5.99 | **< .001** |
| Group (Control vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 4.55 | 0.56 | [3.45, 5.65] | 493.99 | 8.15 | **< .001** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

Similarly, the mean scores on the Delayed Posttest were significantly higher than those on the Pre-test (b = 2.43, SE = 0.41, *p* < .001). The remaining rows of the table shows that there were significant effects in the interaction between group and time (*p* < .001). Results of the pairwise comparison provide further insights into the interaction. Comparison of the scores of each group in the three testing time shows that regardless of the groups, participants' scores on the Pre-test always significantly lower than their scores on the Immediate Posttest (*p* < .001) and the Delayed Posttest (*p* < .001). This means that learning happened for all groups. The pairwise comparison of the Pre-test scores show no significant differences in the Pre-test score of the *Control* group and the Pre-test score of each experimental group: *Motivation* group (*p* > .05), *Noticing* group (*p* > .05), *Retrieval* group (*p* > .05), *Generative Use* group (*p* > .05), and *All TFA Components* group (*p* > .05). However, the pairwise comparison of the Immediate Posttest scores shows that the Control group had significantly lower scores on the Immediate Posttest than all experimental groups: *Motivation* group (*p* < .001), *Noticing* group (*p* < .001), *Retrieval* group (*p* < .001), *Generative Use* group (*p* < .001), and *All TFA Components* group (*p* < .001). Similarly, the pairwise comparison of the Delayed Posttest scores shows that the Control group had significantly lower scores on the Delayed Posttest than all experimental groups: *Motivation* group (*p* < .001), *Noticing* group (*p* < .001), *Retrieval* group (*p* < .001), *Generative Use* group (*p* < .001), and *All TFA Components* group (*p* < .001). The results were in line with the line graph of mean scores by group and time (see also Figure 5.1) presented in Section 5.2.1.1. These results means that the participants from the Control group had similar knowledge of the target words as those from the experimental groups before the treatment (*p* > .05) for all pairwise comparisons). However, after the treatment, the experimental groups had significantly greater gains than the Control group (*p* < .001) for all pairwise comparisons). This indicates that the TFA treatments tend to lead to learning gains.

**Strict scoring scheme**

As described in Section 4.7.1 (see Table 4.5 for the evaluation criteria), for the strict scoring scheme, partially correct answers were not acceptable and graded as an incorrect answer in the current study. Table 5.15 illustrates the mean scores of the six groups on three testing times by using strict scoring criteria.

***Table 5.15. Mean Strict Scores (and Standard Deviations) of the Pre-test, Immediate Posttest, and Delayed Posttest on the FRt***

| Group | n | TFA score | Pre-test | Immediate Posttest | Delayed Posttest |
|---|---|---|---|---|---|
| *Control* | 38 | n/a* | 0.84 (0.79) | 3.16 (2.09) | 3.13 (2.06) |
| *Motivation* | 41 | 7 | 0.63 (0.73) | 4.85 (2.89) | 5.29 (2.03) |
| *Noticing* | 41 | 7 | 0.61 (0.77) | 7.61 (2.29) | 7.63 (2.74) |
| *Retrieval* | 40 | 8 | 0.90 (0.84) | 6.50 (2.67) | 8.07 (2.43) |
| *Generative Use* | 41 | 9 | 0.61 (0.70) | 7.05 (1.83) | 6.44 (3.39) |
| *All TFA Components* | 46 | 15 | 0.85 (0.79) | 7.93 (1.48) | 7.65 (2.11) |

*Note: n = number of participants; Total TFA score is 18; Total test score is 10; n/a* = not applicable as there is no vocabulary activity used in this group (see Section 3.3 for details)*

Even though the mean scores by using strict rating were lower than those of the sensitive rating in all tests, the two scoring scheme showed the same pattern of results. Similar to the sensitive scoring, the upward trend was found from the Pre-test scores to Immediate Posttest, and from Pre-test to Delayed Posttest regardless of the group. Likewise, the experimental groups had higher mean scores than the Control group regardless of the test. The Pre-test strict scores also signify the lack of background knowledge on the form of the target words of the participants. The Immediate Posttest and Delayed Posttest scores of the two scoring scheme tended to be consistent in that the Control group had the lowest mean scores in both posttests compared to the experimental groups: *Motivation, Noticing, Retrieval*, and *Generative Use* (see Table 5.15).

The insight into the differences of strict scores between the six groups was investigated further by using the linear mixed effects model analysis. Table 5.16 shows similar results to the sensitive scoring in all aspects. However, the model presents lower coefficient values: Conditional $R^2$ and Marginal $R^2$. It was found that the model with the Control group and Pre-test as the reference for group and time, which concerns both the fixed and random effects, explained 72% of the variance (conditional $R^2$ = .72) while the fixed effects (group, time, and group by time interaction) explained 70% of the variance in the scores (marginal $R^2$ = .70). According to the results from the second and third rows of Table 5.4, time has significant effects on the participants' test score ($p$ < .001). Their strict mean scores on the Pre-test were significantly lower than those on both the Immediate Posttest (b = 2.32, SE = 0.44, $p$ <.001) and the Delayed Posttest (b = 2.29, SE = 0.44, $p$ <.001). From the fourth row onwards, the results show the significant effects in the interaction between group and time (all $p$ < .001). Further insights into the interaction based on the results of the pairwise comparison confirm that the participants in all groups had the same background knowledge on vocabulary before the experiment. The comparison of the Pre-test strict scores between groups showed no significant difference between the Control group and each experimental group: *Motivation* group ($p$ > .05), *Noticing* group ($p$ > .05), *Retrieval* group ($p$ > .05), *Generative Use* group ($p$ > .05), and *All TFA Components* group ($p$ > .05). However, the comparison between the Pre-test and Immediate Posttest scores, and Pre-test and Delayed Posttest scores shows similar results in that the Control group had significantly lower scores than all experimental groups on both tests: *Motivation* group (Pre vs. Immediate = $p$ < .001; Pre vs. Delayed = $p$ < .001), *Noticing* group (Pre vs. Immediate = $p$ < .001; Pre vs. Delayed = $p$ < .001), *Retrieval* group (Pre vs. Immediate = $p$ < .001; Pre vs. Delayed = $p$ < .001), *Generative Use* group (Pre vs. Immediate = $p$ < .001; Pre vs. Delayed = $p$ < .001), and *All TFA Components* group (Pre vs. Immediate = $p$ < .001; Pre vs. Delayed = $p$ < .001).

***Table 5.16. Linear Mixed-effects Model comparing the strict scores of FRt over the three testing times (Pre-test as Reference Level)***

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.84 | 0.32 | [0.22, 1.47] | 734.85 | 2.64 | **< .001** |
| Group (Control vs. Motivation) | -0.21 | 0.44 | [-1.24, 0.41] | 734.85 | -0.47 | 0.64 |
| Group (Control vs. Noticing) | -0.23 | 0.44 | [-0.90, 0.75] | 734.85 | -0.52 | 0.60 |
| Group (Control vs. Retrieval) | 0.56 | 0.45 | [-0.85, 0.81] | 734.85 | 0.13 | 0.90 |
| Group (Control vs. Generative Use) | -0.23 | 0.44 | [-1.04, 0.61] | 734.85 | -0.52 | 0.60 |
| Group (Control vs. All TFA Components) | 0.01 | 0.43 | [-0.95, 0.66] | 734.85 | 0.01 | 0.99 |
| Time (Pre-test vs. Immediate Posttest) | 2.32 | 0.44 | [1.46, 3.17] | 494.00 | 5.30 | **< .001** |
| Time (Pre-test vs. Delayed Posttest) | 2.29 | 0.44 | [1.43, 3.15] | 494.00 | 5.24 | **< .001** |
| Group (Control vs. Motivation): Time (Pre-test vs. Immediate Posttest) | 1.90 | 0.61 | [0.71, 3.09] | 494.00 | 3.14 | **< .001** |
| Group (Control vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 4.68 | 0.61 | [3.49, 5.87] | 494.00 | 7.73 | **< .001** |
| Group (Control vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | 3.28 | 0.61 | [2.09, 4.48] | 494.00 | 5.38 | **< .001** |
| Group (Control vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | 4.12 | 0.61 | [2.93, 5.31] | 494.00 | 6.80 | **< .001** |
| Group (Control vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 4.77 | 0.59 | [3.61, 5.93] | 494.00 | 8.08 | **< .001** |
| Group (Control vs. Motivation): Time (Pre-test vs. Delayed Posttest) | 2.37 | 0.61 | [1.18, 3.56] | 494.00 | 3.91 | **< .001** |
| Group (Control vs. Noticing): Time (Pre-test vs. Delayed Posttest) | 4.73 | 0.61 | [3.54, 5.93] | 494.00 | 7.81 | **< .001** |
| Group (Control vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 4.89 | 0.61 | [3.69, 6.08] | 494.00 | 8.01 | **< .001** |
| Group (Control vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | 3.54 | 0.61 | [2.35, 4.73] | 494.00 | 5.84 | **< .001** |
| Group (Control vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 4.51 | 0.59 | [3.36, 5.67] | 494.00 | 7.65 | **< .001** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

The evidence demonstrates the positive effects of the TFA treatments on learning gains due to the fact that participants' scores on the Pre-test always significantly lower than their scores on the Immediate Posttest (*p* < .001) and the Delayed Posttest (*p* < .001) regardless of the groups. However, the mean scores between the Immediate Posttest and Delayed Posttet were not significantly different (sensitive: b = -0.08, SE = 0.41, *p* = 0.85; strict: b = -0.03, SE = 0.44, *p* = 0.95). This means that the effect of *time* was mainly due to the differences between the Pre-test and Immediate Poosttest and Pre-test and Delayed Posttest.

Taken together, analysis of the experimental data using both scoring schemes consistently indicated that the five treatments (activities using materials with high TFA scores) resulted in better retention of words in terms of written form recall. The findings help to address the following sub-questions of the Research Question 1. It should be noted that the information below refers to both sensitive and strict scores owing to the indifferent results. The Research Question 1 concerns both short-term and long-term retention of word form resulted from both Immediate and Delayed Posttests scores. The first sub research question paid particular attention to short-term retention.

> *a) Do the TFA-supported groups result in better short-term retention of single words in terms of written form recall than the Control group?*

The Immediate Posttest scores of all experimental groups increased significantly after doing activities with high TFA scores, ranging between 7 and 15 points out of 18 points. The participants in the *Motivation* (Mean = 5.59), *Noticing* (Mean = 7.73), *Retrieval* (Mean = 6.67), and *Generative Use* (Mean = 7.36) could achieve greater scores than the Control group (Mean = 3.61) with no support of TFA components. The results clearly show that the participants in all experimental groups could retain the target words in their short-term store. This means that activities modified by using the TFA framework can effectively promote short-term retention of word form.

Then, the results from the Delayed Posttest were compared with the TFA scores to address the following sub research question related to long-term retention:

*b) Do the TFA-supported groups result in better long-term retention of single words in terms of written form recall than the Control group?*

The Delayed Posttest scores represent long-term retention of word form stored in the participants' memory two weeks after the experiment. The findings were similar to the Immediate Posttest in that the Control group with no TFA support could not perform well in this posttest (Mean = 3.49). The pairwise comparison explained that the treatment groups: *Motivation* (Mean = 5.51), *Noticing* (Mean = 7.84), *Retrieval* (M = 8.23), *Generative Use* (Mean = 6.70), and *All TFA Components* (Mean = 7.89) with high TFA support showed significant higher scores than the Control group ($p$ < .001). Regardless of scoring scheme, the results can confirm that activities with high TFA scores also lead to better long-term retention of single words in terms of written form recall.

In summary, findings from this empirical study shed light on the effectiveness of the TFA framework towards short-term and long-term retention of single-word form in written production. The groups that were treated with high TFA support materials could achieve considerable higher scores on both Immediate and Delayed Posttests than the Control group. This indicates that with high TFA support EFL learners can better recall word form in written production and retain the knowledge to recall later. Further insights into the effects of each TFA component on written form recall were reported in the following section to address the second research question.

#### 5.2.2.2 Research question two: the impact of Motivation, Noticing, Retrieval, and Generative Use on short-term and long-term retention of controlled productive knowledge

**Research Question 2:** *What is the individual role of the different TFA components on promoting vocabulary retention through vocabulary tasks?*

Analysis related to Research Question 1 has indicated that all groups with high TFA support materials could achieve significantly higher scores on both Immediate and Delayed Posttests than the group with non-TFA support materials. However, it is unclear among the groups with high TFA support materials, which groups had significantly higher scores than the other. Therefore, Research Question 2 concerns the effects of *Motivation, Noticing, Retrieval, and Generative Use* on both short-term and long-term retention on controlled productive knowledge. The results regarding short-term retention and long-term retention were presented together with the two scoring schemes (sensitive and strict) which reported minor differences. So, I explain the details of both schemes in this section.

According to Table 5.13 and Table 5.15 above, all experimental groups showed positive effects on the retention of word form. The participants in these groups could retain the knowledge of form to achieve the Immediate Posttest. However, the relative impact of the different components on vocabulary recall is quite complicated so I will describe the findings by first looking at short-term recall sensitive scores and strict scores, respectively.

### Sensitive scoring scheme

The teaching materials which were adapted to facilitate all four TFA components (*Motivation*, *Noticing*, *Retrieval*, and *Generative Use*) led to the highest short-term retention. The results from the Immediate Posttest (sensitive) score in the fifth column of Table 5.13 showed that *Noticing* received the highest mean score (Mean = 7.73, *SD* = 2.24) among the four groups, followed by *Generative Use* (Mean = 7.36, *SD* = 1.85), *Retrieval* (Mean = 6.67, *SD* = 2.62), and *Motivation* (Mean = 5.59, *SD* = 2.38), respectively. The significance of the differences among these components was explained according to the results of the linear Mixed-effects Model analysis. The model with the *Motivation* group and Pre-test as the reference for group and time, respectively shows the analysis results in Table 5.17. I also checked consistency of the results by relevelling the model using each TFA component as the reference level. The linear mixed effects model analysis of sensitive scores showed similar results as when the *Motivation* group was set to be the reference level, so these results were not reported again (see Appendix 8a for details).

The results from the second to fifth rows of Table 5.17 showed that group was not the main effect (*p* > .05 in all rows) because the analysis combined Pre-test and posttests scores of each group together. However, the sixth row presented the effects of time (Pre-test vs. Immediate Posttest) on the participants' test score (*p* < .001). It also showed that Pre-test sensitive score was significantly lower than that of the Immediate Posttest (b = 4.96, SE = 0.40, *p* <.001). The pairwise comparison brings further insights into the impact of each component on short-term retention on form recall. The results show that the *Motivation* group had the least effect on word retention with the smallest effect on retention of form recall and significantly lower than the *Noticing* (*p* < 0.01) and *Generative Use* (*p* < 0.01) materials. However, the pairwise comparison confirmed that the effect of the motivation materials (M = 5.59, *SD* = 2.38) was not significantly different from the *Retrieval* group (M = 6.67, *SD* = 2.62) in the Immediate Posttest (*p* > .05).

When *Noticing* was set as reference level, pairwise comparison results shed light to the significance of the difference in that *Noticing,* though receiving the highest mean score (Mean = 7.73, SD = 2.24) among all groups, was only different from *Motivatio*n (*p* < .05), but not significantly better than the *Retrieval* (*p* > .05) and *Generative Use* (*p* > .05) adaptations. This is correlated to the results from pairwise comparisons of the models setting *Retrieval* and *Generative Use* as reference level. While *Retrieval* had the same effect on form recall as *Motivation*, there was only *Retrieval* that showed no difference from the *Noticing* (Mean = 7.73, SD = 2.24) and *Generative Use* (Mean = 7.36, SD = 1.85) in the Immediate Posttest (*p* > .05). This means that *Motivation* supported short-term retention of controlled productive knowledge to the same extent as *Retrieval*, but less than *Noticing* and *Generative Use* when sensitive scoring was taken into account. However, the analysis of strict score reported an inconsistent result between the *Motivation* and *Retrieval.* I will report this issue according to strict scoring results later in this section.

***Table 5.17. Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest sensitive scores of the experimental groups (Motivation group and Pre-test as Reference Level)***

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.63 | 0.30 | [0.04, 1.22] | 619 | 2.11 | **0.04** |
| Group (Motivation vs. Noticing) | 0.34 | 0.42 | [-0.49, 1.18] | 619 | 0.80 | 0.42 |
| Group (Motivation vs. Retrieval) | 0.40 | 0.43 | [0.44, 1.24] | 619 | 0.93 | 0.35 |
| Group (Motivation vs. Generative Use) | 0.20 | 0.42 | [-0.63, 1.03] | 619 | 0.47 | 0.64 |
| Group (Motivation vs. All TFA Components) | 0.27 | 0.41 | [-0.54, 1.08] | 619 | 0.66 | 0.51 |
| Time (Pre-test vs. Immediate Posttest) | 4.96 | 0.40 | [4.16, 5.76] | 494.00 | 12.15 | **< .001** |
| Group (Motivation vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 1.80 | 0.56 | [0.66, 2.93] | 494.00 | 3.12 | **< .01** |
| Group (Motivation vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | 0.68 | 0.57 | [-0.46, 1.82] | 494.00 | 1.17 | 0.24 |
| Group (Motivation vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | 1.57 | 0.56 | [0.43, 2.70] | 494.00 | 2.72 | **< .01** |
| Group (Motivation vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 2.31 | 0.55 | [1.21, 3.42] | 494.00 | 4.12 | **< .001** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant differenc

***Table 5.18. Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest strict scores of the experimental groups (Motivation group and Pre-test as Reference Level)***

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.63 | 0.31 | [0.02, 1.25] | 620 | 2.02 | **0.04** |
| Group (Motivation vs. Noticing) | -0.02 | 0.44 | [-0.90, 0.85] | 620 | -0.06 | 0.96 |
| Group (Motivation vs. Retrieval) | 0.27 | 0.45 | [-0.61, 1.14] | 620 | 0.60 | 0.55 |
| Group (Motivation vs. Generative Use) | -0.02 | 0.44 | [-0.90, 0.85] | 620 | -0.06 | 0.96 |
| Group (Motivation vs. All TFA Components) | 0.21 | 0.43 | [-0.63, 1.06] | 620 | 0.50 | 0.62 |
| Time (Pre-test vs. Immediate Posttest) | 4.22 | 0.43 | [3.38, 5.06] | 418 | 9.90 | **< .001** |
| Group (Motivation vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 2.78 | 0.60 | [1.60, 3.97] | 418 | 4.61 | **< .001** |
| Group (Motivation vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | 1.38 | 0.61 | [0.19, 2.57] | 418 | 2.27 | **0.02** |
| Group (Motivation vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | 2.22 | 0.60 | [1.03, 3.40] | 418 | 3.68 | **< .001** |
| Group (Motivation vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 2.87 | 0.59 | [1.72, 4.02] | 418 | 4.89 | **< .001** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

After the sensitive scoring was investigated, I analysed the strict scores of the Immediate Posttest by setting the *Motivation* group as reference level for mixed effects model, which was similar to the model used for the analysis of sensitive scoring.

**Strict scoring scheme**

The strict scoring results in the fifth column of Table 5.15 has the same pattern as those of sensitive scoring in that *Noticing* received the highest Mean score (Mean = 7.61, *SD* = 2.29) among the four groups, followed by *Generative Use* (Mean = 7.05, *SD* = 1.83), *Retrieval* (Mean = 6.50, *SD* = 2.67), and *Motivation* (Mean = 4.85, *SD* = 2.89), respectively. The strict scores of all groups, except *Motivation* were not much different from the sensitive scores (see also Table 5.13). The results in Table 5.18 also show that strict Mean scores of the Immediate Posttest are mostly similar to its sensitive mean scores in Table 5.5, except the comparison between *Motivation* and *Retrieval,* presenting in the eighth row. The results in Table 5.18 explains the significance of the differences from the results of the linear mixed effects model that *Motivation* were set as the reference for group, and Pre-test was the reference level for time. In this analysis, each experimental group was also set in turn as the reference for group. Yet, the results were correlated to those of the *Motivation* model in all aspects (see Appendix 8b). Table 5.6 (from the second to fifth rows) shows similar results to sensitive scoring in that group was not the main effects (*p* > .001 in all rows). As mentioned earlier, the result of each group is based on all three tests in combination. However, time still has an impact on the strict scores (b = 4.22, SE = 0.43, *p* < .001). A slight difference was found from further analysis into the differences among these groups.

Pairwise comparisons showed that there were significant differences between the *Motivation* and other TFA components groups: *Noticing* (*p* < .001), *Retrieval* (*p* < .05) and *Generative Use* (*p* < .001). When the test was strictly graded by using the zero-or-one scoring scheme, the score of *Motivation* was considerable lower than the other groups. This can imply that this group could not retain as precise knowledge as the other experimental groups (*Noticing, Retrieval,* and *Generative Use*) in the Immediate Posttest. The difference between *Motivation* and *Retrieval* was confirmed again by the analysis of the

*Retrieval* model. Pairwise comparison explained the superior effects of *Retrieval* over *Motivation* (*p* < .001). The participants who were facilitated with material promoting retrieval could retain form of the target words more correctly than those in the *Motivation* group.

To sum, if full knowledge was considered, *Motivation* showed less effects on short-term retention than the other TFA components. However, *Motivation* tends to have similar power to *Retrieval*, but less than *Noticing* and *Generative Use* when partial knowledge was taken into account. Regardless of scoring scheme, *Retrieval* supported short-term retention of controlled productive knowledge to the same extent as *Noticing* and *Generative Use.*

In terms of long-term retention, sensitive scores and strict scores were presented, respectively to explore the effects of *Motivation, Noticing, Retrieval* and *Generative Use*. Results of mean scores from the Delayed Posttest provide evidence of long-term word retention. Similar to those of the Immediate Posttest, both sensitive and strict scores significantly increased from the Pre-test (see the final column of Table 5.13 and Table 5.15). This means that all experimental groups could retain long-term knowledge of word form for at least two weeks after the experiment. I will first describe the findings of sensitive scores below.

**Sensitive scoring scheme**

The materials designed to support TFA components also showed positive effects on long-term retention of word form. The Delayed Posttest and Immediate Posttest results were corelated in most aspects, except the results of *Retrieval.* The final column of Table 5.13 showed that the *Motivation* group also had the lowest mean score (Mean = 5.51, *SD* = 1.99) while the *Retrieval* had gained the highest mean score (Mean = 8.23, *SD* = 2.34) on this test compared to the other intervention groups. Following to the highest score were *Noticing* (Mean = 7.84, *SD* = 2.67) and *Generative Use* (Mean = 6.70, *SD* = 3.38), respectively. Further insights into the significance of the differences among these components were also explored by using the linear mixed effects model analysis. *Motivation* and Pre-test were set as the reference level of group and time, respectively. Similar to the analysis of Immediate Posttest, the *Motivation* as reference for group was replaced by other experimental groups to check consistency of the results of sensitive scoring. All models showed the same analysis results as those of *Motivation* (see Appendix 9a) and so the results of the other models were not reported again. According to the results of the *Motivation* model (see Table 5.19 from second to fifth row),  there was no effect of group (*p* > .05 in all rows) as each group contained scores from the three tests: Pre-test, Immediate Posttest, and Delayed Posttest. Similar to the Immediate Posttest results, the model explained that there was a significant effect of time on the test scores (b = 4.88, SE = 0.40, *p* <.001) as presented in the sixth row. Pairwise comparison results shed light on the differences among the four groups in the Delayed Posttest in that *Motivation* was significant lower from the *Noticing* (*p* < .001) and *Retrieval* (*p* < .001), but not significant different from the *Generative Use* (*p* > .05), meaning that materials for motivation and generative use can support long-term retention to the same extent. The comparison results of the *Retrieval* model is also consistent to those of the *Motivation, Noticing* and *Generative Use* in that *Retrieval* materials had a greater impact on long-term retention than *Motivation* (*p* < .001) and *Generative Use* (*p* < .001) groups, but no difference from the *Noticing* (*p* > .05). This means that the TFA framework can encourage *Noticing* and *Retrieval* to the same extent, but better than *Motivation*.

I then checked the strict scoring results to compare with those of the sensitive results as reported below.

### Strict scoring scheme

Regardless of the group, there was always an increase in the strict scores from Pre-test to Delayed Posttest. The mean strict scores in the final column of Table 5.15 showed similar results to the mean sensitive scores (see also Table 5.13) in which *Motivation* (M = 5.29, *SD* = 2.03) had the lowest score among the four TFA groups while the *Retrieval* (M = 8.07, *SD* = 2.43) got highest mean scores, followed by *Noticing* (M = 7.63, *SD* = 2.74), and *Generative Use* (M = 6.44, *SD* = 3.39), respectively. Strict Delayed Posttest results of the linear mixed effects model analysis with *Motivation* and Pre-test as the reference for group and time, respectively were presented in Table 5.20. I also check consistency of results from models with different reference levels. *Noticing, Retrieval* and *Generative Use* was also changed as reference level for group in the analysis. Similar results were found from these models (see Appendix 9b) so that I reported here only the *Motivation* group results.

The model with the *Motivation* as reference for group presents the significant difference of time between strict Pre-test and Delayed Posttest scores (b = 4.66, SE = 0.42, *p* < .001) as shown in the sixth row of Table 5.20. This means that time is the main effect of the differences among the experimental groups. The remaining rows of the table present the significant effects (*p* < .005) in the interaction between group (*Motivation* vs. each experimental group) and time (Pre-test vs. Delayed Posttest) of strict scores. Results from pairwise comparison explained that the *Motivation* group was significant different from the *Noticing* (*p* < .001) and *Retrieval* (*p* < .001) groups, but not different from the *Generative Use* (*p* >.05). When relevelling *Noticing, Retrieval* and *Generative Use* as reference for group, pairwise comparisons presented the same results in that *Noticing* had the same effect as *Retrieval* and *Generative Use* (*p* >.05), but greater than *Motivation* (*p* < .05). Although *Generative Use* showed no difference to *Motivation* and *Noticing* (*p* > .05), its score was significant lower than those of the *Retrieval* (*p* < .05)*,* which is similar to the results of sensitive scores.

In brief, both scoring schemes, which consider full and partial knowledge, showed the same results of long-term retention on form recall. Regardless of scoring scheme, *Motivation* had less effects than other TFA components, except *Generative Use*. The results also explained that the materials designed to support *Retrieval* and *Noticing* can facilitate retention of word form to a large extent.

***Table 5.19. Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest sensitive scores of the experimental groups (Motivation group and Pre-test as Reference Level)***

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.63 | 0.30 | [0.04, 1.22] | 619 | 2.11 | **0.04** |
| Group (Motivation vs. Noticing) | 0.34 | 0.42 | [-0.49, 1.18] | 619 | 0.80 | 0.42 |
| Group (Motivation vs. Retrieval) | 0.40 | 0.43 | [0.44, 1.24] | 619 | 0.93 | 0.35 |
| Group (Motivation vs. Generative Use) | 0.20 | 0.42 | [-0.63, 1.03] | 619 | 0.47 | 0.64 |
| Group (Motivation vs. All TFA Components) | 0.27 | 0.41 | [-0.54, 1.08] | 619 | 0.66 | 0.51 |
| Time (Pre-test vs. Delayed Posttest) | 4.88 | 0.40 | [4.08, 5.68] | 418 | 12.27 | **< .001** |
| Group (Motivation vs. Noticing): Time (Pre-test vs. Delayed Posttest) | 1.98 | 0.8 | [0.85, 3.12] | 418 | 3.43 | **< .001** |
| Group (Motivation vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 2.32 | 0.58 | [1.18, 3.46] | 418 | 4.00 | **< .001** |
| Group (Motivation vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | 0.99 | 0.58 | [-0.15, 2.12] | 418 | 1.71 | 0.09 |
| Group (Motivation vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 2.11 | 0.56 | [1.00, 3.21] | 418 | 3.5 | **< .001** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

***Table 5.20. Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest strict scores of the experimental groups (Motivation group and Pre-test as Reference Level)***

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.63 | 0.31 | [0.02, 1.25] | 620 | 2.02 | **0.04** |
| Group (Motivation vs. Noticing) | -0.02 | 0.44 | [-0.90, 0.85] | 620 | -0.06 | 0.96 |
| Group (Motivation vs. Retrieval) | 0.27 | 0.45 | [-0.61, 1.14] | 620 | 0.60 | 0.55 |
| Group (Motivation vs. Generative Use) | -0.02 | 0.44 | [-0.90, 0.85] | 620 | -0.06 | 0.96 |
| Group (Motivation vs. All TFA Components) | 0.21 | 0.43 | [-0.63, 1.06] | 620 | 0.50 | 0.62 |
| Time (Pre-test vs. Delayed Posttest) | 4.66 | 0.42 | [3.82, 5.50] | 418 | 10.92 | **< .001** |
| Group (Motivation vs. Noticing): Time (Pre-test vs. Delayed Posttest) | 2.37 | 0.60 | [1.18, 3.56] | 418 | 3.92 | **< .001** |
| Group (Motivation vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 2.52 | 0.61 | [1.32, 3.71] | 418 | 4.15 | **< .001** |
| Group (Motivation vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | 1.17 | 0.60 | [-0.01, 2.36] | 418 | 1.94 | **0.05** |
| Group (Motivation vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 2.15 | 0.59 | [1.00, 3.30] | 418 | 3.66 | **< .001** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

The below information is provided with an attempt to answer the sub-questions of the Research Question 2:

a. To what extent does *Motivation* support (short-/long-term) retention of controlled productive knowledge?
b. To what extent does *Noticing* support (short-/long-term) retention of controlled productive knowledge?
c. To what extent does *Retrieval* support (short-/long-term) retention of controlled productive knowledge?
d. To what extent does *Generative Use* support (short-/long-term) retention of controlled productive knowledge?

According to the findings, the comparison between Pre-test and Immediate Posttest, and Pre-test and Delayed Posttest yielded a similar result in that *Motivation, Noticing, Retrieval*, and *Generative Use* contributed to the knowledge of form recall. They have fairly similar power on retention. The participants in the experimental groups (*Motivation*, *Noticing*, *Retrieval*, and *Generative Use*) were able to retain word forms after receiving the treatments. There was no significant gain of form recall when the test scores of the Immediate Posttest and Delayed Posttest were compared (*p* >.05), meaning that the learners could retain word knowledge for at least two weeks. The current empirical evidence confirms the positive effects of TFA components on both short-term and long-term retention. However, the effects of *Motivation, Noticing, Retrieval*, and *Generative Use* on retention of form recall were different to some extent as explained in Section 5.2.2.1 and Section 5.2.2.2. When full knowledge was considered the same results were found from different analysis models by using pairwise comparisons in that *Motivation* had the lowest short-term effect compared to *Noticing*, *Retrieval*, and *Generative Use* but had the same long-term effect to *Generative Use* only. If partial knowledge was taken into account, *Motivation* showed less significant short-term effect than, but the same long-term effect to *Generative Use (short-term retention: p < .05; long-term retention: p > .05)*. This means that the relative value of *Generative Use* varied depending on the kind of retention. *Generative Use* resulted in greater short-term retention than *Motivation* but not in long-term retention. Unlike *Retrieval*, materials for *Retrieval* could support long-term retention better than short-term retention. Furthermore, while *Noticing* and *Retrieval* showed the greatest scores on Immediate Posttest and Delayed Posttest, respectively, the amount of support of *Retrieval* was not significant different from the *Noticing* on both short-term (*p* >.05) and long-term (*p* >.05) retention, indicating that *Noticing* and *Retrieval* are likely to have the same power on retention of controlled productive knowledge but better than *Motivation*, especially on long-term retention.

The summary of comparison results to address the Research Question 2 were presented in Table 5.21. The table illustrates the similarities and differences in terms of support on form recall between the four vocabulary leaning components. The tick symbol (✓) represents the differences while the cross symbol (✕) refers to the indifferences in terms of amount of support between each pair. The first to the fourth rows present the comparison results of short-term retention on form recall between each experimental groups whereas the fifth rows onwards show the results of long-term retention. I will first explain the summary results of short-term retention regarding the two scoring scheme: sensitive and strict. The major difference was on *Retrieval.* With regard to sensitive scoring, the findings presented by the cross (✕) symbol (see Table 5.21) in the Short-term Retention section show that *Retrieval* and other TFA components (*Motivation, Noticing*, and *Generative Use*) can provide substantial support on short-term retention of form recall to the same extent. However, the remark symbol (s*) in the first row of Table 5.21 showed contrast results between sensitive scores and strict scores of the *Retrieval.* Apart from this, sensitive and strict scoring show similar results in all aspects, meaning that different scoring systems did not seem to be one of the factors affecting the results of the current study. This is in line with the information related to the reliability between sensitive and strict scoring schemes discussed in Section 4.7.1 (see also Table 4.5). Now, I move on to the summary results of long-term retention in Table 5.21 (see from the fifth rows onwards). As shown in the *Retrieval* column, the materials for retrieval show higher support on long-term retention than *Motivation* and *Generative Use.* This was due to a remarkable increase in the Delayed Posttest scores of the *Retrieval* group as explained earlier in this section. Therefore, *Retrieval* and *Noticing* have the same degree of support on long-term retention as presented by the cross symbol (✕) in the sixth and seventh rows of Table 5.21.

***Table 5.21. Comparison results between four TFA components evaluated by sensitive and strict scoring schemes***

| | | **Sensitive Scoring** | | | | **Strict Scoring** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Reference level** | *Motivation* | *Noticing* | *Retrieval* | *Generative Use* | *Motivation* | *Noticing* | *Retrieval* | *Generative Use* |
| **Short-term Retention** | *Motivation* | n/a | ✓ | ✗(s)* | ✓ | n/a | ✓ | ✓(s)* | ✓ |
| | *Noticing* | ✓ | n/a | ✗ | ✗ | ✓ | n/a | ✗ | ✗ |
| | *Retrieval* | ✗ | ✗ | n/a | ✗ | ✗ | ✗ | n/a | ✗ |
| | *Generative Use* | ✓ | ✗ | ✗ | n/a | ✓ | ✗ | ✗ | n/a |
| **Long-term Retention** | *Motivation* | n/a | ✓ | ✓ | ✗ | n/a | ✓ | ✓ | ✗ |
| | *Noticing* | ✓ | n/a | ✗ | ✗ | ✓ | n/a | ✗ | ✗ |
| | *Retrieval* | ✓ | ✗ | n/a | ✓ | ✓ | ✗ | n/a | ✓ |
| | *Generative Use* | ✗ | ✗ | ✓ | n/a | ✗ | ✗ | ✓ | n/a |

*Note:* ✓ *= difference in degree of support;* ✗ *= no difference in degree of support; (s)* = difference as a result of scoring system*

# Chapter 6
# Discussion

The previous chapter described the findings related to the two research questions. In this chapter, I will discuss these findings to yield more understanding and research contribution to the Technique Feature Analysis (TFA) framework and its components within the framework. Results from previous empirical studies and former knowledge presented in the literature review are discussed along with the implementations in terms of theory, pedagogy, and methodology. This chapter has five main sections. Section 6.1 discusses theoretical contributions regarding the impact of the TFA and other possible factors on word knowledge. It also covers the validity of TFA components: *Motivation*, *Noticing*, *Retrieval*, and *Generative Use* on form recall knowledge with the need for improving the conceptual framework as well as the theoretical implications. Section 6.2 describes pedagogy contributions and implications. Section 6.3 explores research contributions and implications in terms of methodology. Section 6.4 points out key issues on the limitations for pedagogy and methodology of the current study.

## 6.1. Theoretical contribution and implications

This section discusses the effects of Technique Feature Analysis (TFA) framework as a whole and its four components: *Motivation*, *Noticing*, *Retrieval*, and *Generative Use* on form recall knowledge. These effects were discussed in connection with research questions to explain how the findings conceptualise the framework and how the present study contributed to existing knowledge. Section 6.1.1 and Section 6.1.2 discuss the findings related to Research Question 1 and Research Question 2 in turn. Section 6.1.3 and Section 6.1.4 discuss suggestions for further development of the framework and other factors affecting retention of word form, respectively. As presented in Sections 5.2.1 of Chapter 5, the two scoring schemes led to similar findings in most cases. Therefore, in cases where the findings of the two schemes were the same, the discussion will refer to those from the sensitive scoring scheme only. It is because this scheme captures the influence of TFA on form recall knowledge more precisely than the strict scheme that only takes full knowledge into account. Also, the results of partial knowledge indicated that learning happened to some extent. However, in cases where different results were obtained from the two schemes, findings from both schemes were discussed.

### 6.1.1. The effectiveness of TFA as a conceptual framework for vocabulary learning

**Research Question 1:** *Do activities with high TFA scores result in better retention of single words in productive form recall?*

The use of effective framework to evaluate research materials before the experiment will lead to accurate interpretation. This helps to avoid misleading results and pitfalls that may threaten the internal or external validity of the study. The checklist of Technique Feature Analysis (TFA) framework, proposed by Nation and Webb (2011) can help to evaluate vocabulary materials as it can provide effective predictive power to the designed materials as evidence shown in Section 2.3.5. However, it was not clear from the previous studies that the framework can support form recall knowledge in controlled written production. The insights found from this study contribute to theoretical aspects towards the effectiveness of the framework on form recall. According to the test results, the Control group got significant lower scores (sensitive and strict) on both Immediate Posttest and Delayed Posttest than the experimental groups with high TFA support. These findings are consistent with those from a wide range of previous empirical research (e.g., Hu & Nassaji, 2016; Kamali et al., 2020; Khoshsima & Eskandari, 2017; Chaharlang & Farvardin, 2018; Gohar et al., 2018; Zou & Xie, 2018) in that the framework has positive effects on vocabulary gains. The application of this framework will assist language teachers and researchers to select appropriate vocabulary activities to ensure short-term development of form recall knowledge. Regarding its positive predictive power, the framework including eighteen criteria (questions) could help to estimate the validity of the target activities. The investigation into the differences between the Control group and experimental groups helps to confirm the validity of the overall TFA framework. I will discuss details regarding the effects of TFA on short-term and long-term retention of word knowledge, respectively in this section to address the first research question.

This study found that Thai EFL learners in the experimental groups performed better than the learners in the Control group in the Immediate Posttest in term of form recall knowledge. The finding is consistent with previous empirical studies in which participants' knowledge of meaning recall (e.g., Hu & Nassaji, 2016; Chaharlang & Farvardin, 2018; Gohar et al., 2018) and form recall (e.g., Khoshsima & Eskandari, 2017; Zou & Xie, 2018; Kamali et al., 2020) improved after learning with high TFA-support tasks. These consistent findings yielded the significance of TFA framework towards short-term retention of vocabulary knowledge at the recall level in the field of vocabulary.

Similar to the results of short-term retention, all TFA-assisted groups showed significant improvement from Pre-test to two-week Delayed Posttest, meaning that TFA had positive effects on their long-term retention of form recall knowledge. The results are also in line with previous studies using one-week (i.e., Gohar et al., 2018) and two-week Delayed Posttest (i.e., Kamali et al., 2020) in that the groups with high TFA support outperformed the Control group with no TFA support. This finding indicates that the conceptual framework could facilitate learners' ability to retain knowledge of meaning recall for at least one week (Gohar et al., 2018) and of form recall for two weeks (Kamali et al., 2020) after the treatments.

Taken together, the findings from the current study and previous studies have indicated that L2 learners are able to retain single-word form for a short term as well as for a long term after learning with high TFA-support materials. A possible reason for this effectiveness is that TFA involves many components, such as *Noticing, Retrieval* and *Generative Use,* associated with mental process. High quality of explicit vocabulary learning is linked with three learning conditions: noticing, spaced retrieval, and generative use (Nation, 2022).

In terms of *Noticing*, learning happens through this condition because *Noticing* seems to involve memory process, and higher involvement of memory process is likely to result in greater retention. The findings showed positive effect of *Noticing* on both short-term and long-term retention of form recall. The results from *Noticing* (Group 3) are in line with Schmidt' s (1990, 1994,

2001) Noticing Hypothesis in that input (information) can be converted into intake stored in memory with a certain degree of noticing (Schmidt, 1994). In classroom, the *Noticing* group was informed that the list of words was the target vocabulary of the unit before the words were deliberately learned through a wordcards activity. Both form and L1 definition of each word were presented explicitly to them. According to the evidence, it can be assumed that one of the important elements in *Noticing* is awareness of learning. Awareness is required in the primary stages of cognitive processing (Anderson, 1982; Schmidt, 1990). If the input (new word form) is noticed with a considerable degree of awareness in the early stage of acquisition, it seems to be developed and stored in working memory effectively. This is in line with Folwe' s (2002) study in that awareness of learning leads to the growth of vocabulary knowledge. Awareness tends to be connected with the realisation of the importance or aim of learning, leading to higher degree of focus while noticing the target words. According to Nation and Webb (2011), attention and awareness are key elements of *Noticing*. The TFA material used for encouraging *Noticing* had high degree of these two elements, leading to extensive support on vocabulary learning. The findings from the Form Recall test confirm that a certain degree of attention to the target words tends to help learners in acquiring word form and linguistic features. This is also relevant to the findings of Hu and Nassaji' s (2016, p. 36) study in that the target words learnt by using a 'form-focused task' could be remembered greater than the words in a task that did not focus on form. When a target word is focused on with a certain degree of attention, it is likely to be retained in memory at least for two weeks as supported by the results of the precent study. Although the significance of noticing, which includes awareness and attention, on vocabulary gains has been argued by many eye-tracking studies that support incidental learning (e.g., William & Morris, 2004; Pellicer-Sánchez, 2016; 2020), it would seem logical to assume that noticing play a crucial role in vocabulary growth and intentional learning. That is because certain degree of noticing was found in the current study to facilitate both short-term and long-term retention.

In terms of *Retrieval*, the effects of spaced retrieval tends to be one of the factors leading to learning and retention of word form. The findings showed that materials designed to include spaced and multiple retrievals facilitated vocabulary learning and retention, especially long-term retention of form recall. Although the learning was within one session of a three-hour class, rather than massed together each vocabulary learning activity was spaced (for 30 minutes) before moving on to the next one. Spaced retrieval is likely to encourage learners to try harder to memorise new information from the first meet to the next practice. The process of working memory may involve in the practice, helping to consolidate the information to be stored longer in the memory (see more discussion in Section 6.1.2). Also, form recall is demanding in that the learners need to retrieve the words from memory without a provision of choices to see or hear. Therefore, the deeper level of memory process during *Retrieval* practice, resulting from the spacing, would lead to better long-term retention. The increase of the Delayed Posttest score of the *Retrieval* group led to the highest score among the six groups in the experiment, resulting in the significant differences from the *Motivation* and *Generative Use* groups. The findings showed that *Retrieval* is likely to require deeper memory process so that the memory trace lasted longer in the experiment. This can be explained by the Level of Processing Theory (Craik & Lockhart, 1972). Craik and Lockhart divided processing models into three forms: Structural Processing, Phonemic Processing, and Semantic Processing. The Structural Processing and Phonemic Processing are embedded in *Shallow Processing*, the process in which information (i.e., appearance and sound) can be rehearsed and hold for a short time. The knowledge of word form of the *Retrieval* group might be developed deeper into another form which is called Semantic Processing of the memory process. Semantic Processing, which is considered as *Deep Processing*, does not only require knowing the meaning of a word, but also linking the meaning to similar words with similar meaning. This can also validate the results of the meaning recall test of the current study in that the learners could retain the form of a target word longer because they understand its underlying meaning. However, this developmental process of vocabulary knowledge was not only found from the *Retrieval* group, but also from the *Noticing* group. The two-week Delayed Posttest scores of these two groups (*Noticing* and *Retrieval*) raised from the Immediate Posttest scores while the scores of the other groups dropped in the final test. The decrease of the final test scores is commonly found in the field of memory and psycholinguistics. Researchers (e.g., Field, 2004; Mayer, 2014; Schmitt, 2010) found that memory decay and memory lost are associated with the amount of retrieval experience of learners. When the target items are not frequently retrieved, the memory of these items tends to decay. During the two weeks before the Delayed Posttest, the learners did not have exposure to the target words in class, aiming to avoid additional learning. This period of non-exposure could result in the decay of productive vocabulary knowledge even though no significantly difference was found between the Delayed Posttest scores (long-term retention) and Immediate Posttest scores (short-term retention). The increase in the Delayed Posttest scores of the *Noticing* and *Retrieval* groups, however, is likely to be connected with the process of working memory which will also be discussed further in Section 6.1.2.

In terms of *Generative Use*, learning may happen through elaboration rehearsal by thinking and linking form with meaning in a meaningful way. *Generative Use* tends to involve in mental process because elaboration rehearsal occurs in the Deep Processing of working memory (Craik & Lockhart, 1972). One possible reason that the Delayed Posttest of the *Generative Use* group decreased, which makes it significantly lower than the *Retrieval* group, might be due to the insufficient degree of awareness in vocabulary learning which is considered as the primary element of memory retention. The *Retrieval* group was forced to remember the words in sentences appeared quickly (2 seconds, each sentence) on a computer screen in order to do the Part of Speech activity (see also Appendix 2: Lesson Plan for the *Retrieval* group). This can raise higher degree of awareness of learning to the learners even though degree of noticing was unlikely to be high enough for the learners to notice the target words as supported by the questionnaire findings. Also, the *Retrieval* activity tended to have clearer instantiations than the *Generative Use* activity, meaning that the *Generative Use* group may involve in lesser memory processing than the *Retrieval* in the experiment.

In conclusion, *Motivation*, *Noticing, Retrieval* and *Generative Use* in the TFA framework play a significant role in vocabulary learning and retention. Although *Motivation* is likely to have lower effects than other TFA components, the current evidence confirms its effectiveness on form recall knowledge. *Noticing, Retrieval* and *Generative Use* tend to involve in mental process as explained earlier so their effectiveness seems to be higher than *Motivation* to some extent (see also discussion in Section 6.1.2). The materials designed to support *Noticing* and *Retrieval* in the current study do not only include the support on word forms (visual processing), but also the association between form and meaning of the words (sematic processing) through explicit and implicit learning, respectively. The two components are likely to be connected together at an adequate level at the *Shallow Processing* while *Generative Use* might also need higher degree of noticing at the *Shallow Processing* to stimulate deeper memory process. This may be the reason why the Delayed Posttest score of the *Generative Use* group dropped in the final test although it tends to involve in memory process. The evidence tends to emphasise the importance of noticing at the primary stage of working memory. Besides, this evidence seems to be relevant to a memory model of Waugh and Norman (1965), Atkinson and Shiffrin (1968) and Kihlstrom (1984), which proposes that input information (a target word) requires certain degrees of attention to be retained in short-term store and later in long-term store if the information involves the encoding and retrieval process for several times (see also Figure 2.2 in Section 2.3.3.2). So, high levels of support for *Noticing, Retrieval* and *Generative Use* together in the TFA framework are likely to lead to learning and longer retention of word form.

As the overall predictive power of the TFA has been consistently proven to be valid, this framework is likely to be a useful tool to evaluate vocabulary materials. However, expanding on previous research, the present study also found some mismatch between the predictive power of the TFA and the Form Recall test scores. These findings indicated that there is a need for further improvements of this framework. Section 6.1.2 below discusses these issues in detail.

### 6.1.2. The effects of TFA components on controlled productive vocabulary knowledge

***Research Question 2:*** *What is the individual role of the different TFA components on promoting vocabulary retention through vocabulary tasks?*

In addition to finding that all four components of the TFA framework contribute to vocabulary learning at the form recall level, this study also provides further insights into the relative effects of each component in vocabulary retention. In this section, I will discuss the effects of the TFA components on form recall knowledge, and explain the mismatch results between the TFA scores and test scores before I delve into the suggestions for further revision of the framework.

To begin with, this study found that *Motivation* always had the lowest power in both Posttests. Among the four conditions, there was only the *Generative Use* that showed similar power to *Motivation* in the Delayed Posttest. The findings on the effects of *Motivation* and *Noticing* are surprising because the activity for the *Motivation* Group (Group 2) had a similar number of TFA criteria (3) and score (7 out of 18 points) as the activity for the *Noticing* Group (Group 3), which turned out resulted in greater retention. The lowest power of the *Motivation* group among the four experimental group in the Immediate Posttest could be because Nation and Webb's definition of motivation in the TFA is different from that in motivation theories. In the TFA framework, *Motivation* tends to be defined as a mechanism that can stimulate pleasure and challenge in learning. This definition is narrower than those from motivation theories. As presented in Section 2.3.3.1, according to motivation theories, motivation can be broadly divided into intrinsic motivation and extrinsic motivation. Intrinsic motivation means the drive within the learners, triggered by internal factors or inherent satisfaction while extrinsic motivation means a motivation driven by outside factors such as scores, rewards and incentives. Both types of motivation should be included in the TFA framework to drive successful vocabulary learning (see more discussion in Section 6.1.3). Additionally, motivation towards vocabulary materials might be different from motivation towards general materials in terms of quality of attention to vocabulary learning goal. The first concept may require certain degree of consciousness in learning. This may result in the low impact of *Motivation* since the target words were implicitly introduced to this group to avoid high degree of *Noticing*. It means that although the learners know that vocabulary was the learning goal, the activity did not raise awareness of learning a new list of the target words, not to mention the attention on the target words themselves. To increase degree of *Motivation,* the goal in learning may need to be emphasised more explicitly. If the learners realise the importance of the learning, it may drive their intrinsic and extrinsic motivation to learn. As the lower learning gains of the *Motivation* group compared to other experimental groups could be due to the narrow definition of motivation in the TFA, the TFA criteria under *Motivation* component may probably need further development so that the TFA framework could better evaluate vocabulary learning activities (see Section 6.1.3 for further suggestions related to this point).

Of the three remaining conditions (*Noticing, Retrieval,* and *Generative Use*), regardless of the scoring scheme and kind of retention, there was no significant difference between the retention of the *Noticing* group and the other two groups. One reason to this evidence is that these three components are likely to have similar aspects of cognition. Functions of cognition involve attention and memory processing (Tapia & Duñabeitia, 2021). *Noticing*, *Retrieval* and *Generative Use* require attention to the new information and involve complex operations in working memory while learning that information. The results also highlight the significance of noticing, spaced retrieval and generative use suggested by Nation (2001; 2022) as mentioned in Section 6.1.1 in that the three components involve complex memory process that can contribute to deeper understanding. This argument and findings of the current study are also in line with Webb and Nation's (2017) notion. They stated that *Retrieval* and *Generative Use* tends to involve *Noticing* at the primary stage of learning. Thus, the current study confirm that *Noticing, Retrieval*, and *Generative Use* involve mental process, and they can strengthen memory retention to the same extent despite of their different TFA scores. However, when comparing the test scores of the three components with their TFA scores, the mismatches were also found. The *Noticing* Group (Group 3) had similar power to the *Retrieval* (Group 4) and *Generative Use* (Group 5) groups as the evidence from both Posttests while they received different TFA evaluated scores (*Noticing* = 7; *Retrieval* = 8; *Generative Use* = 9 out of 18 points). This shed light on the need for further improvement of the TFA analysis criteria (see discussion in Section 6.1.3).

When *Retrieval* and *Generative Use* were directly compared, the findings showed that *Retrieval* (Group 4) and *Generative Use* (Group 5) could benefit short-term retention to the same degree as evidence by the Immediate Posttest scores. However, *Retrieval* scores on the Delayed Posttest (long-term retention) were significantly higher than those of the *Generative Use* although its TFA evaluated score was lower (TFA scores: *Retrieval* = 8; *Generative Use* = 9). This finding is in line with many previous studies (e.g., Roediger & Butler, 2011; Rowland, 2014) in that retrieval brings more benefit to Delayed Posttest than Immediate Posttest. Recent research (Roelle & Nückles, 2019) comparing Generative Learning and Retrieval Practice also yielded a similar result to the current study in that both components contribute to knowledge gains, but Retrieval Practice greatly benefits long-term retention when the learners have formed substantial mental representations that are linked with their prior knowledge. This empirical research is connected with the current results in that *Retrieval* tended to have superior power than *Generative Use* in long-term retention. The fact that the *Retrieval* group could better facilitate long-term vocabulary retention than the *Generative Use* group even though it received lower TFA score may be because the *Retrieval* activity tends to involve higher attention and deeper memory process than the activity for *Generative Use* as mentioned in Section 6.1.1*.* To clarify, new knowledge can be forgotten if there is the lack of individual need in producing words after finishing the task in classroom and/or when it was not deeply processed in memory for multiple times. The *Generative Use* group only produced the learned-words when they need to involve in the assigned activity in class so that it had less impact on long-term retention than *Retrieval*. Being forced to remember sentences including the target and extra words on the computer screen may trigger personal need of the learners in the *Retrieval* group to pay more attention to learning, resulting in deeper process in memory (see also the Lesson Plan in Appendix 2). Despite of its lower effect on long-term retention, *Generative Use* was found to be as one of the crucial factors affecting vocabulary gains and retention according to its positive outcomes on both posttests. The learners in the *Generative Use* group had been exposed to the target words in the reading passage unconsciously before doing the suffixes with sentence writing exercise that promoted receptive and productive generative use. The findings supports Generative Learning Theory, originally proposed by Bartlett (1932), in that learning is constructed by linking new knowledge or experience to the existing knowledge schemata in a meaningful way in memory. This means that learning and memory are constructive. The association between external (new elements of knowledge to-be-learned) and internal (existing knowledge) connections leads to learning and later fosters retention in memory.

In sum, although all TFA components are likely to have positive effects on form recall, the current study found that the same TFA component may develop form recall knowledge to different degrees at different time. Retrieval may need longer time for vocabulary development than other components. The evidence showed that higher TFA score does not necessarily indicate greater level of knowledge gains and retention. Although the TFA framework has been regarded as an effective conceptual framework for vocabulary knowledge for decades, the mismatches between *Motivation* and *Noticing, Noticing* and *Retrieval,* and *Retrieval* and *Generative Use* scores indicate that there is a need for further development of this framework. The following section provides some possible recommendations for further development of the criteria within the framework based on the findings of this study.

### 6.1.3. The rationales and suggestions for further development of the framework

The above evidence shows the significance of TFA framework and its component towards vocabulary learning; however, the findings also shed some light on possible ways in developing the framework in the future. The five TFA components: *Motivation, Noticing, Retrieval, Generative Use,* and *Retention* were originally guided by prior studies (e.g., Baddeley, 1990; Joe, 1998) on L2 vocabulary acquisition. Nation and Webb (2011) explained the significance of each learning factor included in the framework based on the results of previous studies. After the framework was first introduced, Webb (2013) and Webb and Nation (2017) elaborate more about conditions contributing to vocabulary knowledge gains to support the importance of TFA components. However, their reasons for including various criteria within each component of the framework were not explicitly stated. As reflected by the raters in the present study, some terms used in the framework tended to be difficult to understand. Consequently, different raters provided different TFA scores for the same activity at the first time of the evaluation. The lack of providing a proper training with sufficient explanation of terms in the framework to the raters was found to be one of the obstacles that delayed the evaluation process in the present study. This indicates that it may be difficult to implement the TFA as a tool to evaluate vocabulary learning in real classrooms because language teachers, even experienced teachers, have difficulties in understanding specific terms in this framework. To tackle this problem, further development of the TFA framework is needed. The framework can be (1) refined by giving more explicit information about different extents of *Motivation, Noticing, Retrieval,* and *Generative Use* on different aspects of word knowledge (i.e., meaning recognition and recall, and form recognition and recall) and (2) developed further by adding more criteria to some components such as *Motivation* and *Generative Use* (see Table 6.1).

I propose feasible ways to develop the framework with some additional criteria to *Motivation*, *Noticing, Retrieval* and *Generative Use* as shown in Table 6.1. Each component might be able to provide similar predictive power to each other on word knowledge gains if more criteria are added to some components such as *Motivation*, *Noticing* and *Generative Use*. The suggested additional criteria are presented in *Italic* while those that need more elaboration and/or are suggested to be removed were marked by a star (*). It should be noted that I only aimed to present possible ways in increasing the predictive power of *Motivation*, *Noticing, Retrieval* and *Generative Use* in the TFA framework based on the empirical results of the current study. The details below are driven mainly from the findings of the current study.

***Motivation***

Raters reflected difficulty in understanding the definition of motivation in the TFA framework. In the framework, *Motivation* refers to goal and challenge of learning. However, the TFA definition might be relatively narrow and cannot capture the complexity of learners' motivation. According to the motivation theories, motivation can be classified into two kinds: intrinsic motivation and extrinsic motivation (see Section 2.3.3.1 for details). Again, intrinsic motivation means self-motive that is driven from personal need to accomplish one task while extrinsic motivation means stimulation such as scores or rewards that trigger the need of a person to achieve the goal.

***Table 6.1. Suggested criteria for further development of Motivation, Noticing, Retrieval and Generative Use components of TFA Framework***

| Component | Criteria | scores | |
|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | 0 | 1 |
| *(5 points)* | Does the activity motivate learning?* | 0 | 1 |
| | Do the learners select the words? | 0 | 1 |
| | *Does the activity encourages intrinsic motivation?* | 0 | 1 |
| | *Does the activity encourages extrinsic motivation?* | 0 | 1 |
| ***Noticing***<br>*(5 points)* | Does the activity focus attention on the target words? | 0 | 1 |
| | *Is it form-focused?* | 0 | 1 |
| | *Is it meaning-focused?* | 0 | 1 |
| | Does the activity raise awareness of new vocabulary learning? | 0 | 1 |
| | Does the activity involve negotiation? | 0 | 1 |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 0 | 1 |
| *(5 points)* | Is it productive retrieval? | 0 | 1 |
| | Is it recall?* | 0 | 1 |
| | Are there multiple retrievals of each word? | 0 | 1 |
| | Is there spacing between retrieval? | 0 | 1 |
| ***Generative Use*** | Does the activity involve generative use? | 0 | 1 |
| | Is it productive? | 0 | 1 |
| *(5 points)* | ***Is it task-induced generation?*** | 0 | 1 |
| | Is there a marked change that involves the use of other words? | 0 | 1 |
| | *Does the activity involve instantiation?* | 0 | 1 |

Note: *Italic* questions are additional criteria; questions with a star (*) are those that need more elaboration and/or are suggested to be removed; **bold** and *italic* question under the *Generative Use* component was suggested by Kamali, et al.' s (2020) study

Previous empirical studies on motivation (e.g., Bailey et al. 1999; Amiri & Salehi, 2017) found that the achievement in developing word knowledge linked with extrinsic motivation such as effective materials. Similarly, the connection between intrinsic motivation and learning achievement has been affirmed by a wide range of prior research on motivation (Noels et al., 2000; 2001; Pae, 2008) and vocabulary learning (Zhang et al., 2017). To increase its predictive power, I propose two possible criteria to the *Motivation* component: *Does the activity encourages extrinsic motivation?* and *Does the activity encourages intrinsic motivation?* Although the framework mainly aims to facilitate materials (extrinsic motivation) in vocabulary learning, sequential outcomes of using effective materials may also lead to intrinsic motivation. In addition to this, the terms *extrinsic motivation and intrinsic motivation* should also be defined explicitly in the light of motivation theories. Also, the second criterion in the component: *Does the activity motivate learning?* can be identified within the two suggested criteria so it is suggested to be removed or modified because the term *motivate* tends to be too broad. Some words related to motivation such as *enjoyment* or *challenging* suggested by scholars (i.e., Barry & King, 2000) in the field of motivation should be used to explicitly clarify the criterion.

### *Noticing*

Noticing is regarded as a primary stage of learning new words as claimed by Schmidt (2001). Some L2 studies (e.g., Leow 2001; Philp, 2003) argued against Schmidt' s (1990; 2001) Noticing Hypothesis to highlight a limit on the effects of noticing towards language learning and acquisition. Also, eye-tracking research on word gains (e.g., Pellicer-Sánchez, 2016; 2020) raise awareness of the significance of incidental learning over intentional learning involving elements of noticing (awareness, attention, and consciousness). However, results from the recent empirical study revealed the significance of noticing on long-term word retention. While it is challenging to thoroughly measure noticing, results from the *Noticing* group indicate that attention to the target words with awareness of learning new words is likely to enhance word knowledge and retention of word form for at least two weeks. This emphasises the need of *Noticing* component in the vocabulary framework. Furthermore, results from the current study found that a provision of L1 definitions or L2 synonyms to the target words (word form) can lead to high noticing. Word knowledge involves both knowledge of form and meaning (Nation, 2001; 2013).

According to VanPantten (2020),

> *"…[A] good deal of acquisition is dependent upon learners making appropriate form-meaning connections during the act of comprehension. It is the raw data in the input that learners need to construct a linguistic system."* (p. 105)

This notion is in line with the results of the current project in that with the presence of form and meaning in the materials the *Noticing* group got considerably high score in both Form-Meaning Recall posttests although it received the lowest TFA score among the experimental groups. By adding more questions, its predictive power might be improved as similar to the suggestions regarding the additional plus to the 'evaluation' domain of the ILH given by previous research (e.g., Hazrat, 2020; Yanaginawa & Webb, 2021). One suggested criterion to this component is the presence of only form/meaning or both form and meaning in the materials for gaining word knowledge. This can be useful especially when using the framework to evaluate retention of word knowledge. Following the first existing question *"Does the activity focus attention on the target words?"*, I propose these possible questions: *Is it form-focused?* and *Is it meaning-focused?* Although it is similar to the first criterion under the *Retention* component concerning the link between form and meaning, noticing of form and meaning does not necessarily lead to the successful link between the two elements. Since the results demonstrated significantly high effect of *Noticing* on word recall knowledge, adding only one additional criterion to the component should be sufficient and would not change much of its predictive power, but will show consistency in terms of number of criteria compared to such components as *Retrieval* (with five criteria).

### *Retrieval*

Retrieval is a memory process which requires two prior stages: *encoding* and *storage* (Melton, 1963). Information needs to be encoded and stored in memory to be able to retrieve. With regard to language sustainability, an abundance of research agreed that retrieval strengthens long-term retention (e.g., Roediger & Butler, 2011; Grimaldi & Karpicke, 2014; Rowland, 2014) and knowledge of words and sentences (e.g., Tran et al., 2015; Eglington & Kang, 2018; Hulme et al., 2019). Consolidation and transfer of word items from working (short-term) memory to long-term memory can be due to rehearsal and multiple retrievals. Self-study through using a dictionary after the learning, for example, can benefit retrieval of words. Another influential factor to (recall) retrieval is *visualisation* as explored by previous empirical studies (e.g., Nairne, 2002; Watkins, 1975). The current study found the evidence of retrieval process that involves *visualisation* of word form in the *Retrieval* group. The participant could visualise the form of target words and spell the word correctly. This is in line with the idea about visuo-spatial sketchpad, regarded as a memory system for visual information in working memory. This memory system is the primary process in the short-term store (working memory) that can transfer and retain visuals from short-term store to long-term store if they are associated with any existing information in the brain and so the stored visuals can be retained and revisited over periods of time (Baddeley et al., 2009). Even though *imagination* was added as a criterion under the *Retention* component, it is a different concept to *visualisation*. Regarding to this, the term *visualisation or visual* should be mentioned in any existing criteria within the *Retrieval* component. It is not suggested to add an additional criterion to this component as there are too many questions included in it already, which can affect its predictive power. One possible solution is to integrate the term with the third criterion: *Is it recall?* The underlying concept of *visual recall* on word retention should be explicitly clarified. Prior studies argued that *visual recall* tends to be superior than *verbal recall* (Kosslyn, 1980; Shepard & Cooper, 1982; Hall et al., 2018). If the material involves *visual recall*, a score of one should be added to this question. Besides, the results showed positive effect of *Delayed Retrieval* as Delayed Posttest scores of the *Retrieval* group raised after two weeks of the experiment. The concept about *Delayed Retrieval* is also suggested to be clearly explained along with the existing criterion: *Is there spacing between retrievals?* under the component. The raters also commented on the difficulties in defining terms presented in the *Retrieval* criteria. It is necessary to clearly label each question to clarify the terms included in this component.

### *Generative Use*

The concerns on *Generative Use* component are due to the results obtained from the Delayed Posttest of this study. While the TFA score showed that *Generative Use* (TFA = 9) tends to have higher predictive power on word knowledge than *Motivation* (TFA =7), the two components were not significantly different in the long-term retention results. This means that *Generative Use* appears to facilitate short-term retention better than long-tern retention of single-word form. In recognition to this, future studies on TFA could probably consider the purpose in evaluating *Generative Use* materials whether to facilitate either short-term retention or long-term retention. Taken a consistency of question number into account, adding a relevant question to this component can be one of the possible ways in developing the framework. One potential criterion can be related to the concept of *instantiation*. When linking the findings to the memory model of Waugh and Norman (1965), Atkinson and Shiffrin (1968), and Kihlstrom (1984) presented in Section 2.3.3.2, the rehearsal process in working memory involves *elaborative rehearsal* that is driven by concepts and schema-based processing (Craik & Lockhart, 1972; Hulstijn, 2001). This means that new knowledge can be driven to maintain in long-term store by experiencing examples of the same words recuring in various contexts, and so *Generative Use* is likely to be involved in a working memory process. While *instantiation* is one criterion included in the *Retention* component, it is suggested that this question: *Does the activity involve instantiation?* should rather be relocated under the *Generative Use* component. Also, Kamali and his colleagues (2020) found the evidence that supports the need to increase the number of criterion in this component. They suggested the need in adding *task-induced generation* as a criterion under *Generative Use*. Adding these two criteria to the component can advance its predictive power in assessing word knowledge.

Although I did not focus on *Retention*, there are some suggestions for further development according to the results of the current study.

### *Retention*

As explained in Section 3.4 and Section 4.2.2, retention involves memory process. Some findings emerged from the experiment suggest that the criteria related to retention in the TFA should be revisited in future studies. The findings support that *Noticing, Retrieval*, and *Generative Use* lead to retention of word knowledge, meaning that retention seems to be embedded within other TFA components. So, I propose either to remove *Retention* component from the TFA framework or alter the labelled name of this component to be more relevant to learning factors affecting vocabulary development and word retention. The ideas such as *personalised strategy* or *cognitive strategy* (i.e., *dictionary use*) that occurs while/after using the materials should be rather concerned. The idea is similar to *search,* a component in the Involvement Load Hypothesis (ILH) discussed in Section 2.3.1 and Section 2.3.2, so the inclusion of *search* into the framework may enlarge the effective predictive power of the TFA. As proposed by Schmitt (1997), *cognitive* (i.e., guessing from context, dictionary use) and *metacognitive* (i.e., self-monitoring, selective attention) strategies are considered to be essential in vocabulary learning. Prior vocabulary studies (e.g., Barcroft, 2009; Gu, 2003; Gu & Johnson, 1996; Zhang et al., 2017) has confirmed the association between learning achievement and learning strategies (i.e., *cognitive* and *metacognitive*). Gu (2019) also stated that “For learners of English as a Second (ESL) or foreign language (EFL),…how to treat the learning of each of [English] words are very much strategic tasks” (p. 271). A component regarding vocabulary learning strategies may help with the materials selection process because different types of vocabulary tasks may require either specific learning strategies or a variety of strategies in combination to match the task types and purposes of vocabulary learning.

Apart from the TFA learning components, the current research also found possible factors that seem to be involved in assuring the retention of word recall knowledge. These factors are discussed more in the following section.

### 6.1.4. Other factors affecting retention of form recall

This section describes other factors apart from the TFA components that may possibly affect the learning of form recall in controlled written production. There are two major factors, including type of learning (incidental learning vs. intentional or deliberate learning) and working memory that were also observed to be a trivial influence of form recall knowledge. These factors are discussed based on the findings of the current study.

To begin with, one possible factor is related to how the students learn word knowledge in class. There are two well-known approaches of learning: incidental learning and intentional learning that link with memory retention (Hulstijn, 2001; 2003). The current results confirmed that the *Noticing* group who learned the target words intentionally had the highest test scores, though not significantly different, compared to the five experimental groups in the Immediate Posttest. The *Noticing*, however, could greatly retain knowledge of form longer than the *Generative Use* participants who learned the target word implicitly. While the impacts of incidental learning and intentional learning have been a controversial issue, the results of the present study confirmed positive effects of both learning methods in retaining knowledge of word form. Though, I argue that intentional/deliberate learning is seen to have superior power than incidental learning in long-term retention of form recall giving to the results of the Delayed Posttest. The use of more than one test at one point in time yielded richer data gained from both learning approaches. Previous research on incidental learning (e.g., Day et al., 1991; Cho & Krashen, 1994) is likely to employ only single test to assess word knowledge as criticised by Webb and Nation (2011), and most studies (Paribakht & Wesche, 1999; Webb, 2008; Ponniah, 2011) focused only on the impact of reading task on word knowledge rather than multiple activities. Hence, the present empirical study raises awareness on the significance of intentional learning towards multiple activities that encourage both short-term and long-term retention of form recall.

Additionally, memory process seems to be another factor leading to the potential in retaining word form. The *Noticing, Retrieval*, and *Generative Use* that seem to involve memory processing showed significant higher effect than *Motivation* with no or very little involvement of memory process while learning the target words. The *Noticing* group intentionally learned the target words with the provided definitions (*form-meaning link*), so the link can be connected to their language repertoire in memory. Likewise, *visual recall* in the *Retrieval* component and *productive generative use* in the *Generative Use* component involves memory processing in the brain. With regard to working memory, visuals can be processed in the primary memory system called visuo-spatial sketchpad to be stored over periods of time; however, long-term store requires the association between new information (visual) and existing knowledge in the brain otherwise it will be lost or forgotten (Baddeley et al., 2009). This could be the reason why these three groups got significantly higher scores than the *Motivation* in the Immediate Posttest. If the *Generative Use* component includes more suggested criteria to increase its power on memory process, the three materials encouraging *Noticing, Retrieval*, and *Generative Use* might also give better results than *Motivation* on long-term retention. Therefore, the issues on memory process should be taken into consideration when form recall knowledge in written production is the focus of learning.

### 6.1.5. Implications

As discussed earlier, all TFA components: *Motivation, Noticing, Retrieval* and *Generative Use* can facilitate word learning and retention. However, the impact of each component on retention of form recall showed variety in degrees of support, resulting in the awareness in terms of implications. The components that involve memory process are likely to be *Noticing*, *Retrieval,* and *Generative Use*. Among these, *Noticing* and *Retrieval* contribute to the long-term retention to the same extent, but greater than *Motivation* and *Generative Use*. Long-term memory retention promotes sustainable knowledge of L2 learners. Researchers should take memory process into consideration when aiming to explore or facilitate long-term retention. Sustainable productive knowledge can be facilitated by the use of TFA framework as suggested by the results of the recent project. This will help language teachers to understand how to apply the knowledge of the framework to design instructional materials for a sustainable development of word knowledge. While the framework can be useful as a tool to evaluate vocabulary materials, the evaluation should be done very carefully by first concerning the purpose of learning. If long-term knowledge gain is the main learning objective, *Noticing, Retrieval,* and *Generative Use* that involve substantial memory process should be more focused when evaluating materials. However, an integration of the four TFA components: *Motivation, Noticing, Retrieval,* and *Generative Use* is advisable to enhance working memory (short-term) retention which can eventually strengthen long-term retention of word knowledge.

## 6.2. Pedagogical contribution and implications

The current study also have some pedagogical contribution in terms of instructional material evaluation and development. The positive effects of TFA activities on learning found in this study indicates that until a revised framework is created and validated, the existing TFA is still a useful tool to evaluate materials for promoting EFL learners' word recall knowledge. However, the inconsistence in the predictive power of each TFA component found in the present study also indicates that caution should be taken when implementing the TFA in real classrooms. First, teachers should be aware that this framework is not perfect. It should be seen as a starting point for them to systematically evaluate vocabulary learning activities and other factors such as personalized learning strategies and memory process which are not covered by the framework should also be taken into account. Second, the confusion experienced by the raters when initially rating the TFA scores of the activities indicates that language educators or researchers who are not familiar with these specific terms may find it difficult to use the framework for vocabulary evaluation. This problem could be solved by a training together with a supplementary sheet describing difficult terms (see also Table 2.4, more discussion in Section 6.3 and suggested user's manual is Appendix 10).

Furthermore, language teachers should pay more attention to materials for sustainable (long-term) word learning rather than for short-term retention of words-to-be-learned. The inclusion of *Noticing, Generative Use*, and *Retrieval* components that facilitate memory process to teaching materials for vocabulary can promote long-term word learning. Knowledge in working memory can decay over time if it is not stored in long-term memory. A memory model (see also Figure 2.2 in Section 2.3.3.2) by Waugh and Norman (1965), Atkinson and Shiffrin (1968), and Kihlstrom (1984) illustrated the relationship between attention (*Noticing*) and multiple retrievals (*Retrieval*). This association in working memory (or primary memory) is crucial for further development of knowledge to be stored in long-term memory (or secondary memory). This can be affirmed by the results of long-term retention in that *Retrieval* can facilitate word learning to the same extent as *Noticing*, but better than *Generative Use*. However, the current study found a similar effect of *Generative Use* compared to *Noticing* and *Retrieval* in short-term retention, meaning that *Generative Use* also supports working memory process (or short-term store). This is also in line with the memory model mentioned earlier in that generative rehearsal in working memory requires generative use of words appeared in different contexts before being able to maintain in long-term store. Since these factors assist working memory that can eventually contribute to long-term store, language teachers should create materials that facilitate *Noticing, Generative Use*, and *Retrieval* in combination to promote a sustainable learning. For example, if the activities for *Retrieval* and *Generative Use* are combined and used with the list of target words created for the *Noticing* activity (see Appendices 2 and 3), the material will be more effective, leading to better vocabulary learning and longer retention of form in learners' memory.

According to the pedagogical contribution, this study also provides implications for evaluation practice of the use of TFA framework. Suggestions are made for the purpose of applying the framework to evaluate word knowledge in the field of language education. Firstly, language educators or researchers should avoid bias from self-evaluation by recruiting at least two experienced raters to evaluate materials when using the framework. There are many specific terms in the framework that need more clarification otherwise different raters would use their different background knowledge to justify the terms as found from the first evaluation of this study. The term *Motivation*, for instance, was defined according to the definition from Dörnyei (1994), Sternberg and Williams (2002), and other previous studies on motivation such as Hamada and Kito (2008), Sakai and Kikuchi (2009) and Yoshimura (2017). They defined learning materials as a part of *extrinsic motivation*. Consequently, key terms were defined to match the current research purposes as explained in Section 2.3.5 (see also Table 2.7). In recognition of this problem, a proper training should also be operated prior to the evaluation. Using two- or three-raters system with a provision of an additional training together with a description of terms can help the evaluation process run successfully and so the comparison of results between the raters can ensure the final results of the evaluation (see User's Manual in Appendix 10). Therefore, it is also suggested that language educators or researchers should clarify the terms used in the framework to all raters so that they share the same concept of understanding. This can avoid misleading results of interpretation that may be arisen from bias or difficult concepts appeared in the TFA framework.

Furthermore, instructional materials aiming to promote long-term word learning should support *Noticing, Retrieval* and *Generative Use* as discussed earlier. Materials promoting sustainable productive knowledge should highly promote noticing and retrieval process since they seem to have greater effect than some vocabulary components such as *Motivation* and *Generative Use* in long-term retention of word, especially on form recall. However, knowledge of word form can be successfully transferred from short-term to long-term recall if fun and challenging (*Motivation*) materials involving multiple retrievals (*Retrieval*) and high generation (*Generative Use*) process are designed and promoted to meet adequate attention (*Noticing*) to the target words in learning. This means that a combination of the four TFA components: *Motivation, Noticing, Retrieval*, and *Generative Use* in productive vocabulary materials would provide a greater effect to word learning.

## 6.3. Methodological contribution and implications

This study has two main methodological contributions. The first contribution is the employment of different raters to evaluate the TFA scores. Previous research on the TFA (e.g., Kamali et al., 2020; Khoshsima & Eskandari, 2017; Chaharlang &Farvardin, 2018; Gohar et al., 2018; Zou & Xie, 2018) did not seem to realise that self-rating can cause bias and each TFA component may have different degrees of effectiveness. So, they tended to pay solely attention to the overall slef-evaluation result. This could lead to the variation of TFA score of the same activity. As evidence in the present study, in the first round of rating, different raters provided different scores for the same activities due to the lack of descriptions on some specific terms. By providing training to the raters, the present study offers an innovative solution to this problem. The effectiveness of this solution is supported by the fact that the final evaluation results tended to be consistent after the training. This process should be introduced to raters when using the TFA framework.

The second methodological contribution of this study can be the adaptation of an existing test for the purposes of research objectives and time efficiency.

Although existing vocabulary tests such as the new Computer Adaptive Test of Size and Strength, or CATSS (Aviad-Levitzky et al., 2019) have been proven to be valid and effective, a modification of these tests to match the research and evaluation purposes is crucial. In the current study, the new CATSS was adapted to evaluate the participants' prior knowledge on vocabulary (see Section 3.7.1, Section 4.4.1, and Section 5.2 for details). Before implementing the new CATSS, I found obstacles that the participants might face with when doing the online test through its website. As explained in Section 3.8, I realised that the new CATSS includes some aspects of word knowledge that the current study did not focus on. It also requires test takers to provide their personal information for the registration process which could take time to complete. This might result in loss of interest of the participants and violate the personal protection issue regarding the research ethics. In terms of practicality, the use of an online test in the context where the internet connection seems to be unstable is another problem. The inflexibility of the website that does not allow test takers to go back to select other modalities if they do not complete all levels was also found to be one of the concerns in using it through the online website. Although the test is likely to be useful, it was adapted to use offline to match the purpose of the current project. This issue could reflect the need for adapting the existing vocabulary tests to maximise the use of it, so the test results could help to explore the information of interest, avoid unexpected problems while doing the test, and save time during research.

The present study can provide four practical implications regarding to methodological contribution. These concern (1) the use of more than one TFA rater due to the difficulty of terms in the TFA framework, (2) the adaptation of an existing vocabulary test, (3) the implementation of two posttests (Immediate and Delayed), and (4) the use of two scoring schemes (sensitive and strict). I will give details of each issue in terms of implementations in this section.

First, the evaluation that includes at least two raters has been found to be more effective than one rater. To assess vocabulary materials using the TFA framework, for example, the current study found inconsistent results from different raters before the training was introduced to them. A provision of only lesson plans and sample materials for learning to raters was also found to be insufficient for a valid estimation of the materials' predictive power. A description handout to the terms used in the TFA framework along with a proper training of how to utilise the framework effectively for particular purposes will allow raters to give a corresponding result of material evaluation. Material developers or teachers who attempt to measure vocabulary activities effectively by using the TFA should provide a proper training as explained in Section 3.4.3 with a description handout of the terms in the framework (see also Table 2.7 in Section 2.3.5). For ease of use, they can also follow a step-by-step guide in the User's manual presented in Appendix 10 to apply the TFA framework. This can ensure the internal validity of the assessment results between raters.

The second methodological implication found from the current study concerns the adaptation of the new CATSS (Aviad-Levitzky et al., 2019). The adapted CATSS applied to use in both the Pilot Study and Main Study can be an effective tool to evaluate prior vocabulary knowledge of L2 learners. The test was modified from the new CATSS by removing some items that did not match the purpose of the current project. Some modalities and frequency levels were selected to assess the learners' knowledge on vocabulary. The attempt to modify the existing test helps to reduce the testing time and other negative effects that might occur while taking the test such as test takers' cognitive fatigue from paying long attention to the task (Grandjean, 1968; Hockey, 1983). So, it is advisable that users should realise the purpose of using a test to match the objective of their study even though a wide range of existing vocabulary tests such as the PVLT (Laufer & Nation, 1990), VKS (Paribakht & Wesche, 1993; 1997), and new CATSS (Aviad-Levitzky et al., 2019) have contributed to significance in vocabulary assessment.

Another point is that both Immediate and Delayed Posttests should be implemented not only to assess word gains but also to promote sustainable word learning. The basic principle of learning is to gain knowledge to use in daily life. If the learners forget the words they just learned in a short period of time, learning does not seem to be truly successful. Long-term word knowledge should be promoted for encouraging sustainable learning. The implementation of a Delayed Posttest generally helps to measure long-term retention; however, the delayed posttest should be carefully designed. A recent study by Johanna and her colleagues (2019) claimed that the indifference between the Immediate and fifteen-minute Delayed Posttest results may be due to the length of time between the two tests. While the current project yielded insights to the importance of the two-week Delayed Posttest, I found a similar result to Johanna and her colleagues (2019) in that there was no different in terms of amount of knowledge gain between the two tests. Therefore, it is advisable that the Delayed Posttest should be implemented after two weeks to a few months to measure long-term knowledge.

Lastly, I suggest using not only a strict scoring scheme but also sensitive scoring to measure incremental changes of word knowledge. While strict scoring is useful to evaluate knowledge gains, partial knowledge that can be measured by sensitive scoring should also be considered. Sensitive scoring gives an in-depth result of knowledge gains. This helps both teachers and researchers to ensure that learning happened. Partial knowledge obtained from the learning as a continuum can be further developed to the level of mastery. For this reason, sensitive scoring results can help teachers to monitor the learners' progress more precisely and accurately.

While the results yielded a clearer understanding to the TFA framework and word retention, I am aware that a single project could not be designed to fill all research gaps. In Section 6.5 below, I present information related to the limitations of the current study.

## 6.4. Limitations of the study

This research project investigated the effect of Technique Feature Analysis framework on retention of form recall knowledge. The two main purposes which contribute to the research originality comprise:

1) to explore the effect of TFA on retention of form recall knowledge in productive form recall
2) to find out the individual role of the different TFA components (*Motivation, Noting, Retrieval* and *Generative Use*) promoting vocabulary retention through vocabulary tasks

The study yielded significant contributions which were presented in Section 6.1, Section 6.2 and Section 6.3. However, several limitations should be acknowledged. The following limitations of the current study need further attention from future research. First, due to the time constraints of the Ph.D. and complexity in research design, this study only examined learning of single words. However, because knowledge of multiword sequences is also important for language learners, future studies could include the evaluation of muti-word units to gain more insights into the development of productive vocabulary knowledge and long-term retention of form recall as an influence of different vocabulary learning factors.

The second limitation is the usage of the Form-Meaning Recall test (F-MRt) in a repeated measurement research design study. The same F-MRt was used three times (as Pre-test, Immediate Posttest, and Delayed Posttest) for assessing word knowledge. This is a typical design of this line of research (e.g., Zhang & Graham, 2020; Baleghizadeh & Shafeie, 2017; Webb & Chang, 2015; Keating, 2008). Yet, this could be one of the limitations in terms of methodology. Although items in the test were shuffled every time before distributing it to the participants, there might be some effects (e.g., priming effect) of exposure to the same test for more than one time.

The third limitation concerns the exclusion of Meaning Recall test (MRt) scores for the main analysis. This study follows the dimensional view of word knowledge as mentioned in Section 2.1. Form and meaning are considered as a separate element. Although the F-MRt includes two parts of measurement on form and meaning of word form, the present study did not include the analysis of meaning recall as the main focus of the study was on form recall. It was used only for the purpose of checking accuracy of partial knowledge of word form to ensure that the participants have gained knowledge of the target words after the experiment. Also, it was excluded due to time constraints in doing the research. Therefore, the exclusion of MRt scores is considered as one of the limitations of this study as the provision of both form recall and meaning recall results would give a clearer picture of knowledge gains from the participants.

The fourth limitation is about spacing between each Form-Meaning Recall test. The space between the Immediate and Delayed Pottests was only two weeks. I had to distribute the Delayed Posttest in the third week of the data collection period because the experiment was conducted within three weeks to avoid negative interference to the actual classrooms that a shared course outline was used among fifteen to twenty sections in each term. A delayed posttest of more than two weeks or one month might yield more precise findings to the field of vocabulary and word retention.

The fifth limitation is related to number of the target words in the test. While the current project included ten target words similar to many previous studies mentioned in Section 2.3.5, having more or less items in the test may give different results. The effects of test items and primming effect are beyond the scope of investigation of the current study. Further empirical research should investigate the appropriate numbers of test items for measuring form and meaning knowledge as this will help to provide clearer evidence based on actual findings. Moreover, this study only aimed to explore form recall knowledge on controlled written production so that the results may not be applied to other aspects of productive knowledge such as knowledge in free writing.

The sixth limitation concerns context and proficiency level of leaners. This study was conducted with Thai EFL university students who are considered as intermediate learners of English, future studies could be done with learners in other contexts with similar or different levels of English proficiency. Conducting research in various contexts would give a clearer picture of how vocabulary is learned and retained longer in learners' memory. Learners with different language backgrounds are different in nature. They may require different learning factors to gain word knowledge and retain it in their memory.

The last limitation concerns the inclusion of only quantitative data for the main analysis. The use of post-treatment interview might provide further insights into the findings of the experiment. However, the present study removed the semi-structured interview data from the main analysis due to some limitations. First, the interview was initially conducted to control for *Motivation* that might also have significant effects on activities used with the other experimental groups. The questionnaire items were not primarily designed to answer the two research questions. Even though I could capture some information related to the research questions to support the test results, the qualitative data was carefully considered to be excluded from the main analysis. This was also according to the second limitation regarding the number of participants in the interview. There was only one participant from each group invited to participate in the interview process. This limited number of the participant could lead to misleading results. Importantly, this study aimed to conduct a quasi-experimental design research which focused on quantitative data sets. Due to the insufficient data collection and interpretation of the interview, this current study decided to rely only on the results of quantitative data analysis to address the two research questions. Therefore, the lack of insightful information from qualitative data analysis can be regarded as one limitation of this research project.

In the following chapter, I conclude information in the present study with regard to the two main research questions. Moreover, I will also reflect on my research journey and recommendations for further study in the final chapter.

# Chapter 7
# Conclusion

This study was conducted to address the lack of research on the impact of TFA framework on controlled productive vocabulary knowledge. In the previous chapter, I discussed the issues arisen from the results of the current project. In this chapter, I will summarise the overall information with regard to the two main research questions in Section 7.1 and reflect how I have overcome obstacles in doing this project in Section 7.2. Finally, in Section 7.3 I will give recommendations for further studies based on my findings and experience in conducting this research.

## 7.1 The predictive power of TFA and its components on retention of form recall knowledge

Vocabulary is a fundamental element of a language, and knowing words in L2 languages is essential for communication. Words can be learned successfully by the support of effective vocabulary materials. The Technique Feature Analysis, or TFA framework (Nation & Webb, 2011) has been recommended as a more reliable predictive tool compared to a prior well-known vocabulary framework, Involvement Load Hypothesis, or ILH, proposed by Laufer and Hulstijn (2001). However, earlier research did not pay attention to the effects of each TFA component in gaining controlled productive vocabulary knowledge. There is a lack of research on the impact of vocabulary learning components that can significantly gain knowledge of form and meaning recall on long-term retention. The present study has effectively filled in this gap by showing that the TFA components: *Motivation, Noticing, Retrieval*, and *Generative Use* can facilitate form and meaning recall in controlled written production. This means that the TFA not only facilitates research on receptive word knowledge as suggested by prior research, but can also effectively predict the overall controlled productive word knowledge in terms of meaning and form recall. Also, the TFA appeared to predict both short-term and long-term retention. It is likely to be a useful primary tool to validate components of memory when evaluating vocabulary learning materials.

The development of productive word knowledge involves different components of memory, this study also provide further insight into the relative power of each components on vocabulary retention. To begin with, *Motivation* showed lower effects on short-term retention of word form than *Noticing, Retrieval*, and *Generative Use.* This finding indicates that *Motivation* taps into different cognitive processes from *Noticing, Retrieval* and *Generative Use*, and that the last three components are central to designing materials to promote a sustainable word learning because they involve in memory process which supports long-term retention. The current study provides the insights into the effectiveness of the TFA framework and different extents of its components, which help to raise awareness of users when implementing the framework. Because of its positive effects on vocabulary learning, vocabulary materials can be developed by using the framework to maximise the learning in language classrooms.

Another important finding is related to the effects of each TFA component on form recall. Although we know that *Noticing, Generative Use* and *Retrieval* may be linked together in working memory process, it was unclear about the extent to which each component can facilitate controlled productive vocabulary learning. The current study helps to address this gap by finding that *Noticing* and *Retrieval* have the same degree of impact on long-term retention of word form. Vocabulary materials should rely heavily on these two learning components to promote sustainable learning. In addition, the materials should involve support for *Generative Use* for the purpose of rehearsal as generative rehearsal involve in deep processing of memory. However, certain degree of *Noticing* seems to be required for supporting *Generative Use* to strengthen long-term retention of word form. The development of vocabulary materials should not only put an emphasis on the factors affecting learning, but should also pay considerable attention to the material evaluation process. To avoid misleading results from material evaluation, a training with detailed supplementary handouts describing terms in the TFA seems necessary for making the same concept of understanding to all raters.

Taken together, the present study has provided further evidence supporting the validity of the TFA as a framework to evaluate vocabulary learning activities, and suggested the need for further development of the framework. The development could improve the validity of the framework, which will increase its effectiveness in measuring vocabulary activities and help to avoid mismatch between the TFA score and test result.

In addition to contributing new knowledge to the field of vocabulary, this project also helps me to develop new research knowledge and skills. In Section 7.2, I explain information regarding my research journey to show why long-term retention should be taken into consideration to promote sustainable learning.

## 7.2 My research journey

In my case, as a language lecturer and researcher, independent learning seems to yield benefits to the construction of long-term and in-dept knowledge. Learning components (*Motivation, Noticing, Retrieval* and *Generative Use*) that were the focus of my investigation was found to facilitate my new knowledge and skills in doing the current research. My very first motivation was to pursue a higher education. That stimulation was determined, challenging and self-selected.

With high motivation,  I then began searching for topics of interest. Among those, I noticed the significance of TFA framework in the field of vocabulary. The framework raised my awareness of effective learning materials for language learners as well as its design and evaluation. Due to this, I paid more attention to previous studies related to the TFA framework and synthesise the information gathered from them until I identified a research gap. Working memory process was involved during the synthesis process. Back and forth between remembering and forgetting, I started to retain the newly learned information (or input) when new knowledge was processed and stored in my working memory. The two memory processes which are generative rehearsal and multiple retrievals in combination seem to be a vital stimulators triggering retention of the newly learned knowledge in my long-term store. The information that was rehearsed and retrieved for multiple times was likely to be remembered longer and vice versa. For example, I reviewed all articles related to TFA multiple times so I knew the authors and the years of research related to TFA better than the information of articles in other fields of vocabulary research. The review can be seen as receptive generative use of the same topic in different contexts. As I was noticed and retrieved the information more often, it became easier to recall compared to the very first days of reading and learning. This is because this working memory process involves rehearsal, leading to long-term retention. Therefore, my successful learning journey tends to involve *Motivation, Noticing, Retrieval* and *Generative Use* in combination. However, doing research does not only require a skill of synthesis to summarise and integrate the information, but also need analytical skills to investigate the research questions. I continued to develop my research skills to expand my knowledge on research. This attempt was also self-motivated and self-selected. In the following sections, I will explain my noteworthy research journey that contributes to my new knowledge in doing research in the field of vocabulary.

### 7.2.1 Statistics

When I was first aware that Mixed-effects Model analysis has becoming more well-accepted and useful to recent research in the field of vocabulary, I started to pay considerable attention to the reasons behind its effectiveness. This was driven by both my intrinsic motivation (curiosity and desire to comprehend new knowledge) and extrinsic motivation (current trends and suggestions from my supervisors). The learning process was similar to that of my research journey explained earlier. However, I had to start from scratch since it was a completely new concept to me at the time. Unexpectedly, it took me a whole year to comprehend the fundamental concepts and details related to Mixed-effects Models. Although I was confident that my background knowledge from joining the intensive Math and Science programme in High School would help to facilitate my higher Mathematics skills, I struggled to learn difficult concepts in statistics, namely the Mixed-effects Models. During the long journey of my self-study, I have learned that knowing mathematics helped me with statistics, but not much with advanced statistics for complex data. For the first few months, this awareness pushed me to keep trying and my motivation was high. I then reviewed various articles (i.e., Brown, 2021; Cunnings, 2012) and books (i.e., Winter, 2020; Field et al., 2012) related to mixed-effects analyses. While I was motivated to learn and read more, I started to realise that I could not clearly understand the concept even though I spent more than three months reading the same concept. This showed that only high motivation sometimes could not drive to success. So, I stopped reading about statistics for a while and concentrated more on how to implement the linear Mixed-effect Model for data analysis of my study.

From my reading and my supervisors' advice, I realised that the R programme has been widely used to analyse the Mixed-effect Models. However, I had to start from scratch again since I was not familiar with the programme at all. Another new concept about R programme was focused, learned, and used. My journey in learning about R will be explained further in Section 7.2.2 below. After a while, I began to write a syntax using the Mixed-effects Model to analyse my research data. Unfortunately, with a lack of in-dept understanding about the statistics and enough experience in using R programme, the analysis did not go well as expected. Yet, I kept trying since I was aware that this analysis method would bring more benefits than drawbacks to my study. This makes me realise that awareness, a fundamental element of *Noticing*, is one of the key factors for my success. Back and forth between reading and implementing various models for trials, I could finally achieved my goal. My deeper comprehension started when I paid considerable attention (*Noticing*) to learning the two concepts, reviewed the concepts from various sources (*Generative Use*), and retrieved (*Retrieval*) the knowledge from both reading and trials for multiple times. This illustrated that in-dept understanding requires learning factors that involve working memory process to link the new knowledge to the existing knowledge in the brain. Learning by doing could link the concepts to real practice and create a strong connection between knowledge and usage. Therefore, I recommend that sustainable learning in all levels should rely on the learning components: *Noticing, Retrieval* and *Generative Use* that facilitate working memory.

In the following section, I give information about how I overcome challenges in self-regulated learning and using R programme though the key learning factors: *Motivation, Noticing*, *Retrieval* and *Generative Use*.

### 7.2.2 R programme

As mentioned earlier, I had no experience in using R programme although I had at least six-year teaching experience at that time. To be honest, I had never heard of it prior to the day it was introduced to me by one of my supervisors. However, a surge of interest in using R programme for vocabulary research stimulated my curiosity and attempt in learning how to make use of it after a quick review. Similar to my statistics learning journey, motivation was also found to be the primary factor. Driven from my motivation to know more about the up-to-date statistics and programmes for data analysis, my original intention to employ SPSS, an analysis program I was familiar with, was replaced by the choice to use the R programme because of its benefits to my study. Even though I realised that R programme requires users to write syntax to generate models for analysis, I still had a strong intention to try. My statistics learning journey taught me one thing that reading alone could not facilitate my in-dept understanding. I then started with installing this free programme to practice. This seems to be the first benefit that I can get from R programme. It allows everyone to download for free and is applicable for both Windows and MacOS. However, without enough reading I did not have sufficient knowledge to understand how to create syntax as a model for mixed-effects analysis in the programme. I had to study from Winter's (2013) basic tutorials, articles and other sources including YouTube tutorials. I then started learning step-by-step. A baseline model was created for trial. Unexpectedly, the first stage of trial was unsuccessful. The programme could not present the result in the R output and I could not understand the red warning messages appeared on the output screen. I struggled with modifying the model because I did not understand the meaning of the error messages on the screen. One of the problems is that most tutorials did not describe how to prepare a data file for R programme. The programme requires a long format data in an Excel file for the analysis of repeated-measure designs. It took me a long time to realise this issue from a YouTube tutorial.

Due to this, I realised that I need further reading from other sources. Books written by Winter (2020) and Field and his colleagues (2012) were another useful source to link my knowledge from reading to practice. After reading for a few weeks, I found out that knowledge from reading decays over time, and again reading alone did not help me much with understanding the concepts, but had tired me out. I began to select only relevant topics to read by looking at keywords in the index of the books and did the trials using the baseline model before developing more models for analysis. However, reading from books greatly helped with checking assumptions, another obstacle of the current study. To the best of my knowledge, repeated measurement can be designed differently, so details about checking assumptions are not included in published articles using R for Mixed-effects analyses. While books explains details of checking assumptions, the information tends to be general and too board. Searching for R packages and codes (syntax) for checking each assumption was found to be one of the causes that delays my study. I needed to adapt the general concepts to apply to my specific research design. This was time-consuming because there are many assumptions and many of them could not be simply interpreted. When some assumptions were not satisfied, I had to spend more time reading and searching for alternative options for correcting the problems. To overcome the challenges, my learning process mainly involved trial and error. This eventually led to deeper understanding of how to apply knowledge from various sources to implement with my study. My self-study journey on R programme made me realise the significance of working memory that help to support my life-long learning, which is a never ending journey. My knowledge on R has been profoundly expanded and developed through the processes of generative use and multiple retrievals days by days until now.

Based on my personal experience, R seems to be a useful tool for statistical analysis. Although there are many obstacles during my learning journey, I have learned that using R can also help me to understand more about statistics. The programme requires deeper statistics knowledge than SPSS because users have to write syntax as models for statistics analysis. Syntax models can be modified for several purposes such as releveling the reference level. However, I would recommend that a new user of R should start learning basic statistics before using it to analyse more complex measurements. Last but not least, schools or universities should introduce some basic practices of R programme in a Statistics course for postgraduate students in the field of Education and Linguistics as it has becoming more widely used at present.

In the following section, I identify some potential issues that could be considered as research gaps for further research. Recommendations according to the findings of the current study are also given as a guideline for conducting future studies.

## 7.3 Recommendations for further research

Language teachers and researchers should realise the significance of materials development and evaluation. Vocabulary materials evaluation by using the effective framework is suggested for vocabulary instructions in order to support memory retention for a sustainable learning of language learners. Although the findings confirmed that *Noticing, Retrieval*, and *Generative Use* facilitates memory process of single words in controlled written production, it is still not clear if they can also support muti-word units to the same extent. To gather comprehensive reviews on the effectiveness of the framework, I recommend that more research on the effects of TFA framework and its components on muti-word units should be conducted in the future.

Since noticing is likely to connect with retrieval and generative use at different extent, more studies related to the relationship between noticing and retrieval or noticing and generative use would bring insights to the field of vocabulary learning and memory retention in the future. The extent to which noticing affects retrieval and generative use is still unclear. Knowing this aspect would support language researchers and material developers in designing appropriate language learning lessons and creating effective vocabulary materials.

Moreover, further studies on vocabulary should realise more on the importance of sustainable word learning that associates with cognitive functions in working memory. Longitudinal research on the investigation into the long-term effects of *Motivation, Noticing, Retrieval* and *Generative Use* on form recall in other aspects is still limited. Yet, researchers should aware of primming effect and the impact on implementing the same test over times. It is recommended that test items should be shuffled if the same items are included in the following tests. The insights into other factors such as primming effect that may lead to different results on the effects of *Motivation, Noticing, Retrieval* and *Generative Use* on recall knowledge would help researchers to clearly understand how productive vocabulary knowledge is developed. Therefore, test items analysis together with long-term eye tracking studies that related to form recall knowledge are recommended to further research project in the field of vocabulary.

Also, future research could measure other aspects of vocabulary knowledge in other modalities (e.g., speaking, free writing). This should be conducted in various contexts to explore the similarities and differences of the results. It is recommended that the comparison of knowledge gains from free witing and controlled writing and/or between L2 and EFL learners would provide significant insights into how productive knowledge is learned. Retention of the knowledge would promote sustainable learning. Therefore, I suggest that more research into long-term retention of productive knowledge should be conducted to expand our understanding on its process of development as well as factors affecting retention and memory. Moreover, future research following the other view (i.e., continuum) of word knowledge should pay attention to both form and meaning. This is because the development of productive use may require receptive knowledge regarding the continuum view, meaning that productive use requires that users start with a meaning and then express that meaning in a form along a continuum.

Last but not least, researchers should not apply only single method for analysis to gather comprehensive results. The integration between the analysis of test scores (quantitative data) and interview (qualitative data), which can be called mixed methods (Dang et. al, 2023) might bring more understanding to nature of vocabulary learning and various effects of factors affecting retention of form recall. Previous vocabulary research (Godfroid & Schmidtke, 2013) adopted the mixed methods design to explore the effects of attention and awareness towards test results, and found significant benefits of the integration between quantitative and qualitative data. The value of mixed methods research approach added to experimental studies might better help to ensure the results of both quantitative and qualitative data and the validity of the evaluation of TFA-supported materials as well as the test results. The findings from the mixed methods may provide in-depth information helping to (1) ensure the validity of vocabulary materials, (2) identify factors affecting word retention and (3) fully understand the phenomenon exiting in actual classroom learning. For the purpose of accurate data interpretation, a semi-structure interview and questionnaire should be included to supplement data from the experimental study in future research. However, reporting separate findings of quantitative and qualitative data is not considered as mixed methods design (Dang et. al., 2023) although both types of data are collected in a single study. The integration of the interview data and test data should be carefully analysed for comparison so that appropriate conclusion can be drawn from this methods of analysis. Regarding of its impact, the mixed method has not been widely used in the field of vocabulary (Dang et al., 2023). Most studies, including the current research project relied on quantitative data for measuring word gains and retention. Due to this, I suggest that further studies should consider using the mixed methods research approach for the investigation of vocabulary gains and retention.

## 7.4 Summary

Technique Feature Analysis (TFA) framework has been widely accepted as an effective vocabulary framework for more than a decade. The current study has shed light on its predictive power on form recall in written production. Insights from quantitative data analysis data of this study also give a clearer picture of potential TFA factors affecting vocabulary learning and retention of productive form recall. While *Motivation, Noticing, Retrieval* and *Generative Use* can lead to learning, they differ in degree of support on retention of word form. Importantly, this thesis makes valuable contributions to vocabulary framework evaluation and validation. It captured the mismatch between TFA evaluated scores and test scores, suggesting that further improvement of the criteria within the framework as well as a proper training for raters should be considered. Last but not least, it provides implications for practice and recommendations for further research into vocabulary materials development and form recall knowledge.

## References

AbManan, N.A., Azizan, N. and Nasir, N.F.W.M. 2017. Receptive and Productive Vocabulary Level of Diploma Students from a Public University in Malaysia. *J. Appl. Environ. Biol. Sci.* **7**(1S), pp.53-59.

Adams, R. 2003. L2 Output, Reformulation and Noticing: Implications for IL Development. *Language Teaching Research*. **7**(3), pp.347-376.

Alfaki, I. M. 2015. University Students' English Writing Problems: Diagnosis and Remedy. *International Journal of English Language Teaching*. **3**(3), pp. 40-52.

Alonso, A. C., and Garcia, M, A. 2014. Productive Vocabulary Knowledge of Spanish EFL Learners. *Revista Electronica de Linguistica Aplicada*. **13**(1), pp.39-56.

Amiri, B. and Salehi, H. 2017. Impacts of Applying Crossword Puzzles on Improving Spelling Among Iranian Intermediate EFL Male and Female Learners. *Asian Journal of Education and E-Learning*. **5**(5), pp.159-171.

Anderson, J. 1982.Acquisition of Cognitive skill. *Psychological Review*. **89,** pp. 369-406.

Anderson, J. R. 1995. Cognitive Psychology and its Implications. New York: Free-man.

Atkinson, R. C. and Shiffrin, R. M. 1968. Human Memory: A Proposed System and its Control Processes. *Psychology of Learning and Motivation.* **2**, pp.89-195.

Aviad-Levitzky, T., Laufer, B., and Goldstein, Z. 2019. The New Computer Adaptive Test of Size and Strength (CATSS): Development and Validation. *Language Assessment Quarterly*. **16**(3), pp.345-368.

Avila, A. and Sadoski, M. 1996. Exploring New Application of The Keyword Method to Acquire English Vocabulary. *Language Learning*. **46**(3), pp.379-395.

Baddeley, A. 1990. *Human Memory*. London: Lawrence Erlbaum Associates.

Baddeley, A. 1997. *Human Memory: Theory and Practice*. Hove, UK: Psychology Press.

Bahrick, H. P. 1979. Broader Methods and Narrower Theories for Memory

Research: Comments on the Papers by Eysenck and Cermak. In L. S. Cermak and F. I. M. Craik (Eds). *Levels of Processing in Human Memory.* (pp.141–156). Hillsdale, NJ: Erlbaum.

Bailey, C. M., Hsu, C. T. and Dicarlo, S. E. 1999. Educational Puzzles for Understanding Gastrointestinal Physiology. *Advances in Physiology Education.* **21**(1), pp.S1-S18.

Baleghizadeh, S. and Shafeie, S. 2017. Studying the Effect of Retrieval Direction during Reading on Productive and Receptive Knowledge of Vocabulary. *Issues in Language Teaching*. **6**(2), pp.239-258

Barcroft, J. 2009. Effects of Synonym Generation on Incidental and Intentional L2 Vocabulary Learning During Reading. *TESOL Quarterly.* **43**(1), pp.79-103

Barrett, S. 2001. The impact of training on rater variability. *International Education Journal*. **2**(1), pp.49–58.

Barry, K. and King, L. 2000. *Beginning Teaching and Beyond* (3rd Eds). Katoomba, NSW: Social Science Press.

Battista, J. R. (1978). The science of consciousness. In K. S. Pope & J. L. Singer (Eds.) *The stream of consciousness: Scientific Investigations into the Flow of Human Experience* (pp.55-87). New York: Plenum Press.

Benzies, Y. J. C. 2013. Spanish EFL University Students’ Views on the Teaching of Pronunciation: A Survey-Based Study. *Language Studies Working Papers.* **5**, pp.41-49.

Boonyarattanasoontorn, P. 2017. An investigation of Thai students' English language writing difficulties and their use of writing strategies. *Journal of Advanced Research in Social Sciences and Humanities*. **2**(2), pp.111-118.

Bower, K. 2019. Explaining Motivation in Language Learning: A Framework for Evaluation and Research. *The Language Learning Journal*. **47**(5), pp. 558-574.

Brown, H. D. 1990. M & Ms for Language Classrooms? Another Look at Motivation. In Alatis, J. E. (Ed.) *Georgetown University Round Table on Language and Linguistics* (pp.383–93). Washington: Georgetown University Press.

Brown, R., Waring, R. and Donkaewbua, S. 2008. Incidental Vocabulary Acquisition from Reading, Reading-While-Listening, and Listening to Stories. *Reading in a Foreign Language.* **20**(2), pp.136-163.

Candry, S., Decloedt, J. and Eyckmans, J. 2020. Comparing the Merits of Word Writing and Retrieval Practice for L2 Vocabulary Learning. *System*. **89**, pp.1-11.

Carpenter, S. K. and Olson, K. M. 2012. Are Pictures Good for Learning New Vocabulary in A Foreign Language? Only If You Think They Are Not. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* **38**, pp.92-101.

Chaharlang, N. and Farvardin, M. T. 2018. Predictive Power of Involvement Load Hypothesis and Technique Feature Analysis across L2 Vocabulary Learning Tasks. *International Journal of Foreign Language Teaching and Research*. **6**(24), pp. 127-141.

Chen, Y. 2002. The Problems of University EFL Writing in Taiwan. *The Korea TESOL Journal*. **5**(1), pp.59-79.

Chen, Y. 2011. Dictionary Use and vocabulary learning in the context of reading. *International Journal of Lexicography*. December, pp.1-32.

Cho, K, and Krashen, S. 1994. Acquisition of vocabulary form the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading*. **37**(8), pp.662-667

Clenton, J. 2010. Investigating the construct of productive vocabulary knowledge with Lex30. Thesis, Swansea University. Retrieved from http://cronfa.swan.ac.uk/Record/cronfa42281

Cobb, T. 2015. The Complete Lexical Tutor. [Online]. Available from: https://www.lextutor.ca

Corson, D. 1995. *Using English Words*. Dordrecht, The Netherlands: Kluwer.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly*. **34**(2), pp.213-238

Coyle, D. 2011. Investigating Student Gains: Content and Language Integrated Learning. ITALIC Research Report. pp. 1-112. Edinburgh: University of Aberdeen, Esmée Fairbairn Foundation. Retrieved from https://www.abdn.ac.uk/italic/documents/ITALIC_Report_-_Complete_Version.pdf

Craik, F. I. M. and Lockhart, R. S. 1972. Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*. **11**, pp.671-684.

Craik, F. I. M., and Tulving, E. 1975. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General.* **104**(3), pp.268-294.

Crossley, S. A., Subtirelu, N. and Salsbury, T. 2013. Frequency Effects or Context Effects in Second Language Word Learning. *Studies in Second Language Acquisition.* **35**, pp.727–755.

Dang, T. N. Y. 2020. Vietnamese Non-English Major EFL University Students' Receptive Knowledge of the Most Frequent English Words. *VNU Journal of Foreign Studies*. **36**(3).

Dang, T. N. Y., Coxhead, A., Webb, S. 2017. The Academic Spoken word List. *Language Learning*. **67**, pp.959–997.

Dang, T. N. Y., Lu, C. and Webb, S. 2021. Incidental Learning of Single Words and Collocations Through Viewing an Academic Lecture. *Studies in Second Language Acquisition*. Cambridge University Press, pp.1–29. doi: 10.1017/S0272263121000474.

Dang, T. N. Y., Vu., D. V. and Nguyen, T. M. H. 2023. Researching vocabulary: mixed methods research. In C.A. Chapelle (Ed.) *The Encyclopedia of Applied Linguistics*. John Wiley & Sons Ltd., pp. 1-6. doi: https://doi.org/10.1002/9781405198431.wbeal20015

Dang, T. N. Y. and Webb, S. 2016a. Evaluating Lists of High-Frequency Words. *ITL-International Journal of Applied Linguistics*. **167**(2), pp.132–158.

Dang, T. N. Y. and Webb, S. 2016b. Making an essential word list. In I. S. P. Nation (Ed.) *Making and Using Word Lists for Language Learning and Testing* (pp.153–167). Amsterdam: John Benjamins.

Day, R. R., Omura, C., and Hiramatsu, M. 1991. Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*. **7**(2), pp.541-551

Deci, E. L. and Ryan, R. M. 1985. *Intrinsic Motivation and Self-Determination in Humanbehavior.* New York: Plenum.

DeKeyser, R. M., and Sokalski, K. J. 1996. The Differential Role of

Comprehension and Production Practice. *Language Learning*. **46**, pp. 613–642.

*Dörnyei*, Z. *1994*. Motivation and Motivating in the Foreign Language Classroom. *Modern Language Journal*. **78**(3), pp.273–284.

Dörnyei, Z. 2001a. New Themes and Approaches in Second Language Motivation Research. *Annual Review of Applied Linguistics*. **21**, pp.43-59.

*Dörnyei*, Z. and *Csizer*, K. *1998*. Ten Commandments for Motivating Language Learners Results of an Empirical Study. *Language Teaching Research*. **2**, pp.203-229.

Elley, W. B. 1989. Vocabulary Acquisition from Listening to Stories. *Reading Research Quarterly*. **24**(2), pp174-187.

Ellis, R. 1995. Modified Oral Input and the Acquisition of Word Meanings. *Applied Linguistics*. **16**(4), pp.409-441.

Ellis, R., Tanaka, Y., and Yamazaki, A. 1994. Classroom Interaction, Comprehension, and the Acquisition of L2 Word Meanings. *Language Learning*. **44**, pp.449-491.

Elliott, A. C. and Woodward, W. A., 2007. Statistical Analysis Quick Reference Guidebook. Thousand Oaks, CA: SAGE Publications, Inc.

Ericsson, K. A. and Simon, H. A. 1984. *Protocol analysis*: *Verbal Reports as Data*. Cambridge, MA: The MIT Press.

Erten, I. and Tekin, M. 2008. Effects of Vocabulary Acquisition of Presenting New Words in Semantic Sets Versus Semantically Unrelated Sets. *System*. **36**, pp.407-422.

Fareed, M., Ashraf, A., and Bilal, M. 2016. ESL learners' writing skills: problems, factors and suggestions. *Journal of Education and Social Sciences*. **4**(2), pp.81-92.

Folse, K. 2006. The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*. **4**(2), pp.273-293.

Folse, K. 2011. Applying L2 lexical research findings in ESL teaching. *TESOL Quarterly.* **45**(2). pp. 362-369.

Fowle, C. 2002. Vocabulary Notebooks: implementation and outcomes. *English Language Teaching Journal*. **56**(4), pp.380-388. https://doi.org/10.1093/elt/56.4.380

Gardner, R. C. 2001. Integrative Motivation and Second Language Acquisition. In Z. Dörnyei, and R. Schmidt (Eds.) *Motivation and Second Language Acquisition* (pp.1-19). Hawaii: University of Hawaii Press.

Gardner, R. C. and MacIntyre, P. D. 1993. A student's contributions to second language learning. part II: affective variables. *Language Teaching*. **26**, pp.1-11.

Garnier, M. and Schmitt, N. 2016. Picking up polysemous phrasal verbs: how many do learners know and what facilitates this knowledge? *System.* **59**, pp.29-44.

Ghasemi, A., and Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*. **10**, pp. 486-489.

Gibriel, M. 2017. A Cross-Sectional Study of Egyptian EFL Student-Teachers' Vocabulary Size. *The Journal of Asia TEFL*. **14**(1), pp.189-196.

Godfroid, A. and Schmidtke, J. 2013. What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. In J.M. Bergsleithner, S.N. Frota, and K. Yoshioka (eds) *Noticing and second language acquisition: Studies in honor of Richard Schmidt.* Honolulu, HI: National Foreign Language Resource Center, pp. 183–205

Gohar, M. J., Rahmanian, M., and Soleimani, H. 2018. Technique Feature Analysis or Involvement Load Hypothesis: Estimating Their Predictive Power in Vocabulary Learning. *J Psycholinguist Res*. **47**, pp. 859–869

Goldschneider, J. M. and DeKeyser, R. M. 2001. Explaining the "Natural Order of L2 Morpheme Acquisition" in English: A Meta-analysis of Multiple Determinants. *Language Learning*. **51**(1), pp.1-50.

González-Fernández, B. and Schmitt, N. 2020. Word Knowledge: Exploring the Relationships and Order of Acquisition of Vocabulary Knowledge Components, *Applied Linguistics*. **41**(4), pp.481–505.

Griffin, G. F. and Harley, T. A. 1996. List learning of Second Language Vocabulary. *Applied Psycholinguistics*. **17**, pp.443–460.

Grimm, R., Cassani, G., Gillis, S., and Daelemans, W. 2019. Children

Probably Store Short Rather Than Frequent or Predictable Chunks: Quantitative Evidence from a Corpus Study. *Front. Psychol.* **10**. [Online]. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00080/full

Gu, P. Y. 2003a. Vocabulary Learning in a Second Language: Person, Task, Context and Strategies. *TESL-EJ.* **7**(2), pp.1-28.

Gu, P. Y. 2003b. Fine Brush and Freehand: The Vocabulary-Learning Art of Two Successful Chinese EFL Learners. *TESOL Quarterly*. **37**, pp.73-104. http://dx.doi.org/10.2307/3588466

Gu, P., and Johnson, R. K. 1996. Vocabulary learning strategies and language learning outcomes. *Language Learning*, **46**, 643-679. doi:10.1111/j.1467- 1770.1996.tb01355

Hamada, Y. and Kito, K. 2008. Demotivation in Japanese High Schools. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), JALT2007 Conference Proceedings. Tokyo: JALT.

Hamilton, H. 2012. The Efficacy of Dictionary Use While Reading for Learning New Words. *American Annals of the Deaf*. **157**(4), pp. 358–372. Received from http://www.jstor.org/stable/26234850

Hashemian, M. and Heidari, A. 2013. The Relationship between L2 Learners' Motivation/Attitude and Success in L2 Writing. *Procedia-Social and Behavioral Sciences.* **70**, pp.476-489.

Hazrat, M. 2020.The Involvement Load Hypothesis and Its Impact on Vocabulary Learning. Unpublished doctoral thesis. New Zealand: The University of Auckland. Retrieved from https://researchspace.auckland.ac.nz/handle/2292/51729

Hazrat, M., and Read, J. 2022. Enhancing the involvement load hypothesis as a tool for classroom vocabulary research. *TESOL Quarterly*. **56**(1), pp. 387-400.

Hemchua, S. and Schmitt, N. 2006. An analysis of lexical errors in the English compositions of Thai learners. *Prospects*. **21**(3), pp.3-25.

Henriksen, B. 1999. Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition*. **21**(4), pp.303-317.

Hirata, M. C. 2019. Evaluating Multi-Skill Vocabulary Activities Using the

Technique Feature Analysis (TFA) Framework. *The Journal of Asia TEFL.* **16**(1), pp.377-384.

Hirsh, D. 2015. *Researching Vocabulary* in Paltridge, B and Phakiti, A. Research Methods in Applied Linguistics: A Practical Resource. Bloombery Publishing. London.

Hirsh, D., and Nation, P. 1992. What Vocabulary Size Is Needed to Read Unsimplified Texts for Pleasure. *Reading in a Foreign Language*. **8**, pp.689-696.

Hoyt, W. T. and Kerns, M. 1999. Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*. 4, pp. 403–424.

Hu, H. C. M. and Nassaji, H. 2016. Effective Vocabulary Learning Tasks: Involvement Load Hypothesis versus Technique Feature Analysis. *System*. **56**, pp. 28-39.

Hu, M. and Nation, I.S.P. 2000. Vocabulary Density and Reading Comprehension. *Reading in a. Foreign Language*. **13**(1), 403–30.

Huang, S. 2018. A Verification of Involvement Load Hypothesis on Chinese Adult English Learners. *International Journal of English Linguistics*. **8**(5), pp.125-134.

Hulstijn, J. H. 1993. When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables. *The Modern Language Journal*. **77**, pp.139-147.

Hulstijn, J. H. and Laufer, B. 2001. Some Empirical Evidence for the Involvement Load Hypothesis in Vocabulary Acquisition. *Language Learning*. **51**(3), pp.539-558

Imorde, L., Möltner, A., Runschke, M., and Weberschock, T. 2020. Adaptation and validation of the Berlin questionnaire of competence in evidence-based dentistry for dental students: a pilot study. *BMC Medical Education*. **20**, p.136. https://doi.org/10.1186/s12909-020-02053-0

Joe, A. 1995. Text-based Tasks and Incidental Vocabulary Learning. *Second Language Research*. **11**(2), pp.149-158.

Joe, A. 1998. What Effects do Text-based Tasks Promoting Generation have on Incidental Vocabulary Acquisition? *Applied Linguistics*. **19**(3), pp.

357-377.

Johnson, R. L., Penny, J. A. and Gordon, B. 2008. *Assessing performance*. New York, NY: The Guilford Press

Kachru, B. B. 1985. Standards, Codification and Sociolinguistic Realism: English Language in the Outer Circle. In R. Quirk and H. Widowson (Eds.) *English in the World: Teaching and Learning the Language and Literatures* (p.11-36). Cambridge: Cambridge University Press.

Kahneman, D. 1973. *Attention and Effort.* Englewood Cliffs, NJ: Prentice-Hall.

Karpicke, J. D. and Bauerenschmidt, A. 2011. Spaced Retrieval: Absolute Spacing Enhances Learning Regardless of Relative Spacing. *Journal of experimental psychology. Learning, Memory, and Cognition.* **37**(5), pp.1250-1257.

Karpicke, J. D. and Roediger, H. L. 2008. The Critical Importance of Retrieval for Learning. *Science.* **319**, pp.966-968.

Keating, G. D. 2008. Task Effectiveness and Word Learning in a Second Language: The involvement Load Hypothesis on Trial. *Language Teaching Research*. **12**(3), pp.365–386.

Khoshsima, H. and Eskandari, Z. 2017. Task Effectiveness Predictors: Technique Feature Analysis Versus Involvement Load Hypothesis. *Iranian Journal of English for Academic Purposes.* **6**(2), pp.50-69.

Kihlstrom, J. F. 1984. Conscious, Subconscious, Unconscious: A Cognitive Perspective. In K. S. Bowers and D. Meichenbaum (Eds.) The Unconscious Reconsidered (pp.149-211). New York: Wiley.

Kiliç, M. 2019. Vocabulary Knowledge as a Predictor of Performance in Writing and Speaking: A Case of Turkish EFL Learners, *PASAA*. **57**, pp.133-164.

Kim, Y. 2008. The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition. *Language Learning*. **58**(2), pp.100-140.

Knight, S. 1994. Dictionary: The tool of last resort in foreign language reading? *Modern Language Journal*. **78**(3), 285-299.

Kosslyn, S. M. 1980. *Images and Mind*, Harvard University Press, Cambridge, MA.

Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics*. **33**, 159-174.

Larsen-Freeman, D. E. 1976. An Explanation for the Morpheme Acquisition Order of Second Language Learners. *Language Learning*. **26**, pp.125-134.

Laufer, B. 1992. How Much Lexis Is Necessary for Reading Comprehension. In P. J. L. Arnaud, and H. Bejoing (Eds.) *Vocabulary and Applied Linguistics* (pp.129-132). London: Macmillan.

Laufer, B. 1998. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*. **19**(2), pp.255-271.

Laufer, B., Elder, C., Hill, K. and Congdon, P. 2004. Size and Strength: Do We Need Both to Measure Vocabulary Knowledge? *Language Teaching*. **21**, pp.202-226.

Laufer, B. and Goldstein, Z. 2004. Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning.* **54**(3), pp.399–436.

Laufer, B. and Nation, P. 1995. Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied linguistics.* **16**(3), pp.307-322.

Laufer, B. and Nation, P. 1999. A Vocabulary-Size Test of Controlled Productive Ability. *Language testing.* **16**(1), pp.33-51.

Laufer, B. and Paribakht, T. S. 1998. The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context. *Language Learning*. **48**(3), pp.365-391.

Laufer, B. and Waldman, T. 2011. Verb-noun Collocations in Second Language Writing: A Corpus Analysis of Learners’ English. *Language Learning.* **61**(2), pp.647-672.

Lumley, T. and McNamara, T. F. 1995. Rater characteristics and rater bias: Implications for training. *Language Testing*. (12), pp.54–72.

Luppescu, S. and Dat, R. 1993. Reading, dictionaries, and vocabulary learning, Language Learning. **43**(2), pp.263-287.

McCauley, S. M., and Christiansen, M. H. 2014. Acquiring Formulaic Language: A Computational Model. *Ment. Lex.* **9**, pp.419-436.

McCowan, R.J. and McCowan, S.C. 1999. Item Analysis for Criterion-Referenced Tests. Available at: https://eric.ed.gov/?id=ED501716. Accessed November 22, 2022.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*. **22**(3), pp.276-282.

Meara, P. 1996. The Vocabulary Knowledge Framework. *Vocabulary Acquisition Research Group Virtual Library.*

Meara, P. and Bell, H. 2001. P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*. **16**(3), pp.5-19.

Meara, P. and Fitzpatrick, T. 2000. Lex30: An Improved Method of Assessing Productive Vocabulary in an L2. *System*. **28**(1), pp.19-30.

Mehrpour, S. and Rahimi, M. 2010. The Impact of General and Specific Vocabulary Knowledge on Reading and Listening Comprehension: A Case of Iranian EFL Learners. *System*. **38**(2), pp.292-300.

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

Milton, J. and Fitzpatrick, T. 2014. *Dimensions of Vocabulary Knowledge.* Basingstoke: Palgrave Macmillan.

Moon, R. 1997. Vocabulary Connections: Multi-word Items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy.* (pp. 40-63). Cambridge: Cambridge University Press.

Mungkonwong, P. and Wudthayagorn, J. 2017. An Investigation of Vocabulary Size of Thai Freshmen and Its Relationship to Years of English Study. *LEARN Journal: Language Education and Acquisition Research Network*. **10**(2), pp.1-9.

Myers, G. C. 1914. A Comparative Study of Recognition and Recall. *Psychological Review*, **21**(6), pp.442-456.

Nakata, T. and Webb, S. A. 2016. Vocabulary Learning Exercises: Evaluating a Selection of Exercises Commonly Featured in Language Learning Materials. In B. Tomlinson (Ed.) *SLA Research and Materials Development for Language Learning.* (pp. 123-138). Oxon, UK: Routledge.

Nation, I. S. P. 1990. *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, I. S. P. 1997. L1 and L2 use in the classroom: a systematic approach. *TESL Reporter*. **30**(2), pp.19-27.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. 2006. How Large a Vocabulary Is Needed for Reading and Listening? *Canadian Modern Language Review*. **63**(1), pp.59-81.

Nation, I. S. P. 2007. The Four Strands. Innovation in Language Learning and Teaching. **1**(1), pp.1-12.

Nation, I. S. P. 2013. *Learning Vocabulary in Another Language*. (Eds) New York: Cambridge University Press.

Nation, P. and Beglar, D. 2007. A vocabulary size test. *The Language Teacher*. **31**(7), pp. 9-13.

Nation, I.S.P. and Meara, P. 2010. *Vocabulary*. In N. Schmitt (ed.) An Introduction to Applied Linguistics. Edward Arnold. Second edition. pp. 34-52

Nation, I.S.P. and Waring, R. 1997. Vocabulary Size, Text Coverage and Word Lists. In: N. Schmitt and M. McCarthy (Eds) *Vocabulary Description, Acquisition and Pedagogy.* (pp.6-19). Cambridge: Cambridge University Press.

Nation, I. S. P. and Webb, S. 2011. *Researching and analyzing vocabulary*. Boston, MA: Heinle. Newton, J. 2013. Incidental Vocabulary Learning in Classroom Communication Tasks. Language. *Teaching Research*. **17**(2), pp.164-187.

Noels, K.A., Pelletier, L.G., Clement, R. and Vallerand, R.J. 2000. Why are You Learning a Second Language? Motivational Orientations and Self-determination Theory. *Language Learning*. **50**, pp.57–85.

Noor, N.M. (2011). Reading habits and preferences of EFL post graduates: a case study. *Indonesian Journal of Applied Linguistics*. **1**(1), pp.1-9.

Oosterhof, A.C. and Coats, P.K. 1984. Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. *Applied Psychological Measurement*. **8**, pp. 287-294. doi:10.1177/014662168400800305

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), pp.554-566.

Özönder, Ö. 2016. Student EFL Teachers’ Receptive Vocabulary Size.

*Procedia-Social and Behavioral Sciences*. **232**. pp.444-450. doi: 10.1016/j.sbspro.2016.10.061

Palmer, H. E. 1921. *The Principles of Language-study.* World Book Company.

Paribakht, T. S. and Wesche M. 1997. Reading Comprehension and Second Language Development in a Comprehension-based ESL Program. *TESL Canada Journal*. **11**(1), pp.9-29.

Paribakht, T. S. and Wesche M. 1997. Vocabulary Enhancement Activities and Reading for Meaning in Second Language Vocabulary Development. In J. Coady and T. Huckin (Eds.) *Second Language Vocabulary Acquisition: A Rationale for Pedagogy.* (pp.174-200.) New York: Cambridge University Press.

Paribakht, T. S. and Wesche, M. 1999. Reading and “incidental” L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition*. **21**(2). pp.195-224. http://dx.doi.org/10.1017/ S027226319900203X.

Park, S. E. 2011. Learner-generated Noticing of Written L2 Input: What do Learners Notice and Why? *Language Learning.* **61**(1), pp.146–186.

Peacock, M. 1997. The Effect of Authentic Materials on the Motivation of EFL Learners. *ELT Journal*. **51**(2), pp.144-156.

Peters, E. 2016. The Learning Burden of Collocations: The Role of Interlexical and Intralexical Factors. *Language Teaching Research*. **20**(1), pp.113-138.

Phonna, R. 2014. An Analysis of Students’ Free Writing. *Englisia: Journal of Language*, *Education, and Humanities*. **1**(2), pp.270-279.

Philp J. 2003. Constraints on “noticing the gap” *Studies in Second Language Acquisition*. **25**, pp.99-126.

Ponniah, R. J. 2011. Incidental acquisition of vocabulary by reading. *The Reading Matrix*. **11**(2), pp.135-139.

Postman, L. and Rau, L. 1957. Retention as a Function of the Method of Measurement. *University of California Publications in Psychology*. **8**, pp.217-270.

Postman, L., Kruesi, E. and Regan, J. 1975. Recognition and Recall as Measures of Long-Term Retention. *The Quarterly Journal of Experimental Psychology*. **27**(3), pp.411-418.

Pressley, M. 1977. Children' s Use of the Keyword Method to Learn Simple Spanish Vocabulary Words. *Journal of Educational Psychology.* **69**(5), pp.465-472.

Puimège, E. and Peters, E. 2019. Learning L2 Vocabulary from Audiovisual Input: An Exploratory Study into Incidental Learning of Single Words and Formulaic Sequences, *The Language Learning Journal*. **47**(4), pp.424-438

Qi, D. S. and Lapkin, S. H. 2001. Exploring the Role of Noticing in a Three-Stage Second Language Writing Task. *Journal of Second Language Writing*. **10**, pp.277-303.

Qian, D. D. 2002. Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning*. **52**, pp.513-536.

Ramachandran, S. D. and Rahim, H. A. 2004. Meaning Recall and Retention: The Impact of the Translation Method on Elementary Level Learners' Vocabulary Learning. *RELC Journal*. **35**(2), pp.161-178.

Rattanadilok Na Phuket, P. and Othman, N. B. 2015. Understanding EFL Students' Errors in Writing. *Journal of Education and Practice*. **6**(32), pp.99-106.

Razali, N. M. and Wah, Y. B. 2011. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *Journal of Statistical Modeling and Analytics*. **2**, pp.21-33.

Rezaei, M and Davoudi, M. 2016. The Influence of Electronic Dictionaries on Vocabulary Knowledge Extension. *Journal of Education and Learning*. **5**(3), pp.139-148. URL: http://dx.doi.org/10.5539/jel.v5n3p139

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

Rott, S. 1999. The Effect of Exposure Frequency on Intermediate Language Learners' Incidental Vocabulary Acquisition and Retention Through Reading. *SSLA.* **21**, pp.589-619.

Sabri S. 2013. Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. *Int. J. Educ. Res.* **1**(12), pp.1-14.

Sakai, H., and Kikuchi, K. 2009. An Analysis of Demotivators in the EFL

Classroom. *System*. **37**(1), pp.57-69.

Sawir, E. 2005. Language difficulties of international students in Australia: The Effects of Prior Learning Experience. *International Education Journal*. **6**(5), pp. 567-580.

Schmidt, R. 1990. The Role of Consciousness in Second Language Learning. *Applied Linguistics*. **11**, pp.129 - 158.

Schmitt, N. 1997. Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), Vocabulary: Description, acquisition and pedagogy (pp.199-227). Cambridge: Cambridge University.

Schmidt, R. and Frota, S. 1986. Developing Basic Conversational Ability in a Second Language: A Case Study of an Adult Learner. In R. Day (Ed) *Talking to Learn*. Rowley, Mass: Newbury House.

Schmitt, N. 2000. *Vocabulary in language teaching.* Cambridge. New York: Cambridge University Press.

Schmitt, N. 2007. Current trends in vocabulary learning and teaching. In J. Cummins & C. Davison (Eds.), The International Handbook of English Language Teaching. (pp. 745-759). Springer.

Schmitt, N. 2008. Instructed Second Language Vocabulary Learning. *Language Teaching Research*. **12**(3), pp.329-363.

Schmitt, N. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test. *Language Testing*. **18**(1), pp.55-88.

Sharwood Smith, M. 1981. Conscious-Raising and the Second Language Learner. *Applied Linguistics*. **2**, pp.159-68.

Shen, Z. 2013. The Effects of Vocabulary Knowledge and Dictionary Use on EFL Reading Performance. *English Language Teaching*. **6**(6), pp.77-85. URL: http://dx.doi.org/10.5539/elt.v6n6p77

Shepard, R. N. and Cooper, L. A. 1982. *Mental Images and Their Transformations*. MIT Press, Cambridge, MA

Shintani, N. 2011. A Comparative Study of the Effects of Input-based and Production-based Instruction on Vocabulary Acquisition by Young EFL Learners. *Language Teaching Research.* **15**(2), pp.137-158.

Shiotsu, T. and Weir, C. J. 2007. The Relative Significance of Syntactic Knowledge and Vocabulary Breadth in the Prediction of Reading Comprehension Test Performance. *Language Testing*. **24**(1), pp.99-128.

Shokrpour, N. and Fallahzadeh, M. 2007. A Survey of the Students and Interns' EFL Writing Problems in Shiraz University of Medical Sciences. *Asian EFL Journal*. **9**(1), pp.147-163.

Shuman, J. H. 2014. Foreword. In Z. Dörnyei, P.D. MacIntyre, and A. Henry. (Ed) *Motivational Dynamics in Language Learning* (pp.xv–xix). Bristol: Multilingual Matters.

Simon, E. and Taverniers, M. 2011. Advanced EFL Learners' Beliefs about Language Learning and Teaching: A Comparison Between Grammar, Pronunciation, and Vocabulary. *English Studies*. **92**(8), pp. 896–922.

Smith, G. G., Li, M., Drobisz, J., Park, H. R., Kim, D. and Smith, S. D. 2013. Play Games or Study? Computer Games in Ebooks to Learn English Vocabulary. *Computers and Education*. **69**, pp.274-286.

Snelling, P., De Glopper, K. and Van Gelderen, A. 2004. The Effects of Enhanced Lexical Retrieval on Second Language Writing: A Classroom Experiment. *Applied Psycholinguistics*. **25**(2), pp.175-200.

Snoder, P. and Laufer, B. 2022. EFL Learners' Receptive Knowledge of Derived Words: The Case of Swedish Adolescents. *TESOL Quarterly*. **56**(4), pp. 1242-1265.

Sorace, A. 1985. Metalinguistic Knowledge and Language Use in Acquisition-poor Environments. *Applied Linguistics*. **6**(3), pp.239–254

Stæhr, L. S. 2008. Vocabulary Size and the Skills of Listening, Reading and Writing. *The Language Learning Journal*. **36**(2), pp.139-152

Stæhr, L. S. 2009. Vocabulary Knowledge and Advanced Listening Comprehension in English as a Foreign Language. *Studies in Second Language Acquisition*. **31**(4), pp.577-607.

Steinel, M. P., Hulstijn, J. H. and Steinel, W. 2007. Second Language Idiom Learning in a Paired-Associate Paradigm: Effects of Direction of Learning, Direction of Testing, Idiom Imageability, and Idiom Transparency. *Studies in Second Language Acquisition*. **29**(03), pp.449–484.

Steinskog, D.J., Tjøstheim, D. and Kvamstø, N.G. 2007. A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality. *Monthly Weather Review*. **135***,* pp.1151-1157.

Strauber, C. B., Sorcar, P., Howlett, C. and Goldman, S. 2020. Using a Picture-Embedded Method to Support Acquisition of Sight Words. *Learning and Instruction*. **65**, pp.1-9.

Sun, Y. and Dang, T.N.Y. 2020. Vocabulary in High-School EFL Textbooks: Texts and Learner Knowledge. *System.* **93**, pp.1-41

Suresh K. 2011. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of Human Reproductive Sciences*. **4**(1), pp.8-11.

Tang, C. and Treffers-Daller, J. 2016. Assessing incidental vocabulary learning by Chinese EFL learners: a test of the Involvement Load Hypothesis. In: Yu, G. and Jin, Y. (eds.) Assessing Chinese learners of English: language constructs, consequences and conundrums. Palgrave, London, pp.121-149. Available at http://centaur.reading.ac.uk/39023/

Tapia, J. L. and Duñabeitia, J. A. 2021. Improving Language Acquisition and Processing With Cognitive Stimulation. Front. Psychol. **12**, pp.1-5 https://doi.org/10.3389/fpsyg.2021.663773

Thompson, B., and Levitov, J. E. 1985. Using microcomputers to score and evaluate test items. *Collegiate Microcomputer*, **3**. Pp.163-168.

Thornbury, S. 2002. *How to teach vocabulary.* Harlow: Longman.

Tinsley, H. E. A., and Weiss, D. J. 2000. Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). New York: Academic Press.

Tollefson, N. 1987. A comparison of the item difficulty and item discrimination of multiple-choice items using the "none of the above" and one correct response options. Educational and Psychological Measurement. **47**(2), pp.377–383. https://doi.org/10.1177/0013164487472010

Tomczak, E. and Robert, L. 2019. The Song of Words: Teaching Multi- Word

Units with songs. *3L: Language, Linguistics, Literature.* **25**(4), pp.16-33.

Tomlinson, B. 2012. Materials Development for Language Learning and Teaching. *Language Teaching*. **45**(2), pp.143-179.

Tsuai, A. S. and Barry, B. 1986. Interpersonal affect and rating errors. *The Academy of Management Journal*. **29**(3), pp.586–599

Underwood, B. J., Zimmerman, J. and Freund, J. S. 1971. Retention of Frequency Information with Observations on Recognition and Recall. *Journal of Experimental Psychology.* **87**(2), pp.149–162.

Ushioda, E. 2003. Motivation as a Socially Mediated Process". In D. Little, J. Ridley and E. Ushioda (Eds.) *Learner Autonomy in the Foreign Language Classroom: Teacher, Learner, Curriculum and Assessment.* (pp.90-102) Dublin, Ireland: Authentik.

Unsworth S., Persson L., Prins T. and De Bot K. 2015. An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*. **36**(5), pp.527–548. https://doi.org/10.1093/applin/amt052

Van den Broek, S. E. G., Takashima, A., Segers, E. and Verhoeven, L. 2018. Contextual Richness and Word Learning: Context Enhances Comprehension but Retrieval Enhances Retention. *Language Learning*. **68**(2), pp.546-585

Van Gelderen, A., Oostdam, R. and Van Schooten, E. 2011. Does Foreign Language Writing Benefit from Increased Lexical Fluency? Evidence from a Classroom Experiment. *Language Learning*. **61**(1), pp.281-321.

Vanichvasin, P. 2021. Effects of Visual Communication on Memory Enhancement of Thai Undergraduate Students, Kasetsart University. *Higher Education Studies*, **11**(1), pp. 34-41

Vidal, K. 2011. A Comparison of the Effects of Reading and Listening on Incidental Vocabulary Acquisition. *Language Learning*. **61**(1), 219-258.

Vidal, K. 2003. Academic Listening: A Source of Vocabulary Acquisition? *Applied Linguistics*. **24**(1), pp.56–89.

Wang, J. 2011. The use of e-dictionary to read e-text by intermediate and advanced learners of Chinese. *Computer Assisted Language Learning*. **24**, pp.1-13.

Wang, X. M., Wong, K. F. E. and Kwong, J. Y. Y. 2010. The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology*. **95**(3), pp.546–561.

Ward, J. 2009a. A Basic Engineering English Word List for Less Proficient Foundation Engineering Undergraduates. *English for Specific Purposes*. **28**(3), pp.170-182.

Ward, J. 2009b. EAP Reading and Lexis for Thai Engineering Undergraduates. *Journal of English for Academic Purposes.* **8**(4), pp.294-301.

Waring, R. 1997a. A Study of Receptive and Productive Learning from Word Cards. *Studies in Foreign Languages and Literature.* **21**, pp.94–114.

Waring, R. 1997b. A Comparison of the Receptive and Productive Vocabulary Sizes of Some Second Language Learners. *Immaculata* (Notre Dame Seishin University, Okayama)*.* **1**, pp.53-68.

Waring, R., and Takaki, M. 2003. At What Rate do Learners Learn and Retain New Vocabulary from Reading a Graded Reader? *Reading in a Foreign Language*. **15**, pp.1-27.

Watcharapunyawong, S. and Usaha, S. 2013. Thai EFL Students' Writing Errors in Different Text Types: The Interference of the First Language. *English Language Teaching*. **6**(1), pp. 67-78.

Waugh, N. C. and Norman, D. A. 1965. Primary Memory. *Psychological Review*. **72**(2), pp.89–104.

Webb, S. 2005. Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition*. **27**, pp.33–52.

Webb, S. 2007a. The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics*, **28**(1), pp.46-65.

Webb, S. 2009. The Effects of Receptive and Productive Learning of Word Pairs on Vocabulary Knowledge. *RELC Journal*. **40**(3), pp. 360-376.

Webb, S. and Chang, A. 2015. How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition.* **37**(4), pp. 651-675. Cambridge University Press

Webb, S. and Nation, P. 2017. *How Vocabulry Is Learned*. Oxford, England:

Oxford University Press.

Webb, S., Sasao, Y., and Balance, O. 2017. The Updated Vocabulary Levels Test: Developing and Validating Two New Forms of the VLT. *International Journal of Applied Linguistics.* **168**(1), pp.33-69.

Wei, Z., and Nation, I. S. P. 2013. The Word Part Technique: A Very Useful Vocabulary Teaching Technique. *Modern English Teacher*. **22**(1), pp.12- 16.

Weigle, S. C. 1998. Using FACETS to model rater training effects. *Language Testing*. **15**, pp.263–287.

Wesche, M., and Paribakht, T. S. 1996. Assessing Second Language Vocabulary Knowledge: Depth vs. Breadth. *Canadian Modern Language Review*. **53**, pp.13-39.

Williams, M. and R.L. Burden. 1997. *Psychology for Language Teachers: A Social Constructivist Approach.* New York: Cambridge University Press.

Yamashita, J. and Jiang, N. 2010. L1 Influence on the Acquisition of L2 Collocations: Japanese ESL Users and EFL Learners Acquiring English Collocations. *TESOL Quarterly*. **44**, p.647668.

Yanagisawa, A., and Webb, S. 2021. To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*. **71**(2), pp. 487-536.

Yanagisawa, A., and Webb, S. 2022. Involvement load hypothesis plus: Creating an improved predictive model of incidental vocabulary learning. *Studies in Second Language Acquisition*. **44**(5), pp. 1279-1308.

Yeh, Y. and Wang, C. 2003. Effects of Multimedia Vocabulary Annotations and Learning Styles on Vocabulary Learning. *CALICO Journal*. **21**(1), pp.131–144.

Yoshimura, C. G. 2017. Teaching Communication and Conflict as a Game. *Communication Teacher*. 31(4), pp. 231-238.

Zhang, P. and Graham, S. 2020. Learning vocabulary through listening: the role of vocabulary knowledge and listening proficiency. *Language Learning*. **70**(4), pp.1017-1053. Available at https://centaur.reading.ac.uk/88908/

Zhang, Y; Lin, C-H; Zhang, D. 2017. Motivation, strategy, and English as a foreign language vocabulary learning: A structural equation modelling study. *British Journal of Educational Psychology*. pp.1-39. [online] Available at http://hdl.handle.net/10871/28123

Zhang S. and Zhang X. 2020. The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research.* **26**(4,. pp.1–30.

Zhou, S. 2010. Comparing Receptive and Productive Academic Vocabulary Knowledge of Chinese EFL Learners. *Asian Social Science*. **6**(10), pp.14-19.

Zimmerman, B. J. and Kitsantas, A. 1999. Acquiring Writing Revision Skill Shifting from Process to Outcome Self-Regulatory Goals. *Journal of Educational Psychology*. **91**(2), pp.241–250.

Zou, D. 2017. Vocabulary Acquisition Through Cloze Exercises, Sentence-writing And Composition- writing: Extending the Evaluation Component of the Involvement Load Hypothesis. *Language Teaching Research*. **21**(1), pp.54-75.

Zou, D. and Xie, H. 2018. Personalized Word-Learning Based on Technique Feature Analysis and Learning Analytics. *Journal of Educational Technology & Society*. **21**(2), pp.233-244.

# Appendices

## Appendix 1: The Analysis of Vocabulary Activities for Five Experiments

### TFA analysis of a reading only for Group 1 (Control)

This group received a total score of zero because learning is based on contents in the coursebook and there is no vocabulary activity included in the Unit 1 to be evaluated.

| Component | Criteria | Result | % |
|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 0 | **0** |
| | Does the activity motivate learning? | 0 | |
| | Do the learners select the words? | 0 | |
| **Noticing** | Does the activity focus attention on the target words? | 0 | **0** |
| | Does the activity raise awareness of new vocabulary learning? | 0 | |
| | Does the activity involve negotiation? | 0 | |
| **Retrieval** | Does the activity involve retrieval of the word? | 0 | **0** |
| | Is it productive retrieval? | 0 | |
| | Is it recall? | 0 | |
| | Are there multiple retrievals of each word? | 0 | |
| | Is there spacing between retrieval? | 0 | |
| **Generative use** | Does the activity involve generative use? | 0 | **0** |
| | Is it productive? | 0 | |
| | Is there a marked change that involves the use of other words? | 0 | |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 0 | **0** |
| | Does the activity involve instantiation? | 0 | |
| | Does the activity involve imagination? | 0 | |
| | Does the activity avoid interference? | 0 | |

*Note: The satisfy level of 'high support' of each component is equal to or greater than 66.68%*

**TFA analysis of a reading plus fill-in activity for Group 2 (*Motivation*)**

| Component | Criteria | Result | % |
|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 1 | |
| | Does the activity motivate learning? | 1 | **66.68** |
| | Do the learners select the words? | 0 | |
| **Noticing** | Does the activity focus attention on the target words? | 1 | |
| | Does the activity raise awareness of new vocabulary learning? | 0 | 33.34 |
| | Does the activity involve negotiation? | 0 | |
| **Retrieval** | Does the activity involve retrieval of the word? | 1 | |
| | Is it productive retrieval? | 0 | 40 |
| | Is it recall? | 1 | |
| | Are there multiple retrievals of each word? | 0 | |
| | Is there spacing between retrieval? | 0 | |
| **Generative use** | Does the activity involve generative use? | 0 | |
| | Is it productive? | 0 | 0 |
| | Is there a marked change that involves the use of other words? | 0 | |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 0 | |
| | Does the activity involve instantiation? | 1 | 50 |
| | Does the activity involve imagination? | 0 | |
| | Does the activity avoid interference? | 1 | |

*Note: The satisfy level of ‘high support’ of each component is equal to or greater than 66.68%*

**TFA analysis of a reading with L2 glosses for Group 3 (*Noticing*)**

| Component | Criteria | Result | % |
|---|---|---|---|
| **Motivation** | Is there a clear vocabulary learning goal? | 1 | 33.34 |
| | Does the activity motivate learning? | 0 | |
| | Do the learners select the words? | 0 | |
| **Noticing** | Does the activity focus attention on the target words? | 1 | **100** |
| | Does the activity raise awareness of new vocabulary learning? | 1 | |
| | Does the activity involve negotiation? | 1 | |
| **Retrieval** | Does the activity involve retrieval of the word? | 1 | 40 |
| | Is it productive retrieval? | 0 | |
| | Is it recall? | 0 | |
| | Are there multiple retrievals of each word? | 1 | |
| | Is there spacing between retrieval? | 0 | |
| **Generative use** | Does the activity involve generative use? | 0 | 0 |
| | Is it productive? | 0 | |
| | Is there a marked change that involves the use of other words? | 0 | |
| **Retention** | Does the activity ensure successful linking of form and meaning? | 0 | 25 |
| | Does the activity involve instantiation? | 1 | |
| | Does the activity involve imagination? | 0 | |
| | Does the activity avoid interference? | 0 | |

*Note: The satisfy level of 'high support' of each component is equal to or greater than 66.68%*

**TFA analysis of reading and identifying word parts of Group 4 (*Retrieval*)**

| Component | Criteria | Result | % |
|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | 1 | 33.33 |
| | Does the activity motivate learning? | 0 | |
| | Do the learners select the words? | 0 | |
| ***Noticing*** | Does the activity focus attention on the target words? | 0 | 33.33 |
| | Does the activity raise awareness of new vocabulary learning? | 1 | |
| | Does the activity involve negotiation? | 0 | |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 1 | **80** |
| | Is it productive retrieval? | 0 | |
| | Is it recall? | 1 | |
| | Are there multiple retrievals of each word? | 1 | |
| | Is there spacing between retrieval? | 1 | |
| ***Generative use*** | Does the activity involve generative use? | 0 | 0 |
| | Is it productive? | 0 | |
| | Is there a marked change that involves the use of other words? | 0 | |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | 0 | 50 |
| | Does the activity involve instantiation? | 1 | |
| | Does the activity involve imagination? | 0 | |
| | Does the activity avoid interference? | 1 | |

*Note: The satisfy level of 'high support' of each component is equal to or greater than 66.68%*

**TFA analysis of reading and sentence writing of Group 5 (*Generative Use*)**

| Component | Criteria | Result | % |
|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | 1 | 33.33 |
| | Does the activity motivate learning? | 0 | |
| | Do the learners select the words? | 0 | |
| ***Noticing*** | Does the activity focus attention on the target words? | 0 | 33.33 |
| | Does the activity raise awareness of new vocabulary learning? | 1 | |
| | Does the activity involve negotiation? | 0 | |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 1 | 40 |
| | Is it productive retrieval? | 1 | |
| | Is it recall? | 0 | |
| | Are there multiple retrievals of each word? | 0 | |
| | Is there spacing between retrieval? | 0 | |
| ***Generative use*** | Does the activity involve generative use? | 1 | **100** |
| | Is it productive? | 1 | |
| | Is there a marked change that involves the use of other words? | 1 | |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | 0 | 50 |
| | Does the activity involve instantiation? | 1 | |
| | Does the activity involve imagination? | 0 | |
| | Does the activity avoid interference? | 1 | |

*Note: The satisfy level of 'high support' of each component is equal to or greater than 66.68%*

**TFA analysis of reading, wordcards and sentence writing of Group 6 (*All TFA Components*)**

| Component | Criteria | Result | % |
|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | 1 | **66.68** |
| | Does the activity motivate learning? | 1 | |
| | Do the learners select the words? | 0 | |
| ***Noticing*** | Does the activity focus attention on the target words? | 1 | **66.68** |
| | Does the activity raise awareness of new vocabulary learning? | 1 | |
| | Does the activity involve negotiation? | 0 | |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 1 | **100** |
| | Is it productive retrieval? | 1 | |
| | Is it recall? | 1 | |
| | Are there multiple retrievals of each word? | 1 | |
| | Is there spacing between retrieval? | 1 | |
| ***Generative use*** | Does the activity involve generative use? | 1 | **100** |
| | Is it productive? | 1 | |
| | Is there a marked change that involves the use of other words? | 1 | |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | 1 | **75** |
| | Does the activity involve instantiation? | 1 | |
| | Does the activity involve imagination? | 0 | |
| | Does the activity avoid interference? | 1 | |

*Note: The satisfy level of ‘high support’ of each component is equal to or greater than 66.68%*

## Appendix 2: Lesson Plans

*Please note that there are five activities (five experiments from Groups 2-6). Vocabulary tasks in each activity are highlighted in gray.*

### (Group 1: Control)

**TFA scores**

| **Controlled Group:** *Reading only (learning is based on Unit 1)* | | | | | | |
|---|---|---|---|---|---|---|
| ***Component*** | motivation (3) | noticing (3) | retrieval (5) | generative use (3) | retention (4) | Total 18 |
| ***Interpretation*** | **Low** | **Low** | **Low** | **Low** | **Low** | |
| ***Received score*** | **(0)** | **(0)** | **(0)** | **(0)** | **(0)** | 0 |

**Lesson Plan**

| | | | |
|---|---|---|---|
| **TFA** | (Control group: no vocabulary activity) | **Group** | 1 |
| **Course** | TU 105 | | |
| **Unit** | 1. Culture and Society | | |
| **Duration** | 90 minutes *(+ 90 minutes for an immediate F-MRt)* | | |
| **Materials** | TU105 Coursebook | | |
| Objectives: Students should be able to<br>1.) better understand cultural differences<br>2.) differentiate between various aspects of culture<br>3.) form open ended Wh-questions<br>4.) skim an essay for general information | | | |
| **Activities** | **Description of activity** | **Time** | |
| Warm-up/ Lead-in activity | 1.) have students discuss about the meanings of culture and society.<br>2.) have students take a look at the provided definitions on page 2 in the coursebook.<br>3.) have students take a short quiz (6 questions) to test knowledge of the world. | 10 | |
| Language function: Wh-questions | 4.) have students do an activity in section A on page 4 by answering the questions about Thai government poster. | 60 | |

| | | |
|---|---|---|
| and vocabulary | 5.) have students read information in section B (Asking Questions) on page 5 in the coursebook before doing Activity 1 and Activity 2.<br>6.) have students work in pairs to form questions for the missing information and use the questions to ask each other until they get all missing information (student A turns to page 8 and student B turns to page 16). | |
| Reading: Skimming | 9.) have students read information about scanning and skimming in the coursebook on pages 6 and 7.<br>10.) have them look at some examples of how to skim a reading passage by using the skimming activity on page 7. Then, have them go over the answers together. | 20 |

**(Group 2: *Motivation*)**

**TFA scores**

| **Reading plus fill-in** | | | | | | |
|---|---|---|---|---|---|---|
| ***Component*** | **motivation (3)** | noticing (3) | retrieval (5) | generative use (3) | retention (4) | Total 18 |
| ***Interpretation*** | **High** | Low | Low | Low | Low | |
| ***Received score*** | **(2)** | (1) | (2) | (0) | (2) | 7 |

**Lesson Plan**

| | | | |
|---|---|---|---|
| **TFA** | Motivation | **Group** | 2 |
| **Course** | TU 105 | | |
| **Unit** | 1. Culture and Society | | |
| **Duration** | 90 minutes *(+ 90 minutes for an immediate F-MRt)* | | |
| **Materials** | TU105 Coursebook, supplementary handouts, and a Kahoot game | | |
| Objectives: At the end of this class, students should be able to<br>1.) better understand cultural differences<br>2.) form open ended Wh-questions<br>3.) form questions to find missing words in a reading passage<br>4.) skim an essay for general information | | | |
| **Activities** | **Description of activity** | | **Time** |
| Warm-up/ Lead-in activity | 1.) have students discuss about the meanings of culture and society.<br>2.) have students take a look at the provided definitions on page 2 in the coursebook.<br>3.) have students take a short quiz (6 questions) to test knowledge of the world. | | 10 |
| Language function: Wh-questions and vocabulary | 4.) have students read information in section B (Asking Questions) before doing Activity 1 and Activity 2 on page 5 in the coursebook.<br>5.) have students take a look at a reading passage in a supplementary handout to notice missing information in blanks. They will be reminded that these missing words will help them comprehend the reading passage. | | 60 |

| | | |
|---|---|---|
| | 6.) They have to brainstorm ideas about open-ended questions that can be used to find the missing information for the first three items of each version together. Then, have students write at least three Wh-questions for the missing information of Item 4, Item 5, and Item 6 in Activity 1 in the handout.<br>7.) have students work in pairs by doing vocabulary Activity 2 in the handout (handout A for student A and handout B for student B). Student A has to find the missing words by using the questions from Activity 1 to ask their partner (Student B). However, students will be remined that they do not have to write the words in the blanks as a complete version of the passage will be presented to the class later. Also, they could not use the same question for the next missing words.<br>8.) have students take turn until they get all missing information.<br>9.) have students play a Kahoot game by choosing the missing words from the previous activity to match with sentences from the book that will be presented on a computer screen during the game. The winner will get a reward.<br>For teacher:<br>https://create.kahoot.it/details/a9d97248-9fe4-40ee-94c0-c4dd8145f27f<br>For student: https://kahoot.it/ (PIN: 3227828) | |
| Reading: Skimming | 9.) have students read information about scanning and skimming in the coursebook on pages 6 and 7.<br>10.) have them look at some examples of how to skim a reading passage by using the skimming activity on page 7. Then, have them go over the answers together. | 20 |

### (Group 3: *Noticing*)

**TFA scores**

| **Reading with glosses** | | | | | | |
|---|---|---|---|---|---|---|
| ***Component*** | motivation (3) | **noticing (3)** | retrieval (5) | generative use (3) | retention (4) | Total 18 |
| ***Interpretation*** | Low | **High** | Low | Low | Low | |
| ***Received score*** | (1) | **(3)** | (2) | (0) | (1) | 7 |

**Lesson Plan**

| | | | |
|---|---|---|---|
| **TFA** | Noticing | **Group** | 3 |
| **Course** | TU 105 | | |
| **Unit** | 1. Culture and Society | | |
| **Duration** | 90 minutes *(+ 90 minutes for an immediate F-MRt)* | | |
| **Materials** | TU105 Coursebook, a supplementary handout | | |
| Objectives: Students should be able to<br>1.) better understand cultural differences<br>2.) form open ended Wh-questions<br>3.) comprehend meanings of the target words<br>4.) skim an essay for general information | | | |
| **Activities** | **Description of activity** | **Time** | |
| Warm-up/ Lead-in activity | 1.) have students discuss about the meanings of culture and society.<br>2.) have students take a look at the provided definitions on page 2 in the coursebook.<br>3.) have students take a short quiz (6 questions) to test knowledge of the world. | 10 | |
| Language function: Wh-questions and vocabulary | 4.) have students read information in section B (Asking Questions) before doing Activity 1 and Activity 2 in the coursebook.<br>5.) have students take a look at a list of ten target words with L2 definitions in a supplementary handout. They will be remined that this is a list of new target words in this unit activity that helps them comprehend a reading passage of Unit 1. | 60 | |

| | | |
|---|---|---|
| | 6.) have students find sentences that contain the target words in the provided passage and underline them.<br>7.) have students work in pairs by doing Activity 1 in the handout to discuss and match a correct synonym with each target word. | |
| Reading: Skimming | 8.) have students read information about scanning and skimming in the coursebook on pages 6 and 7.<br>9.) have students practise skimming the passage 'The History of Pad Thai' in the handout to get answers for the skimming activity on page 7 in the coursebook. Then, have them go over the answers together. | 20 |

### (Group 4: *Retrieval*)

**TFA scores**

| **Reading and word parts** | | | | | | |
|---|---|---|---|---|---|---|
| ***Component*** | motivation (3) | noticing (3) | **retrieval (5)** | generative use (3) | retention (4) | Total 18 |
| ***Interpretation*** | Low | Low | **High** | Low | Low | |
| ***Received score*** | (1) | (1) | **(4)** | (0) | (2) | 8 |

**Lesson Plan**

| | | | |
|---|---|---|---|
| **TFA** | Retrieval | **Group** | 4 |
| **Course** | TU 105 | | |
| **Unit** | 1. Culture and Society | | |
| **Duration** | 90 minutes *(+ 90 minutes for an immediate F-MRt)* | | |
| **Materials** | TU105 Coursebook, a supplementary handout | | |
| Objectives: Students should be able to<br>1.) better understand cultural differences<br>2.) identify part of speech (POS) and suffixes<br>3.) skim an essay for general information<br>4.) form open ended Wh-questions | | | |
| **Activities** | **Description of activity** | **Time** | |
| Warm-up/ Lead-in activity | 1.) have students discuss about the meanings of culture and society.<br>2.) have students take a look at the provided definitions on page 2 in the coursebook.<br>3.) have students take a short quiz (6 questions) to test knowledge of the world. | 10 | |
| Language function: Wh-questions and vocabulary | 4.) have students take a look at a supplementary handout to learn about word parts.<br>5.) have students identify POS and suffixes of words that include both ten target words and several words taken from the reading passage in Activity 1.<br>6.) have students read information about forming questions in section B (Asking Questions) before doing activities (Activity 1 and Activity 2) in the coursebook. | 40 | |

| | | |
|---|---|---|
| Reading:<br>Skimming | 7.) have students read information about scanning and skimming in the coursebook on pages 6 and 7.<br>8.) have students practise skimming the passage ‘The History of Pad Thai’ in a provided handout to get answers for the skimming activity on page 7 in the coursebook. Then, have them go over the answers together. | 20 |
| Wrap-up:<br>Vocabulary | 9.) have students pay attention to a list of words (target words and extra words) on a computer screen and try to catch the words that will appear quickly (2 seconds, each). There will be both the list of words they have encountered and a new list of words taken from the passage.<br>10.) have them put a mark on a happy face (J) icon of an item in Activity 2 if they see that item appears on the screen.<br>11.) have students recall the items by saying them out loud without looking at the handout. | 20 |

### (Group 5: *Generative Use*)

**TFA scores**

| **Reading, word parts, and writing** | | | | | | |
|---|---|---|---|---|---|---|
| ***Component*** | motivation (3) | noticing (3) | retrieval (5) | **generative use (3)** | retention (4) | Total 18 |
| ***Interpretation*** | Low | Low | Low | **High** | Low | |
| ***Received score*** | (1) | (1) | (2) | **(3)** | (2) | 9 |

**Lesson Plan**

| | | | |
|---|---|---|---|
| **TFA** | Generative Use | **Group** | 5 |
| **Course** | TU 105 | | |
| **Unit** | 1. Culture and Society | | |
| **Duration** | 90 minutes *(+ 90 minutes for an immediate F-MRt)* | | |
| **Materials** | TU105 Coursebook, a supplementary handout | | |
| Objectives: Students should be able to<br>1.) better understand cultural differences<br>2.) identify part of speech (POS) and suffixes<br>3.) skim an essay for general information<br>4.) form open ended Wh-questions | | | |
| **Activities** | **Description of activity** | **Time** | |
| Warm-up/ Lead-in activity | 1.) have students discuss about the meanings of culture and society.<br>2.) have students take a look at the provided definitions on page 2 in the coursebook.<br>3.) have students take a short quiz (6 questions) to test knowledge of the world. | 10 | |
| Language function: Wh-questions and vocabulary | 4.) have students read information about forming questions in section B (Asking Questions) before doing activities (Activity 1 and Activity 2) in the coursebook.<br>5.) have students take a look at a supplementary handout to learn about word parts.<br>6.) have students find the words with suffixes in a provided reading passage and add the correct suffixes to the words taken from the passage in | 60 | |

| | | |
|---|---|---|
| | Activity 1 in a supplementary handout. Then, have them identify POS of each word.<br>7.) have students form questions and provide answers for each question by using the words in the provided handout. These words must be included either in questions or answers. | |
| Reading: skimming | 8.) have students read information about scanning and skimming in the coursebook on pages 6 and 7.<br>9.) have students practise skimming the passage 'The History of Pad Thai' in the handout to get answers for the skimming activity on page 7 in the coursebook. Then, have them go over the answers together. | 20 |

### (Group 6: *All TFA Components*)

**TFA scores**

| **Reading, wordcards and writing** | | | | | | |
|---|---|---|---|---|---|---|
| ***Component*** | **motivation (3)** | **noticing (3)** | **retrieval (5)** | **generative use (3)** | **retention (4)** | Total 18 |
| ***Interpretation*** | **High** | **High** | **High** | **High** | **High** | |
| ***Received score*** | **(2)** | **(2)** | **(5)** | **(3)** | **(3)** | 15 |

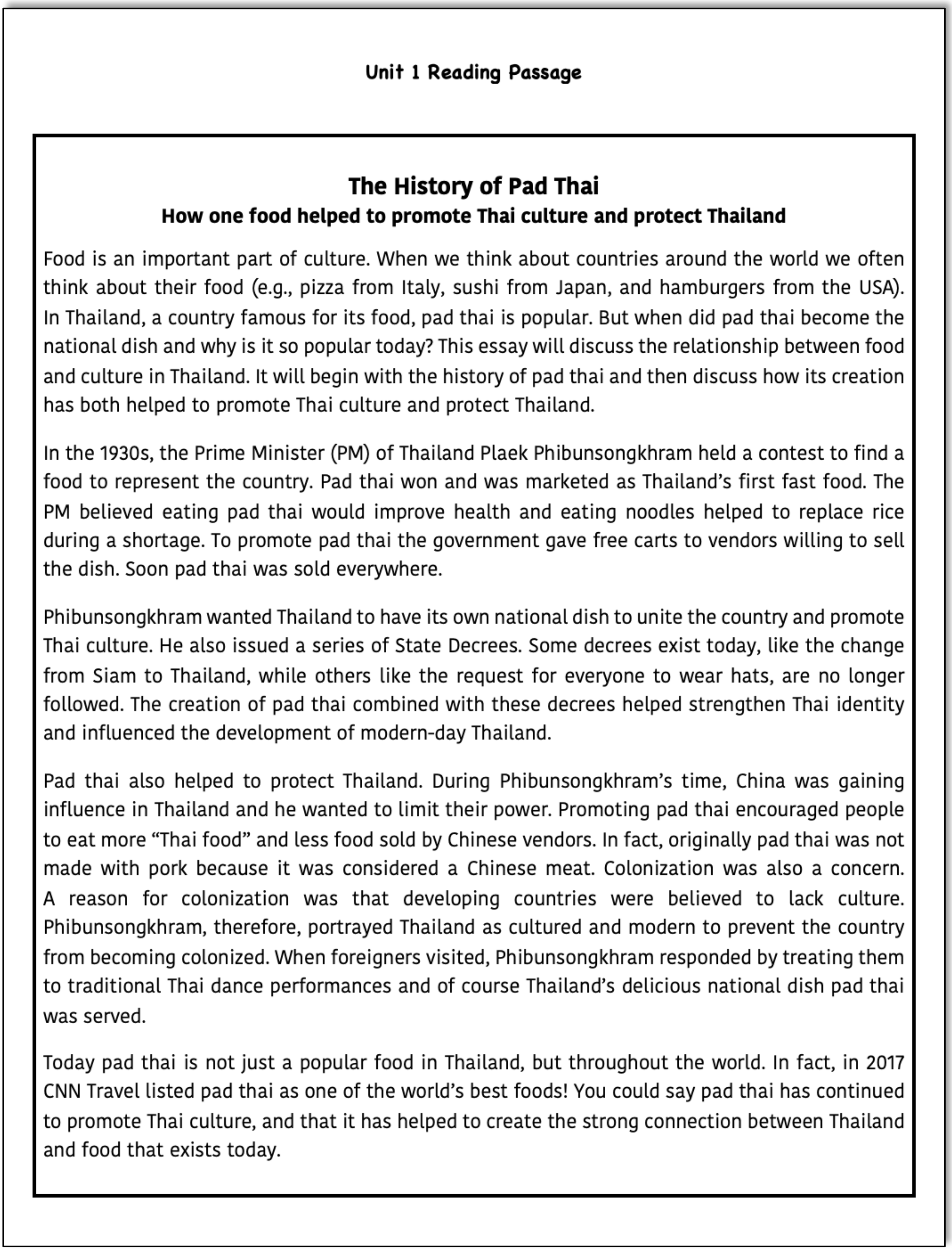
**Lesson Plan**

| | | | |
|---|---|---|---|
| **TFA** | All components | **Group** | 6 |
| **Course** | TU 105 | | |
| **Unit** | 1. Culture and Society | | |
| **Duration** | 90 minutes *(+ 90 minutes for an immediate F-MRt)* | | |
| **Materials** | TU105 Coursebook, wordcards, a supplementary handout, and a Kahoot game | | |
| Objectives: Students should be able to<br>1.) better understand cultural differences<br>2.) form open ended Wh-questions<br>3.) comprehend meanings of the target words<br>4.) skim an essay for important information | | | |
| **Activities** | **Description of activity** | **Time** | |
| Warm-up/ Lead-in activity | 1.) have students discuss about the meanings of culture and society.<br>2.) have students take a look at the provided definitions on page 2 in the coursebook.<br>3.) have students take a short quiz (6 questions) to test knowledge of the world. | 10 | |
| Reading: skimming and vocabulary | 4.) have students read information about scanning and skimming in the coursebook on pages 6 and 7.<br>5.) have students learn ten target vocabulary (taken from the reading passage) from wordcards with L1 definitions. They will be reminded that these words will help them comprehend the reading passage.<br>6.) have students recall their memory through a Kahoot game by choosing the correct word they have | 35 | |

| | | |
|---|---|---|
| | just learned to match with its correct definition presented in the game.<br>7.) have students practise skimming the passage 'The History of Pad Thai' in a provided handout to get answers for the skimming activity on page 7 in the coursebook. Then, have them go over the answers together. | |
| Language function: Wh-questions and vocabulary | 8.) have students read through the information about forming questions in section B (Asking Questions) in the coursebook before doing the following activities (Activity 1 and Activity 2) on page 5.<br>9.) have students write questions and provide answers for each question in Activity 1 in a supplementary handout by using the target words they have learned. These words must be included either in questions or answers.<br>10.) have some students volunteer to present their work. | 45 |

## Appendix 3: Supplementary Handouts for Intervention Groups

**Unit 1 Reading Passage** *(used or adapted to use with all groups):* no other supplementary handout



**Unit 1 Reading Passage**

### The History of Pad Thai

**How one food helped to promote Thai culture and protect Thailand**

Food is an important part of culture. When we think about countries around the world we often think about their food (e.g., pizza from Italy, sushi from Japan, and hamburgers from the USA). In Thailand, a country famous for its food, pad thai is popular. But when did pad thai become the national dish and why is it so popular today? This essay will discuss the relationship between food and culture in Thailand. It will begin with the history of pad thai and then discuss how its creation has both helped to promote Thai culture and protect Thailand.

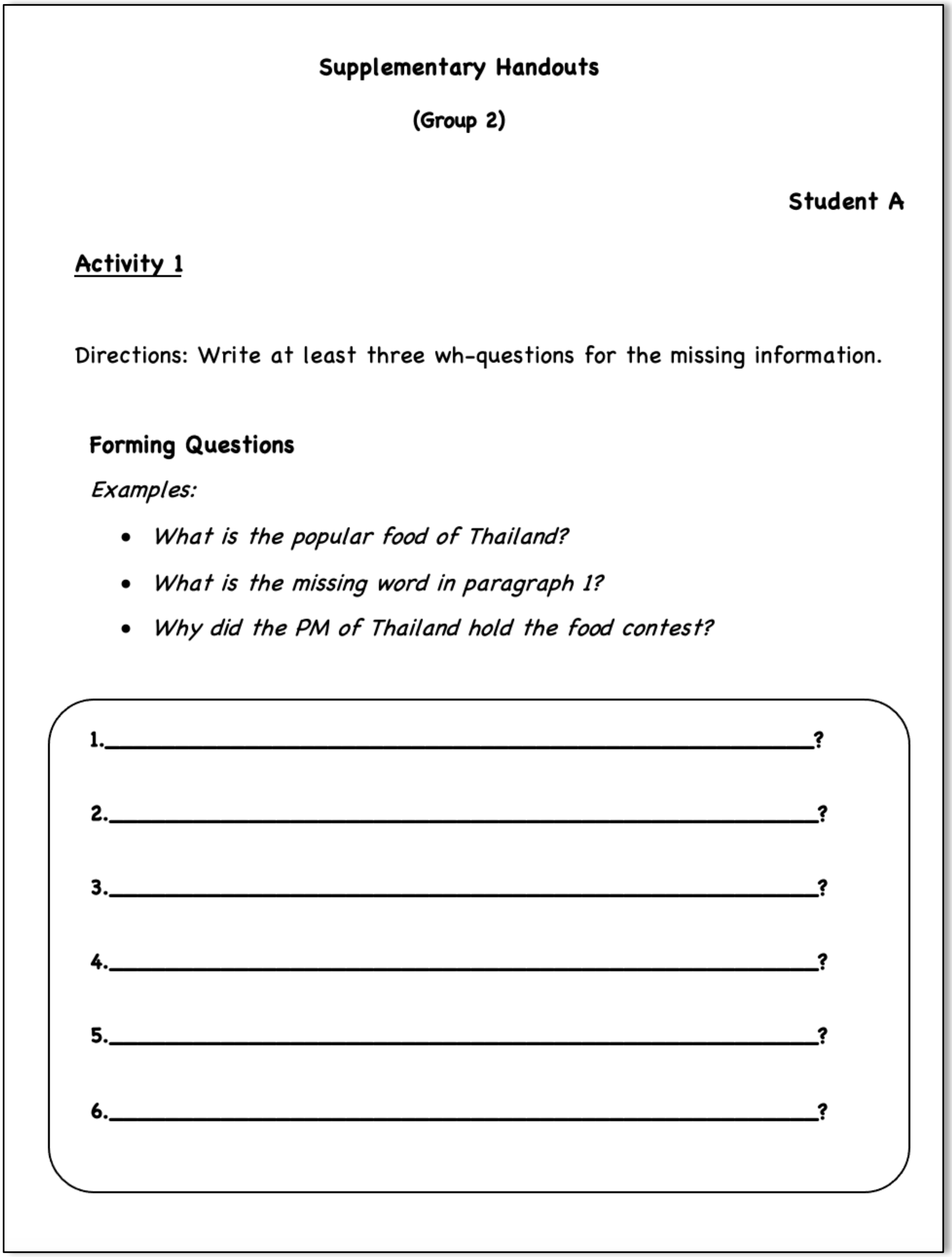
In the 1930s, the Prime Minister (PM) of Thailand Plaek Phibunsongkhram held a contest to find a food to represent the country. Pad thai won and was marketed as Thailand's first fast food. The PM believed eating pad thai would improve health and eating noodles helped to replace rice during a shortage. To promote pad thai the government gave free carts to vendors willing to sell the dish. Soon pad thai was sold everywhere.

Phibunsongkhram wanted Thailand to have its own national dish to unite the country and promote Thai culture. He also issued a series of State Decrees. Some decrees exist today, like the change from Siam to Thailand, while others like the request for everyone to wear hats, are no longer followed. The creation of pad thai combined with these decrees helped strengthen Thai identity and influenced the development of modern-day Thailand.

Pad thai also helped to protect Thailand. During Phibunsongkhram's time, China was gaining influence in Thailand and he wanted to limit their power. Promoting pad thai encouraged people to eat more "Thai food" and less food sold by Chinese vendors. In fact, originally pad thai was not made with pork because it was considered a Chinese meat. Colonization was also a concern. A reason for colonization was that developing countries were believed to lack culture. Phibunsongkhram, therefore, portrayed Thailand as cultured and modern to prevent the country from becoming colonized. When foreigners visited, Phibunsongkhram responded by treating them to traditional Thai dance performances and of course Thailand's delicious national dish pad thai was served.

Today pad thai is not just a popular food in Thailand, but throughout the world. In fact, in 2017 CNN Travel listed pad thai as one of the world's best foods! You could say pad thai has continued to promote Thai culture, and that it has helped to create the strong connection between Thailand and food that exists today.

**Handouts for Group 2: reading plus fill-in** (TFA support for *Motivation*)



**Supplementary Handouts**

**(Group 2)**

**Student A**

**<u>Activity 1</u>**

Directions: Write at least three wh-questions for the missing information.

**Forming Questions**

*Examples:*

- *What is the popular food of Thailand?*
- *What is the missing word in paragraph 1?*
- *Why did the PM of Thailand hold the food contest?*

**1.**______________________________________________**?**

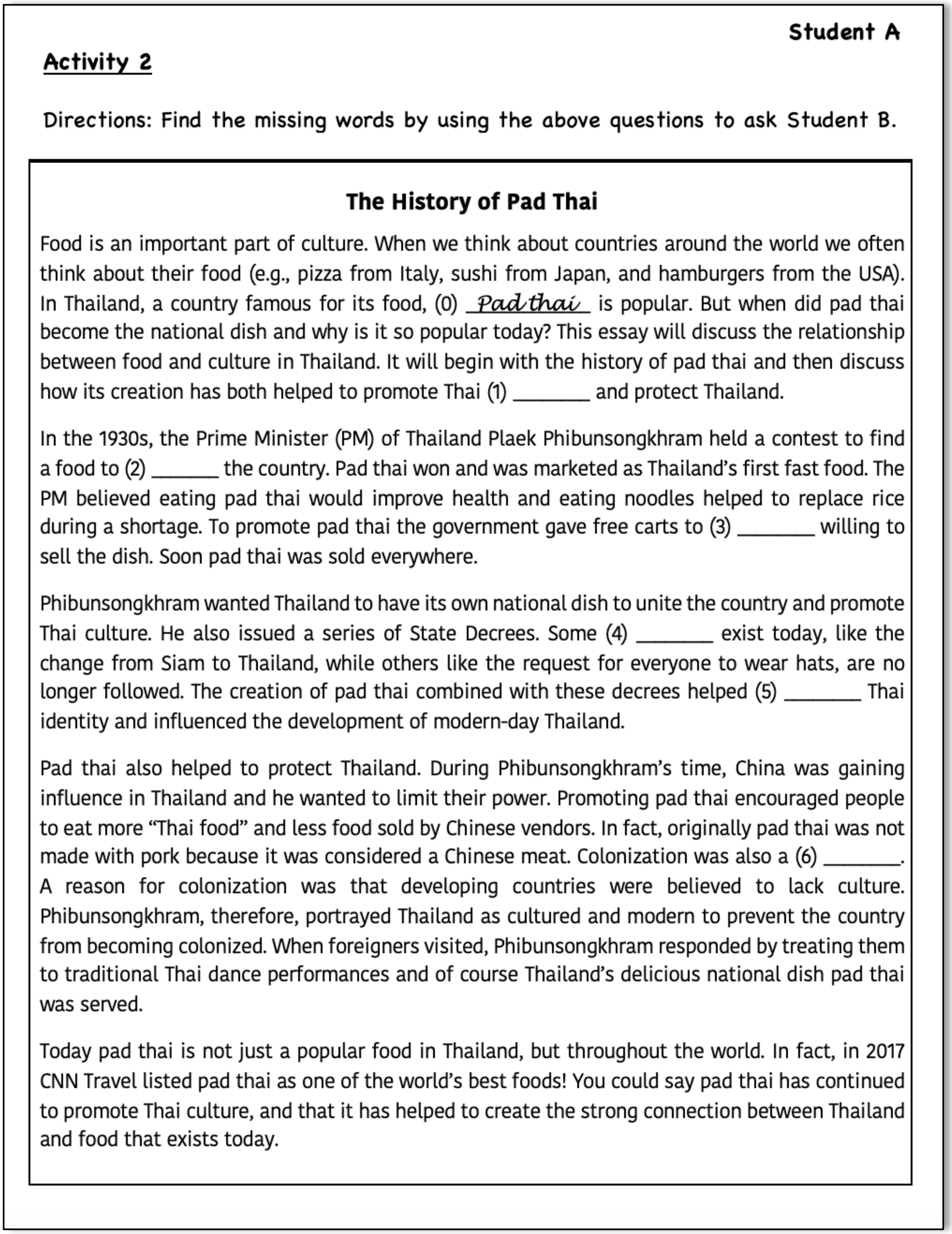
**2.**______________________________________________**?**

**3.**______________________________________________**?**

**4.**______________________________________________**?**

**5.**______________________________________________**?**

**6.**______________________________________________**?**



**Student A**

**Activity 2**

Directions: Find the missing words by using the above questions to ask Student B.

## The History of Pad Thai

Food is an important part of culture. When we think about countries around the world we often think about their food (e.g., pizza from Italy, sushi from Japan, and hamburgers from the USA). In Thailand, a country famous for its food, (0) _Pad thai_ is popular. But when did pad thai become the national dish and why is it so popular today? This essay will discuss the relationship between food and culture in Thailand. It will begin with the history of pad thai and then discuss how its creation has both helped to promote Thai (1) ________ and protect Thailand.

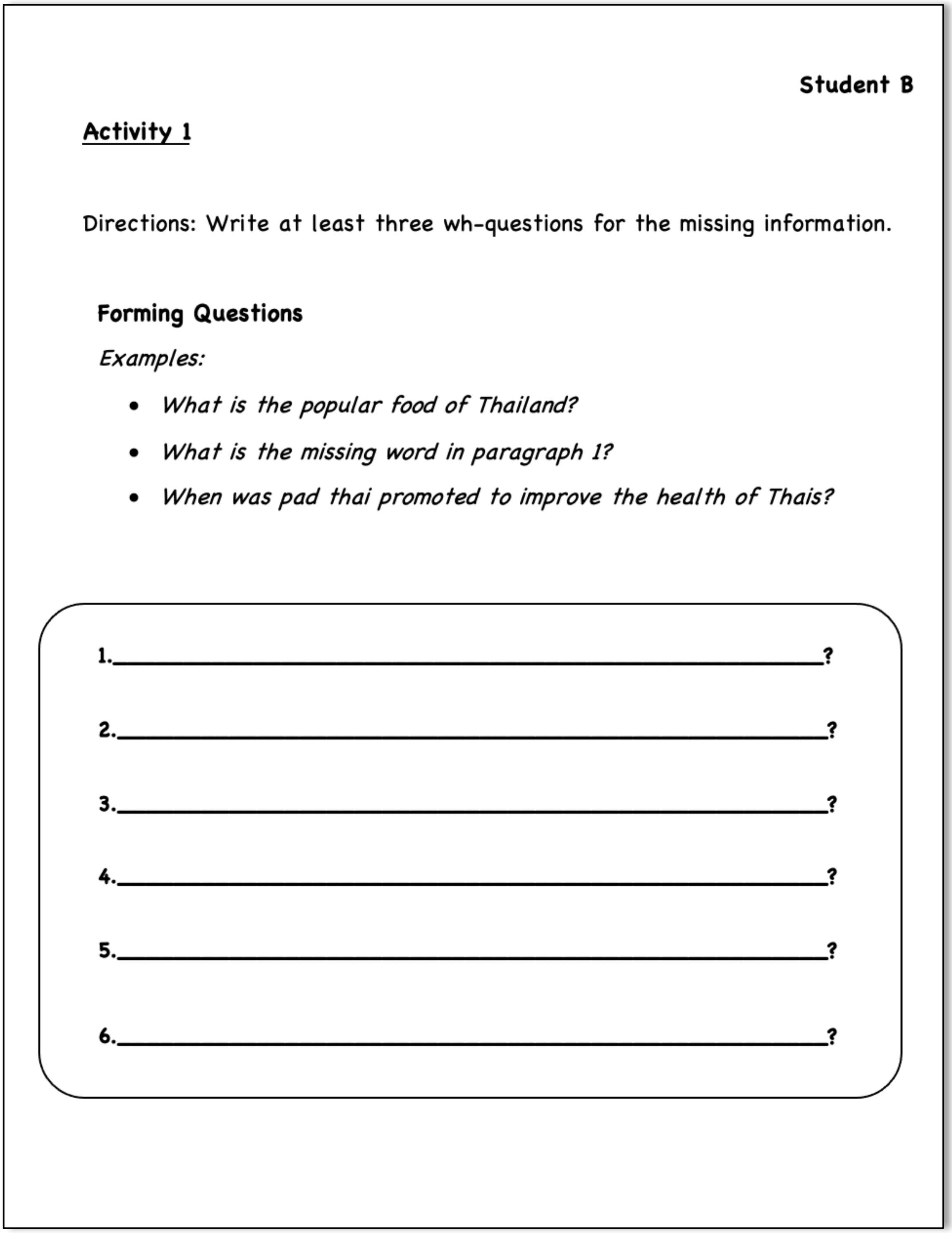
In the 1930s, the Prime Minister (PM) of Thailand Plaek Phibunsongkhram held a contest to find a food to (2) _______ the country. Pad thai won and was marketed as Thailand's first fast food. The PM believed eating pad thai would improve health and eating noodles helped to replace rice during a shortage. To promote pad thai the government gave free carts to (3) ________ willing to sell the dish. Soon pad thai was sold everywhere.

Phibunsongkhram wanted Thailand to have its own national dish to unite the country and promote Thai culture. He also issued a series of State Decrees. Some (4) ________ exist today, like the change from Siam to Thailand, while others like the request for everyone to wear hats, are no longer followed. The creation of pad thai combined with these decrees helped (5) ________ Thai identity and influenced the development of modern-day Thailand.

Pad thai also helped to protect Thailand. During Phibunsongkhram's time, China was gaining influence in Thailand and he wanted to limit their power. Promoting pad thai encouraged people to eat more "Thai food" and less food sold by Chinese vendors. In fact, originally pad thai was not made with pork because it was considered a Chinese meat. Colonization was also a (6) ________. A reason for colonization was that developing countries were believed to lack culture. Phibunsongkhram, therefore, portrayed Thailand as cultured and modern to prevent the country from becoming colonized. When foreigners visited, Phibunsongkhram responded by treating them to traditional Thai dance performances and of course Thailand's delicious national dish pad thai was served.

Today pad thai is not just a popular food in Thailand, but throughout the world. In fact, in 2017 CNN Travel listed pad thai as one of the world's best foods! You could say pad thai has continued to promote Thai culture, and that it has helped to create the strong connection between Thailand and food that exists today.

*Note: extra words are added to avoid noticing*



**Student B**

**Activity 1**

Directions: Write at least three wh-questions for the missing information.

**Forming Questions**

*Examples:*

- *What is the popular food of Thailand?*
- *What is the missing word in paragraph 1?*
- *When was pad thai promoted to improve the health of Thais?*

**1.**______________________________________________**?**

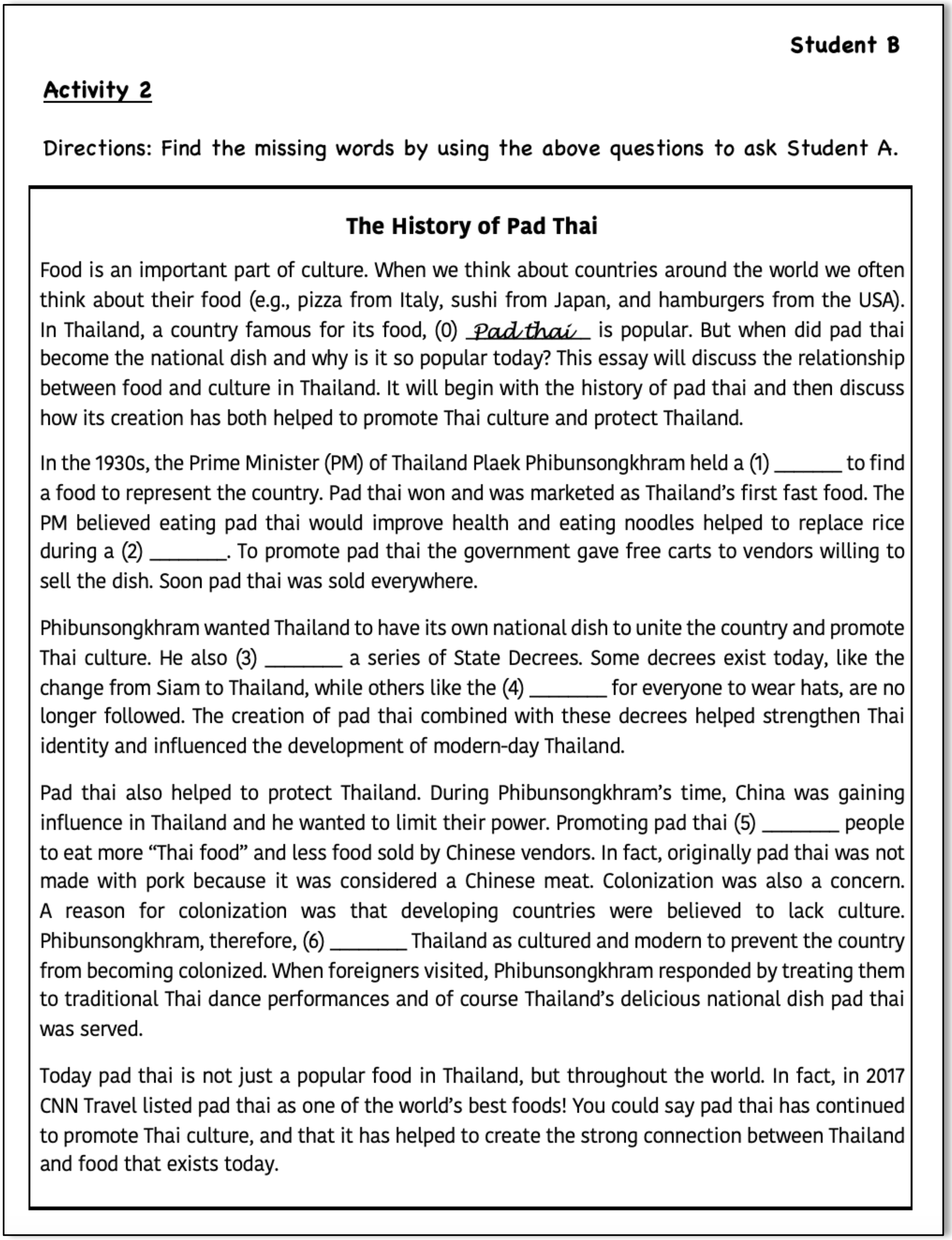
**2.**______________________________________________**?**

**3.**______________________________________________**?**

**4.**______________________________________________**?**

**5.**______________________________________________**?**

**6.**______________________________________________**?**



**Student B**

**Activity 2**

Directions: Find the missing words by using the above questions to ask Student A.

**The History of Pad Thai**

Food is an important part of culture. When we think about countries around the world we often think about their food (e.g., pizza from Italy, sushi from Japan, and hamburgers from the USA). In Thailand, a country famous for its food, (0) Pad thai is popular. But when did pad thai become the national dish and why is it so popular today? This essay will discuss the relationship between food and culture in Thailand. It will begin with the history of pad thai and then discuss how its creation has both helped to promote Thai culture and protect Thailand.

In the 1930s, the Prime Minister (PM) of Thailand Plaek Phibunsongkhram held a (1) _______ to find a food to represent the country. Pad thai won and was marketed as Thailand's first fast food. The PM believed eating pad thai would improve health and eating noodles helped to replace rice during a (2) _______. To promote pad thai the government gave free carts to vendors willing to sell the dish. Soon pad thai was sold everywhere.

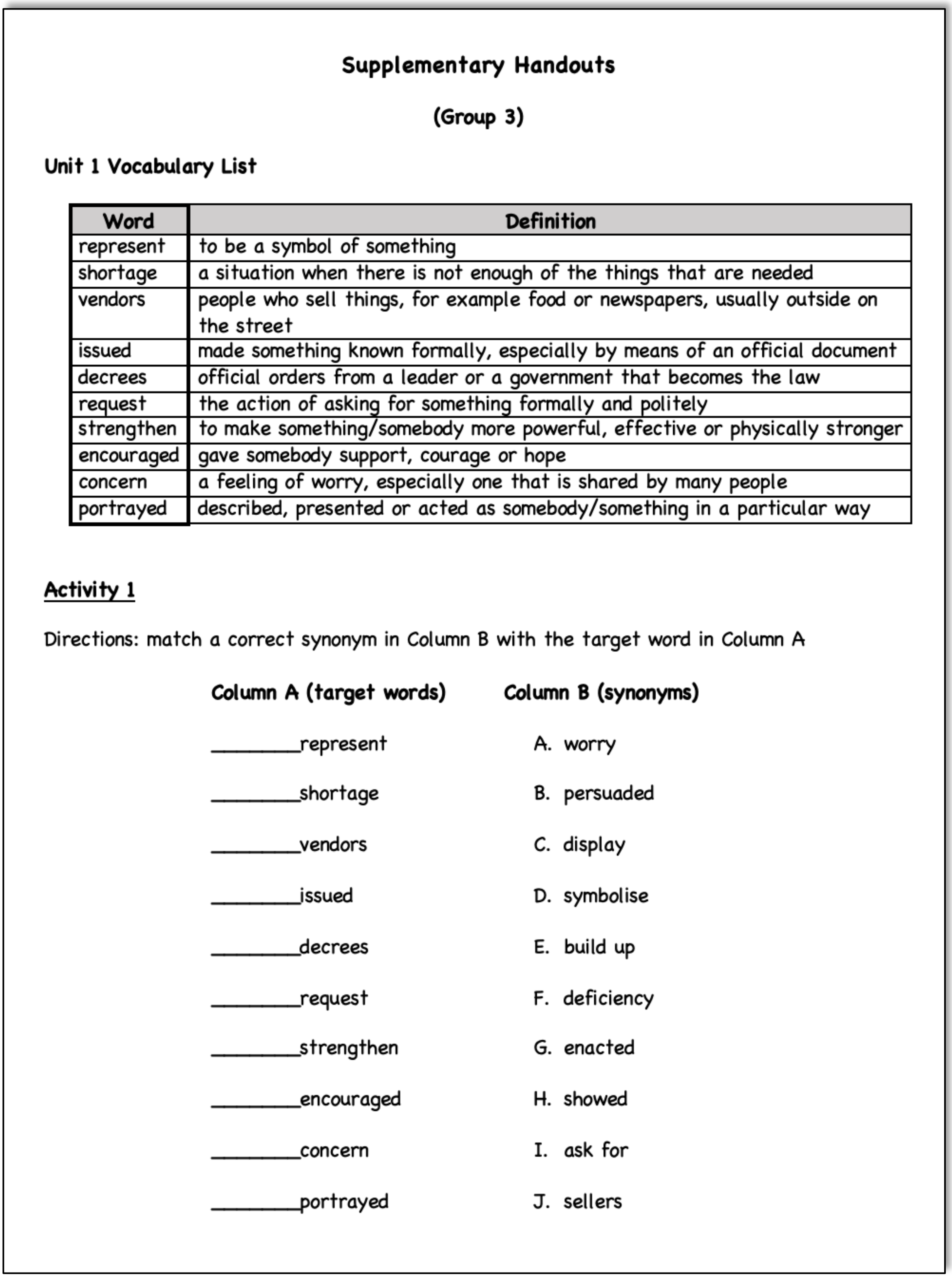
Phibunsongkhram wanted Thailand to have its own national dish to unite the country and promote Thai culture. He also (3) _______ a series of State Decrees. Some decrees exist today, like the change from Siam to Thailand, while others like the (4) _______ for everyone to wear hats, are no longer followed. The creation of pad thai combined with these decrees helped strengthen Thai identity and influenced the development of modern-day Thailand.

Pad thai also helped to protect Thailand. During Phibunsongkhram's time, China was gaining influence in Thailand and he wanted to limit their power. Promoting pad thai (5) _______ people to eat more "Thai food" and less food sold by Chinese vendors. In fact, originally pad thai was not made with pork because it was considered a Chinese meat. Colonization was also a concern. A reason for colonization was that developing countries were believed to lack culture. Phibunsongkhram, therefore, (6) _______ Thailand as cultured and modern to prevent the country from becoming colonized. When foreigners visited, Phibunsongkhram responded by treating them to traditional Thai dance performances and of course Thailand's delicious national dish pad thai was served.

Today pad thai is not just a popular food in Thailand, but throughout the world. In fact, in 2017 CNN Travel listed pad thai as one of the world's best foods! You could say pad thai has continued to promote Thai culture, and that it has helped to create the strong connection between Thailand and food that exists today.

*Note: extra words are added to avoid noticing*

**Handouts for Group 3: reading with glosses** (TFA support for *Noticing*)



## Supplementary Handouts

**(Group 3)**

**Unit 1 Vocabulary List**

| Word | Definition |
|---|---|
| represent | to be a symbol of something |
| shortage | a situation when there is not enough of the things that are needed |
| vendors | people who sell things, for example food or newspapers, usually outside on the street |
| issued | made something known formally, especially by means of an official document |
| decrees | official orders from a leader or a government that becomes the law |
| request | the action of asking for something formally and politely |
| strengthen | to make something/somebody more powerful, effective or physically stronger |
| encouraged | gave somebody support, courage or hope |
| concern | a feeling of worry, especially one that is shared by many people |
| portrayed | described, presented or acted as somebody/something in a particular way |

**Activity 1**

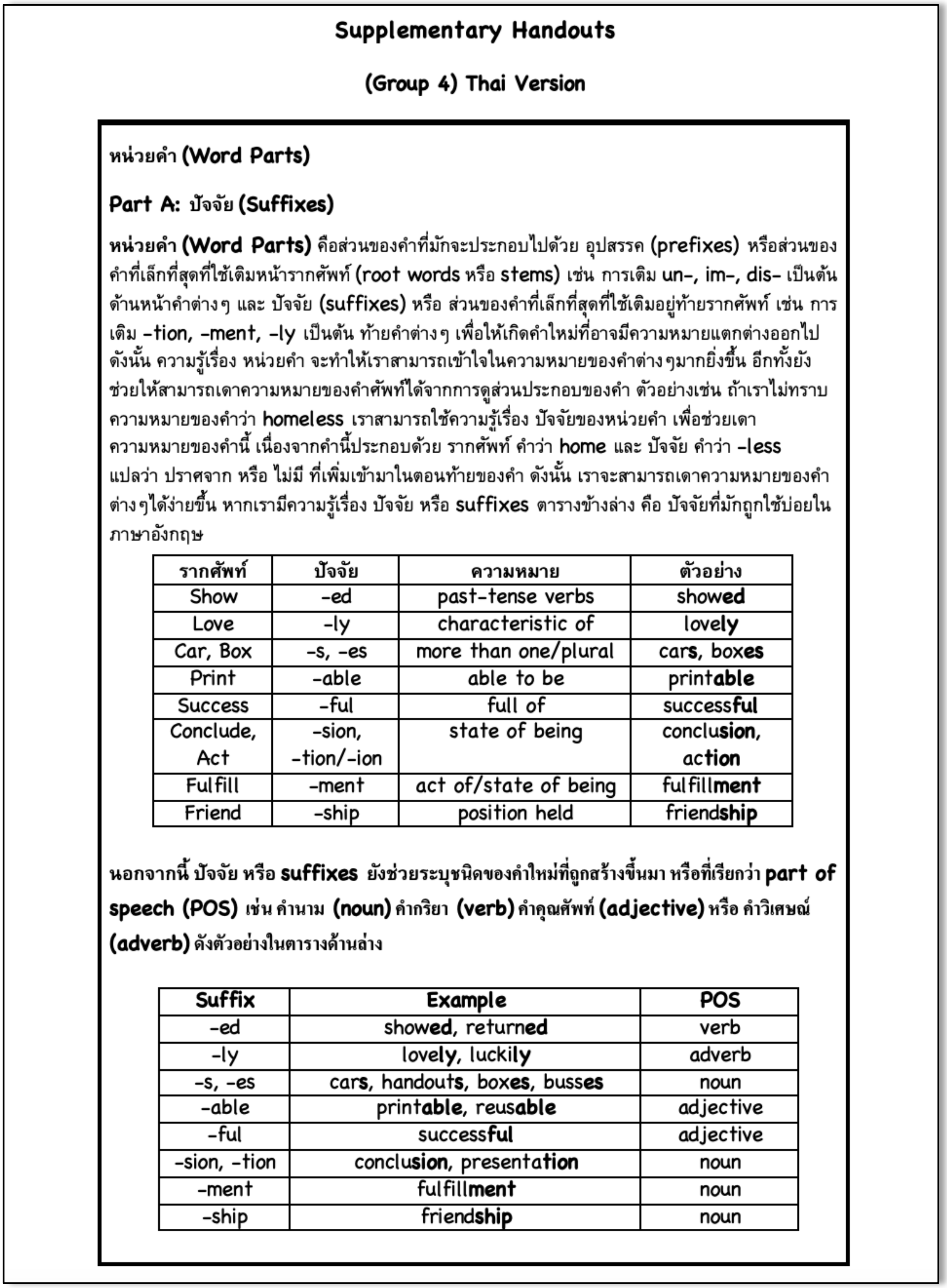
Directions: match a correct synonym in Column B with the target word in Column A

| Column A (target words) | Column B (synonyms) |
|---|---|
| ________represent | A. worry |
| ________shortage | B. persuaded |
| ________vendors | C. display |
| ________issued | D. symbolise |
| ________decrees | E. build up |
| ________request | F. deficiency |
| ________strengthen | G. enacted |
| ________encouraged | H. showed |
| ________concern | I. ask for |
| ________portrayed | J. sellers |

*Note: the words are listed in the order they appeared in the passage*

**Handouts for Group 4: reading and word parts** (TFA support for *Retrieval*)

## Thai Version



## Supplementary Handouts

### (Group 4) Thai Version

**หน่วยคำ (Word Parts)**

**Part A:** ปัจจัย **(Suffixes)**

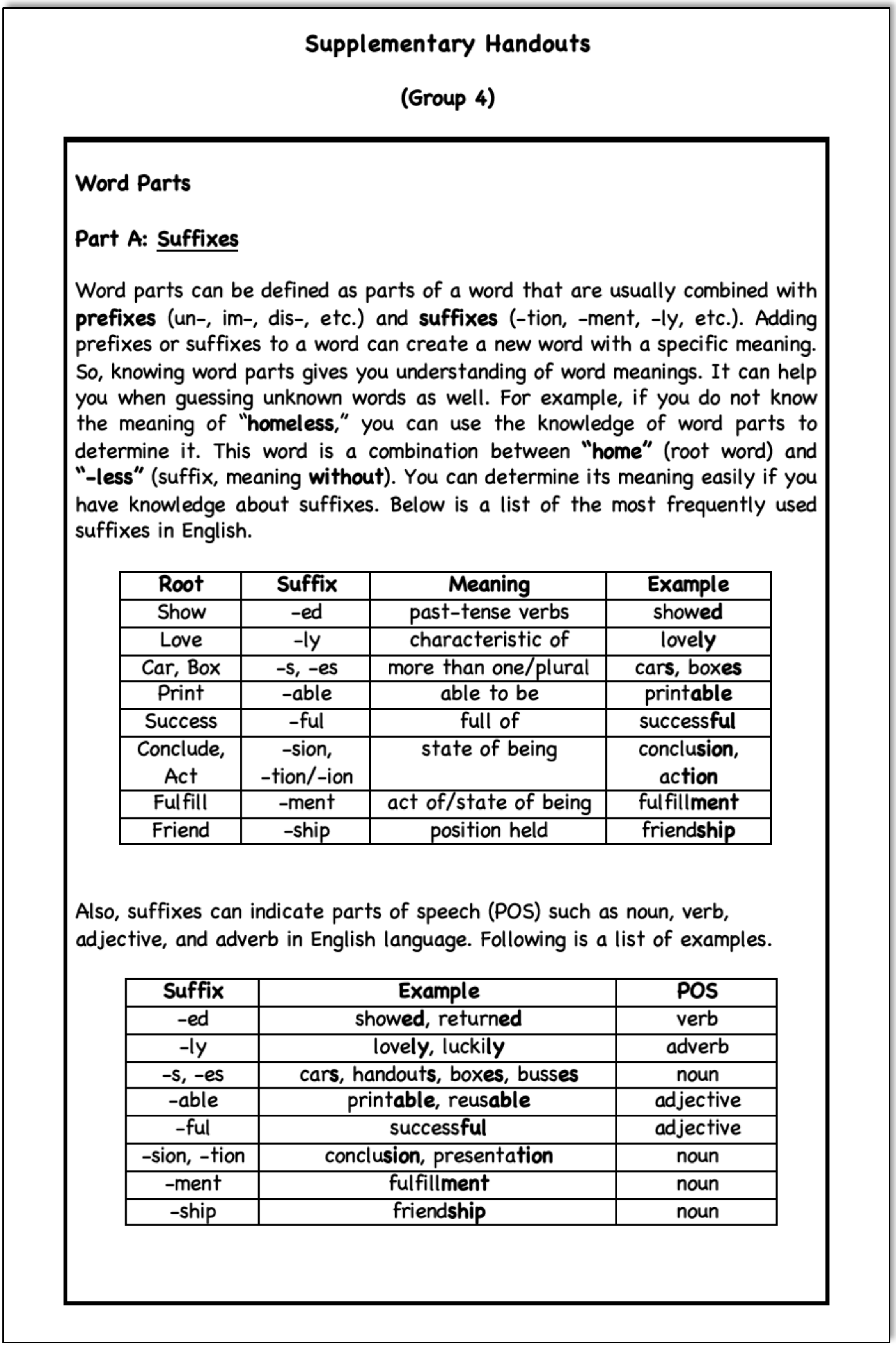
**หน่วยคำ (Word Parts)** คือส่วนของคำที่มักจะประกอบไปด้วย อุปสรรค (**prefixes**) หรือส่วนของคำที่เล็กที่สุดที่ใช้เติมหน้ารากศัพท์ (**root words** หรือ **stems**) เช่น การเติม **un-, im-, dis-** เป็นต้น ด้านหน้าคำต่างๆ และ ปัจจัย (**suffixes**) หรือ ส่วนของคำที่เล็กที่สุดที่ใช้เติมอยู่ท้ายรากศัพท์ เช่น การเติม **-tion, -ment, -ly** เป็นต้น ท้ายคำต่างๆ เพื่อให้เกิดคำใหม่ที่อาจมีความหมายแตกต่างออกไป ดังนั้น ความรู้เรื่อง หน่วยคำ จะทำให้เราสามารถเข้าใจในความหมายของคำต่างๆมากยิ่งขึ้น อีกทั้งยังช่วยให้สามารถเดาความหมายของคำศัพท์ได้จากการดูส่วนประกอบของคำ ตัวอย่างเช่น ถ้าเราไม่ทราบความหมายของคำว่า **homeless** เราสามารถใช้ความรู้เรื่อง ปัจจัยของหน่วยคำ เพื่อช่วยเดาความหมายของคำนี้ เนื่องจากคำนี้ประกอบด้วย รากศัพท์ คำว่า **home** และ ปัจจัย คำว่า **-less** แปลว่า ปราศจาก หรือ ไม่มี ที่เพิ่มเข้ามาในตอนท้ายของคำ ดังนั้น เราจะสามารถเดาความหมายของคำต่างๆได้ง่ายขึ้น หากเรามีความรู้เรื่อง ปัจจัย หรือ **suffixes** ตารางข้างล่าง คือ ปัจจัยที่มักถูกใช้บ่อยในภาษาอังกฤษ

| รากศัพท์ | ปัจจัย | ความหมาย | ตัวอย่าง |
|---|---|---|---|
| Show | -ed | past-tense verbs | show**ed** |
| Love | -ly | characteristic of | love**ly** |
| Car, Box | -s, -es | more than one/plural | car**s**, box**es** |
| Print | -able | able to be | print**able** |
| Success | -ful | full of | success**ful** |
| Conclude, Act | -sion, -tion/-ion | state of being | conclu**sion**, ac**tion** |
| Fulfill | -ment | act of/state of being | fulfill**ment** |
| Friend | -ship | position held | friend**ship** |

**นอกจากนี้ ปัจจัย หรือ suffixes ยังช่วยระบุชนิดของคำใหม่ที่ถูกสร้างขึ้นมา หรือที่เรียกว่า part of speech (POS) เช่น คำนาม (noun) คำกริยา (verb) คำคุณศัพท์ (adjective) หรือ คำวิเศษณ์ (adverb) ดังตัวอย่างในตารางด้านล่าง**

| Suffix | Example | POS |
|---|---|---|
| -ed | show**ed**, return**ed** | verb |
| -ly | love**ly**, luck**ily** | adverb |
| -s, -es | car**s**, handout**s**, box**es**, buss**es** | noun |
| -able | print**able**, reus**able** | adjective |
| -ful | success**ful** | adjective |
| -sion, -tion | conclu**sion**, presenta**tion** | noun |
| -ment | fulfill**ment** | noun |
| -ship | friend**ship** | noun |

## English Version



### Supplementary Handouts

### (Group 4)

**Word Parts**

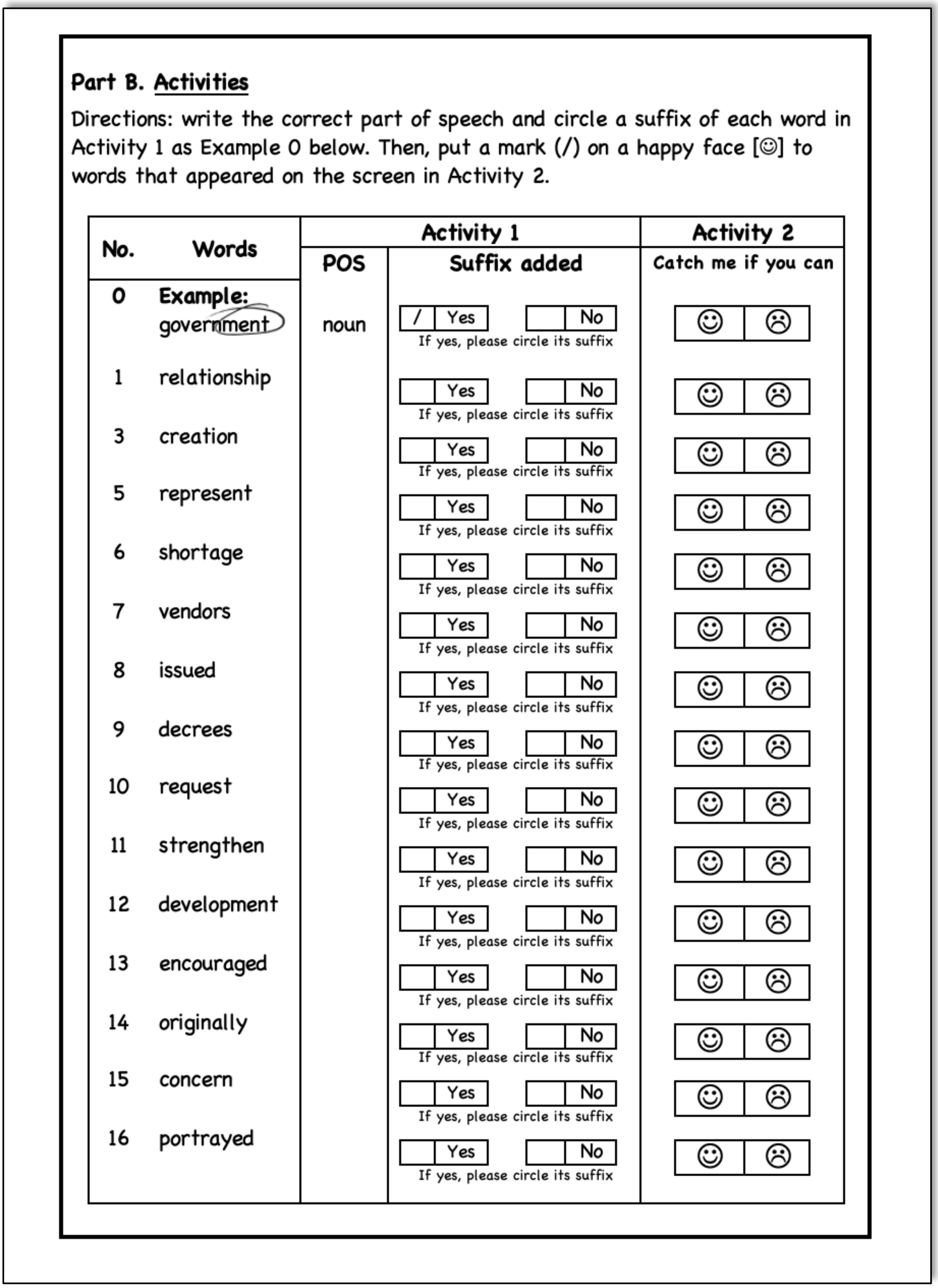
**Part A: Suffixes**

Word parts can be defined as parts of a word that are usually combined with **prefixes** (un-, im-, dis-, etc.) and **suffixes** (-tion, -ment, -ly, etc.). Adding prefixes or suffixes to a word can create a new word with a specific meaning. So, knowing word parts gives you understanding of word meanings. It can help you when guessing unknown words as well. For example, if you do not know the meaning of "**homeless**," you can use the knowledge of word parts to determine it. This word is a combination between "**home**" (root word) and "**-less**" (suffix, meaning **without**). You can determine its meaning easily if you have knowledge about suffixes. Below is a list of the most frequently used suffixes in English.

| Root | Suffix | Meaning | Example |
|---|---|---|---|
| Show | -ed | past-tense verbs | show**ed** |
| Love | -ly | characteristic of | love**ly** |
| Car, Box | -s, -es | more than one/plural | car**s**, box**es** |
| Print | -able | able to be | print**able** |
| Success | -ful | full of | success**ful** |
| Conclude, Act | -sion, -tion/-ion | state of being | conclu**sion**, ac**tion** |
| Fulfill | -ment | act of/state of being | fulfill**ment** |
| Friend | -ship | position held | friend**ship** |

Also, suffixes can indicate parts of speech (POS) such as noun, verb, adjective, and adverb in English language. Following is a list of examples.

| Suffix | Example | POS |
|---|---|---|
| -ed | show**ed**, return**ed** | verb |
| -ly | love**ly**, luck**ily** | adverb |
| -s, -es | car**s**, handout**s**, box**es**, buss**es** | noun |
| -able | print**able**, reus**able** | adjective |
| -ful | success**ful** | adjective |
| -sion, -tion | conclu**sion**, presenta**tion** | noun |
| -ment | fulfill**ment** | noun |
| -ship | friend**ship** | noun |



**Part B. Activities**

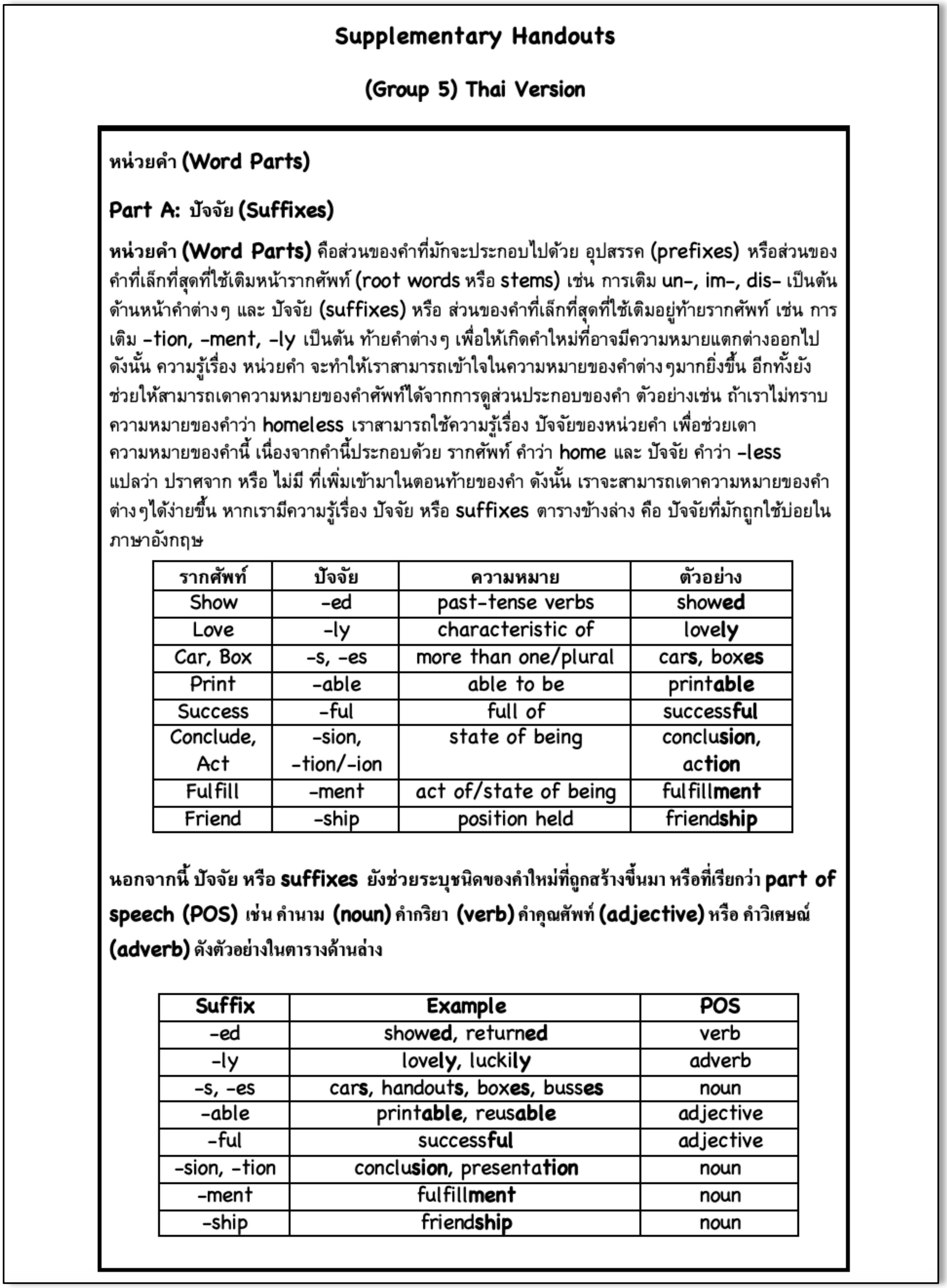
Directions: write the correct part of speech and circle a suffix of each word in Activity 1 as Example 0 below. Then, put a mark (/) on a happy face [☺] to words that appeared on the screen in Activity 2.

| No. | Words | Activity 1 | | Activity 2 |
|---|---|---|---|---|
| | | POS | Suffix added | Catch me if you can |
| 0 | Example:<br>government | noun | ☑ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 1 | relationship | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 3 | creation | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 5 | represent | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 6 | shortage | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 7 | vendors | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 8 | issued | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 9 | decrees | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 10 | request | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 11 | strengthen | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 12 | development | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 13 | encouraged | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 14 | originally | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 15 | concern | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |
| 16 | portrayed | | ☐ Yes ☐ No<br>If yes, please circle its suffix | ☺ ☹ |

*Note: the words are listed in the order they appeared in the passage and extra words are added to avoid noticing*

**Handouts for Group 5: reading and word parts** (TFA support for *Generative Use*)

## Thai Version



### Supplementary Handouts

### (Group 5) Thai Version

**หน่วยคำ (Word Parts)**

**Part A:** ปัจจัย **(Suffixes)**

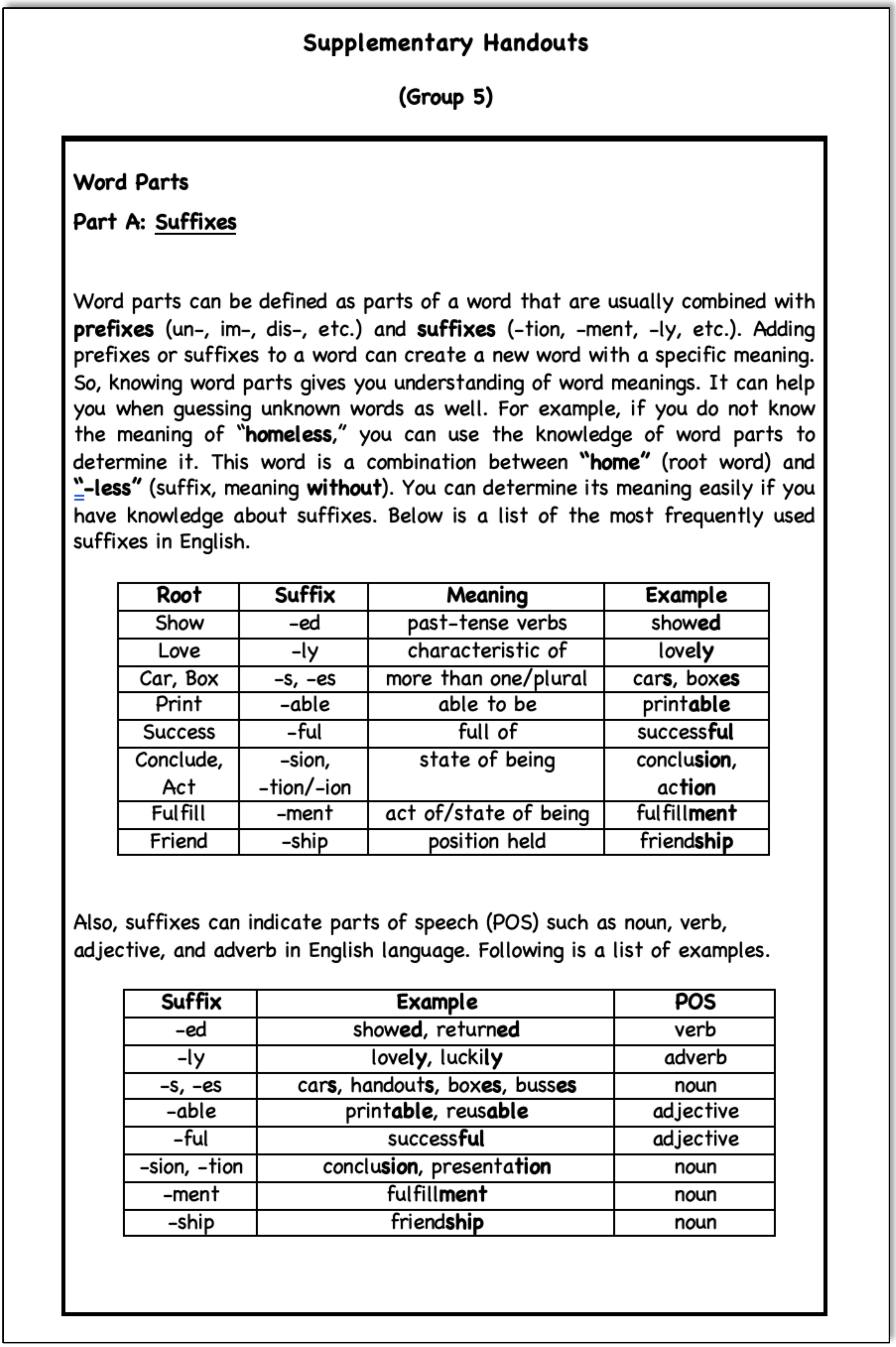
**หน่วยคำ (Word Parts)** คือส่วนของคำที่มักจะประกอบไปด้วย อุปสรรค (**prefixes**) หรือส่วนของคำที่เล็กที่สุดที่ใช้เติมหน้ารากศัพท์ (**root words** หรือ **stems**) เช่น การเติม **un-, im-, dis-** เป็นต้น ด้านหน้าคำต่างๆ และ ปัจจัย (**suffixes**) หรือ ส่วนของคำที่เล็กที่สุดที่ใช้เติมอยู่ท้ายรากศัพท์ เช่น การเติม **-tion, -ment, -ly** เป็นต้น ท้ายคำต่างๆ เพื่อให้เกิดคำใหม่ที่อาจมีความหมายแตกต่างออกไป ดังนั้น ความรู้เรื่อง หน่วยคำ จะทำให้เราสามารถเข้าใจในความหมายของคำต่างๆมากยิ่งขึ้น อีกทั้งยังช่วยให้สามารถเดาความหมายของคำศัพท์ได้จากการดูส่วนประกอบของคำ ตัวอย่างเช่น ถ้าเราไม่ทราบความหมายของคำว่า **homeless** เราสามารถใช้ความรู้เรื่อง ปัจจัยของหน่วยคำ เพื่อช่วยเดาความหมายของคำนี้ เนื่องจากคำนี้ประกอบด้วย รากศัพท์ คำว่า **home** และ ปัจจัย คำว่า **-less** แปลว่า ปราศจาก หรือ ไม่มี ที่เพิ่มเข้ามาในตอนท้ายของคำ ดังนั้น เราจะสามารถเดาความหมายของคำต่างๆได้ง่ายขึ้น หากเรามีความรู้เรื่อง ปัจจัย หรือ **suffixes** ตารางข้างล่าง คือ ปัจจัยที่มักถูกใช้บ่อยในภาษาอังกฤษ

| รากศัพท์ | ปัจจัย | ความหมาย | ตัวอย่าง |
|---|---|---|---|
| Show | -ed | past-tense verbs | show**ed** |
| Love | -ly | characteristic of | love**ly** |
| Car, Box | -s, -es | more than one/plural | car**s**, box**es** |
| Print | -able | able to be | print**able** |
| Success | -ful | full of | success**ful** |
| Conclude,<br>Act | -sion,<br>-tion/-ion | state of being | conclu**sion**,<br>ac**tion** |
| Fulfill | -ment | act of/state of being | fulfill**ment** |
| Friend | -ship | position held | friend**ship** |

นอกจากนี้ ปัจจัย หรือ **suffixes** ยังช่วยระบุชนิดของคำใหม่ที่ถูกสร้างขึ้นมา หรือที่เรียกว่า **part of speech (POS)** เช่น คำนาม **(noun)** คำกริยา **(verb)** คำคุณศัพท์ **(adjective)** หรือ คำวิเศษณ์ **(adverb)** ดังตัวอย่างในตารางด้านล่าง

| Suffix | Example | POS |
|---|---|---|
| -ed | show**ed**, return**ed** | verb |
| -ly | love**ly**, luckil**y** | adverb |
| -s, -es | car**s**, handout**s**, box**es**, buss**es** | noun |
| -able | print**able**, reus**able** | adjective |
| -ful | success**ful** | adjective |
| -sion, -tion | conclu**sion**, presenta**tion** | noun |
| -ment | fulfill**ment** | noun |
| -ship | friend**ship** | noun |

## English Version



## Supplementary Handouts

(Group 5)

### Word Parts

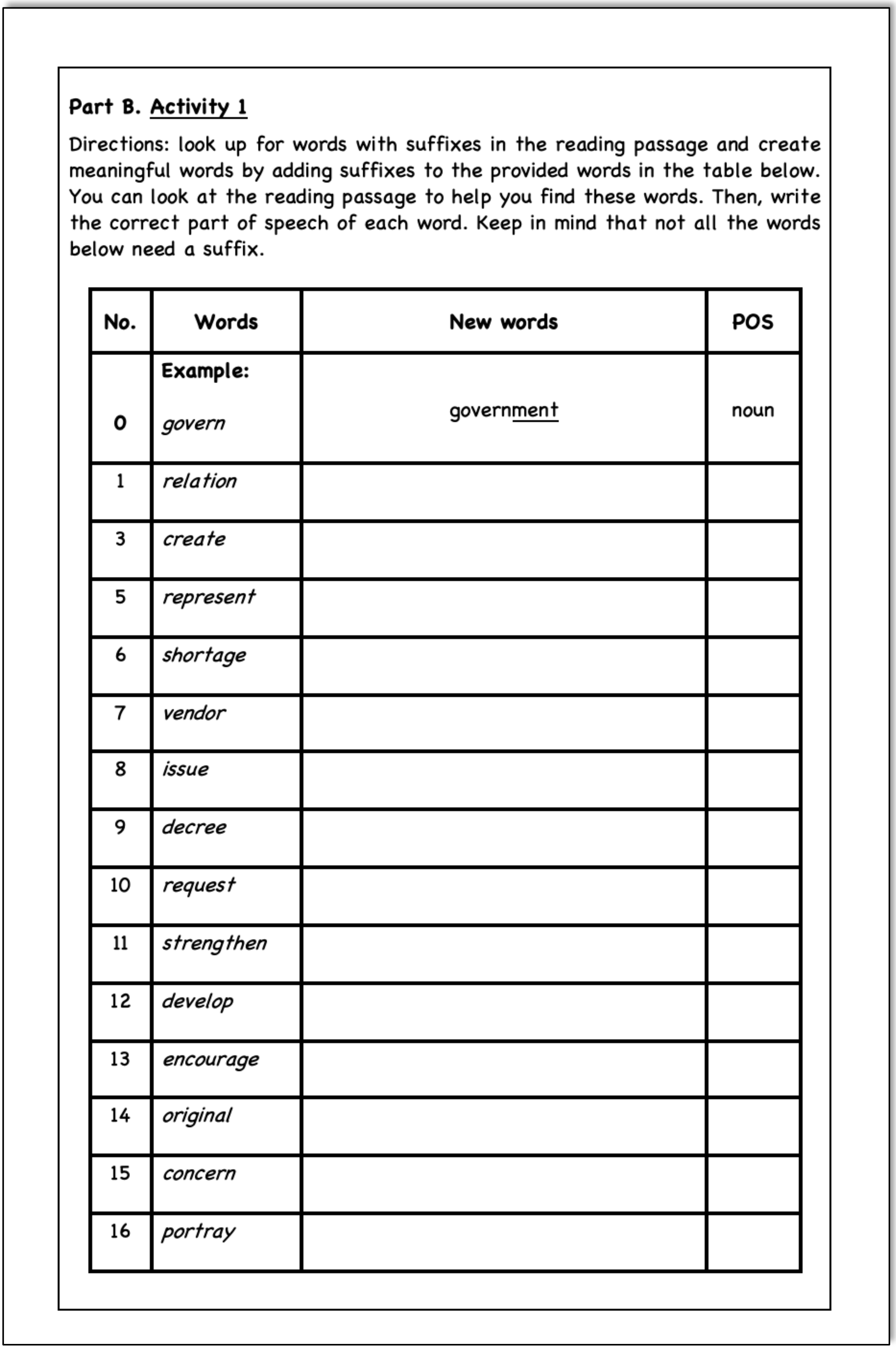
### Part A: Suffixes

Word parts can be defined as parts of a word that are usually combined with **prefixes** (un-, im-, dis-, etc.) and **suffixes** (-tion, -ment, -ly, etc.). Adding prefixes or suffixes to a word can create a new word with a specific meaning. So, knowing word parts gives you understanding of word meanings. It can help you when guessing unknown words as well. For example, if you do not know the meaning of "**homeless**," you can use the knowledge of word parts to determine it. This word is a combination between **"home"** (root word) and **"-less"** (suffix, meaning **without**). You can determine its meaning easily if you have knowledge about suffixes. Below is a list of the most frequently used suffixes in English.

| Root | Suffix | Meaning | Example |
|---|---|---|---|
| Show | -ed | past-tense verbs | show**ed** |
| Love | -ly | characteristic of | love**ly** |
| Car, Box | -s, -es | more than one/plural | car**s**, box**es** |
| Print | -able | able to be | print**able** |
| Success | -ful | full of | success**ful** |
| Conclude,<br>Act | -sion,<br>-tion/-ion | state of being | conclu**sion**,<br>ac**tion** |
| Fulfill | -ment | act of/state of being | fulfill**ment** |
| Friend | -ship | position held | friend**ship** |

Also, suffixes can indicate parts of speech (POS) such as noun, verb, adjective, and adverb in English language. Following is a list of examples.

| Suffix | Example | POS |
|---|---|---|
| -ed | show**ed**, return**ed** | verb |
| -ly | love**ly**, luck**ily** | adverb |
| -s, -es | car**s**, handout**s**, box**es**, buss**es** | noun |
| -able | print**able**, reus**able** | adjective |
| -ful | success**ful** | adjective |
| -sion, -tion | conclu**sion**, presenta**tion** | noun |
| -ment | fulfill**ment** | noun |
| -ship | friend**ship** | noun |

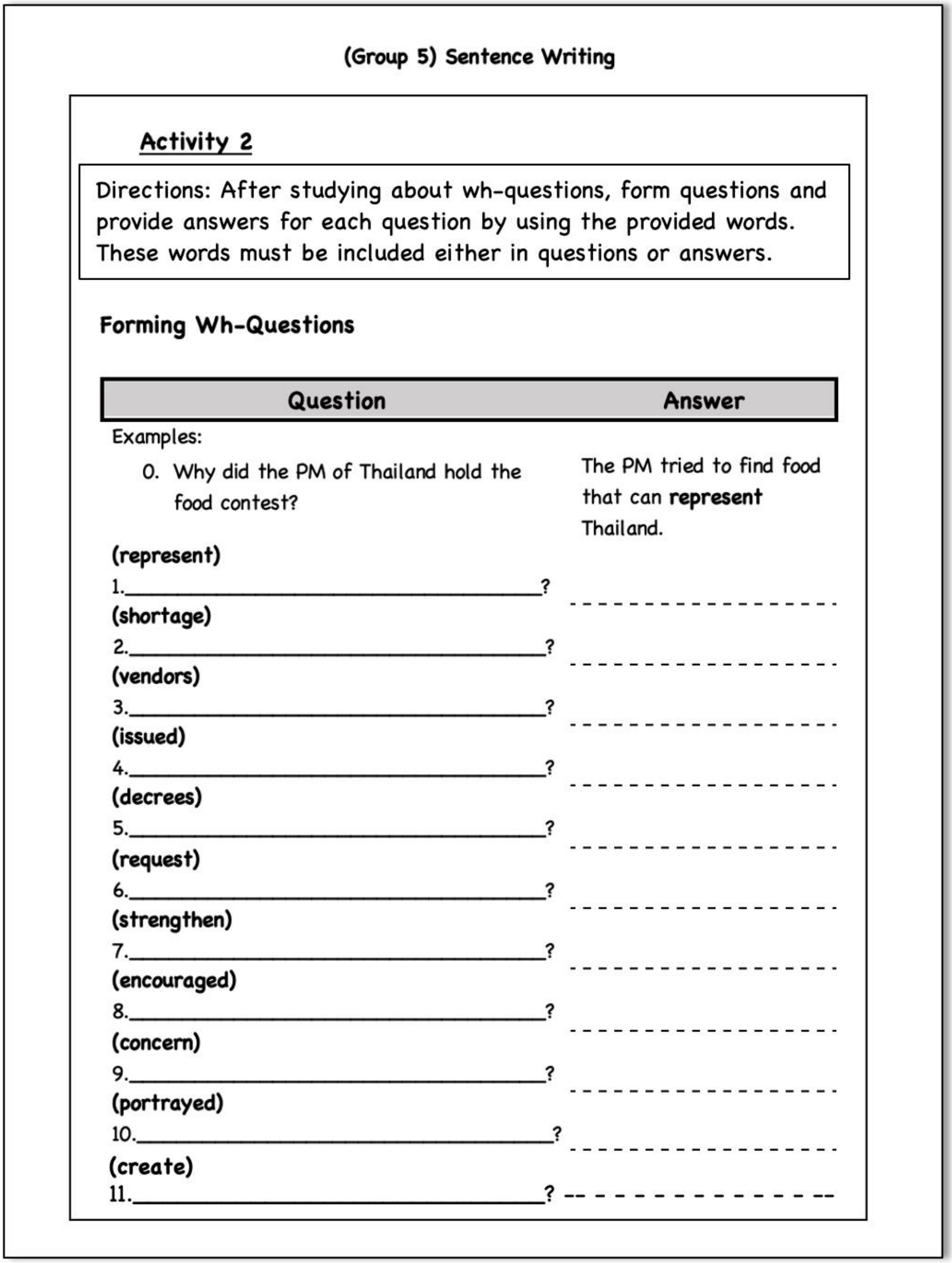


**Part B. Activity 1**

Directions: look up for words with suffixes in the reading passage and create meaningful words by adding suffixes to the provided words in the table below. You can look at the reading passage to help you find these words. Then, write the correct part of speech of each word. Keep in mind that not all the words below need a suffix.

| No. | Words | New words | POS |
|---|---|---|---|
| 0 | **Example:**<br>*govern* | government | noun |
| 1 | *relation* | | |
| 3 | *create* | | |
| 5 | *represent* | | |
| 6 | *shortage* | | |
| 7 | *vendor* | | |
| 8 | *issue* | | |
| 9 | *decree* | | |
| 10 | *request* | | |
| 11 | *strengthen* | | |
| 12 | *develop* | | |
| 13 | *encourage* | | |
| 14 | *original* | | |
| 15 | *concern* | | |
| 16 | *portray* | | |

*Note: the words are listed in the order they appeared in the passage and extra words are added to avoid noticing*



(Group 5) Sentence Writing

## Activity 2

Directions: After studying about wh-questions, form questions and provide answers for each question by using the provided words. These words must be included either in questions or answers.

### Forming Wh-Questions

| Question | Answer |
|---|---|
| Examples:<br>0. Why did the PM of Thailand hold the food contest? | The PM tried to find food that can **represent** Thailand. |

**(represent)**
1.________________________________? - - - - - - - - - - - - - - - - - -

**(shortage)**
2.________________________________? - - - - - - - - - - - - - - - - - -

**(vendors)**
3.________________________________? - - - - - - - - - - - - - - - - - -

**(issued)**
4.________________________________? - - - - - - - - - - - - - - - - - -

**(decrees)**
5.________________________________? - - - - - - - - - - - - - - - - - -

**(request)**
6.________________________________? - - - - - - - - - - - - - - - - - -

**(strengthen)**
7.________________________________? - - - - - - - - - - - - - - - - - -

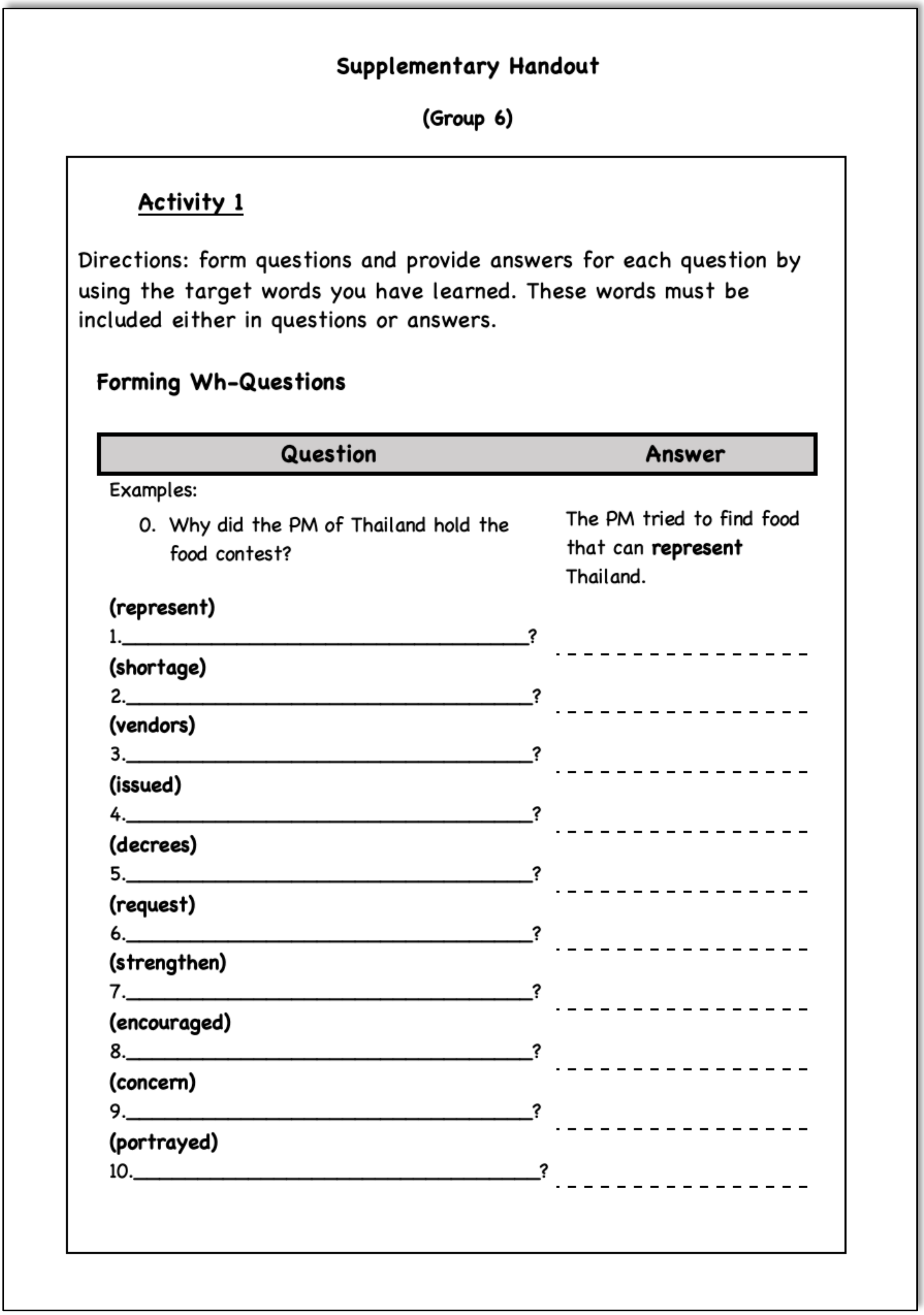
**(encouraged)**
8.________________________________? - - - - - - - - - - - - - - - - - -

**(concern)**
9.________________________________? - - - - - - - - - - - - - - - - - -

**(portrayed)**
10.________________________________? - - - - - - - - - - - - - - - - - -

**(create)**
11.________________________________? -- - - - - - - - - - - - - --

**Handouts for Group 6 reading, wordcards, and writing** (Full TFA support for *All TFA Components* group)



**Supplementary Handout**

**(Group 6)**

**<u>Activity 1</u>**

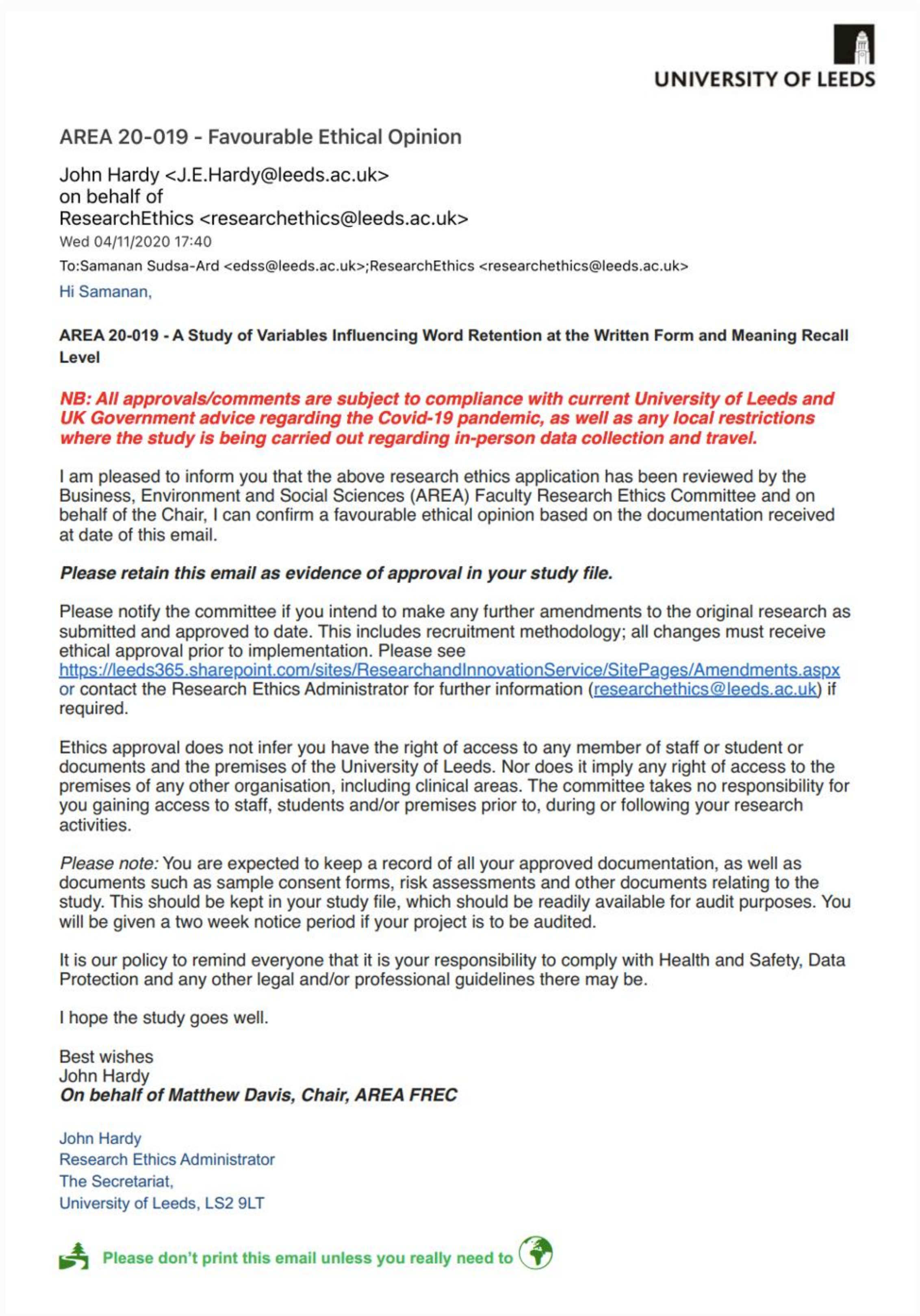
Directions: form questions and provide answers for each question by using the target words you have learned. These words must be included either in questions or answers.

**Forming Wh-Questions**

| Question | Answer |
|---|---|
| Examples: | |
| 0. Why did the PM of Thailand hold the food contest? | The PM tried to find food that can **represent** Thailand. |
| **(represent)**<br>1.______________________________? | ---------------- |
| **(shortage)**<br>2.______________________________? | ---------------- |
| **(vendors)**<br>3.______________________________? | ---------------- |
| **(issued)**<br>4.______________________________? | ---------------- |
| **(decrees)**<br>5.______________________________? | ---------------- |
| **(request)**<br>6.______________________________? | ---------------- |
| **(strengthen)**<br>7.______________________________? | ---------------- |
| **(encouraged)**<br>8.______________________________? | ---------------- |
| **(concern)**<br>9.______________________________? | ---------------- |
| **(portrayed)**<br>10.______________________________? | ---------------- |

*Note: the words are listed in the order they appeared in the passage*

## Appendix 4a: Ethics Approval Document of the Pilot Study



UNIVERSITY OF LEEDS

AREA 20-019 - Favourable Ethical Opinion

John Hardy <J.E.Hardy@leeds.ac.uk>
on behalf of
ResearchEthics <researchethics@leeds.ac.uk>
Wed 04/11/2020 17:40
To:Samanan Sudsa-Ard <edss@leeds.ac.uk>;ResearchEthics <researchethics@leeds.ac.uk>

Hi Samanan,

**AREA 20-019 - A Study of Variables Influencing Word Retention at the Written Form and Meaning Recall Level**

***NB: All approvals/comments are subject to compliance with current University of Leeds and UK Government advice regarding the Covid-19 pandemic, as well as any local restrictions where the study is being carried out regarding in-person data collection and travel.***

I am pleased to inform you that the above research ethics application has been reviewed by the Business, Environment and Social Sciences (AREA) Faculty Research Ethics Committee and on behalf of the Chair, I can confirm a favourable ethical opinion based on the documentation received at date of this email.

***Please retain this email as evidence of approval in your study file.***

Please notify the committee if you intend to make any further amendments to the original research as submitted and approved to date. This includes recruitment methodology; all changes must receive ethical approval prior to implementation. Please see https://leeds365.sharepoint.com/sites/ResearchandInnovationService/SitePages/Amendments.aspx or contact the Research Ethics Administrator for further information (researchethics@leeds.ac.uk) if required.

Ethics approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

*Please note:* You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, risk assessments and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

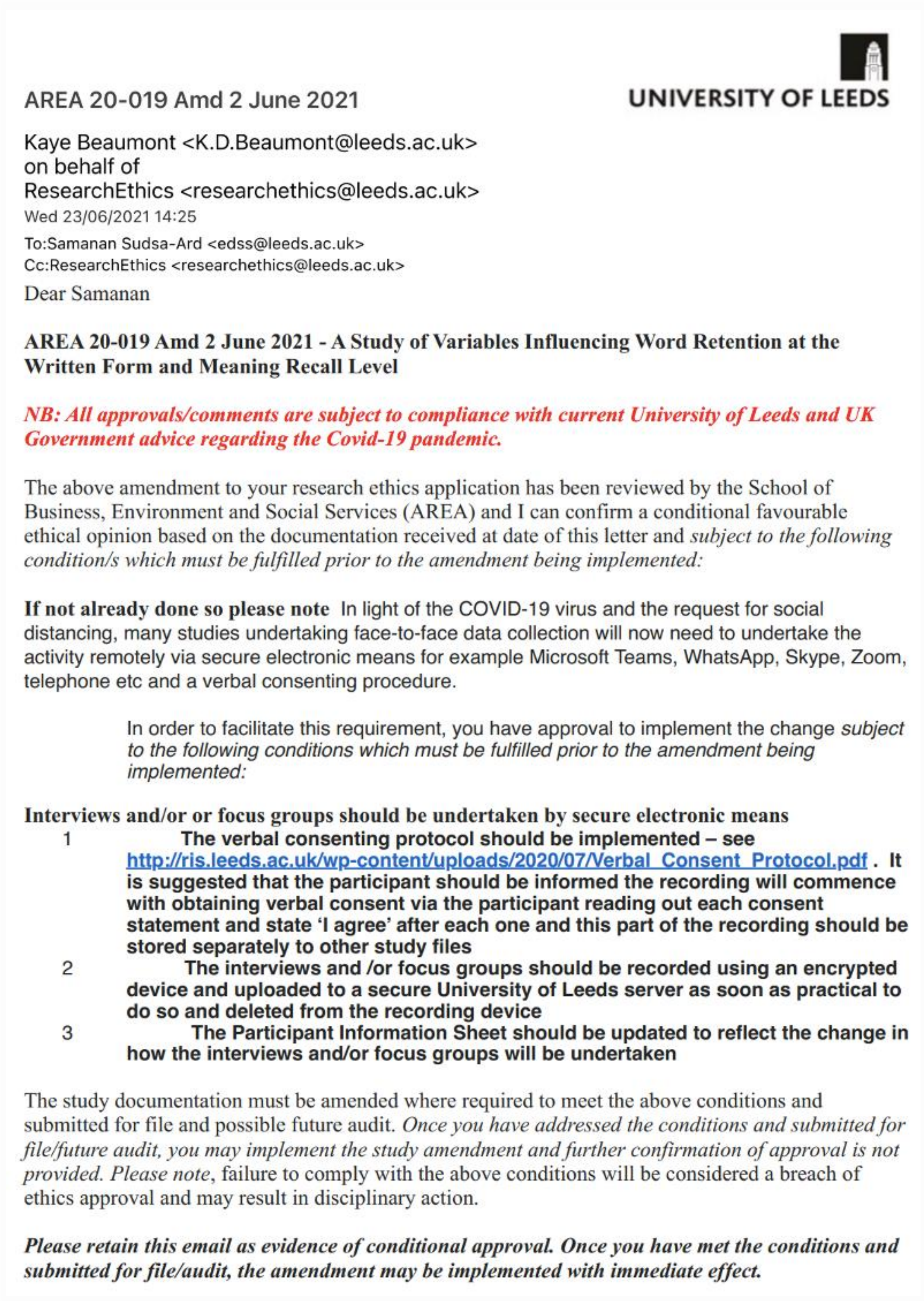
I hope the study goes well.

Best wishes
John Hardy
***On behalf of Matthew Davis, Chair, AREA FREC***

John Hardy
Research Ethics Administrator
The Secretariat,
University of Leeds, LS2 9LT

**Please don't print this email unless you really need to**

## Appendix 4b: Ethics Approval Document of the Main Study



AREA 20-019 Amd 2 June 2021

UNIVERSITY OF LEEDS

Kaye Beaumont <K.D.Beaumont@leeds.ac.uk>
on behalf of
ResearchEthics <researchethics@leeds.ac.uk>
Wed 23/06/2021 14:25
To:Samanan Sudsa-Ard <edss@leeds.ac.uk>
Cc:ResearchEthics <researchethics@leeds.ac.uk>

Dear Samanan

**AREA 20-019 Amd 2 June 2021 - A Study of Variables Influencing Word Retention at the Written Form and Meaning Recall Level**

***NB: All approvals/comments are subject to compliance with current University of Leeds and UK Government advice regarding the Covid-19 pandemic.***

The above amendment to your research ethics application has been reviewed by the School of Business, Environment and Social Services (AREA) and I can confirm a conditional favourable ethical opinion based on the documentation received at date of this letter and *subject to the following condition/s which must be fulfilled prior to the amendment being implemented:*

**If not already done so please note** In light of the COVID-19 virus and the request for social distancing, many studies undertaking face-to-face data collection will now need to undertake the activity remotely via secure electronic means for example Microsoft Teams, WhatsApp, Skype, Zoom, telephone etc and a verbal consenting procedure.

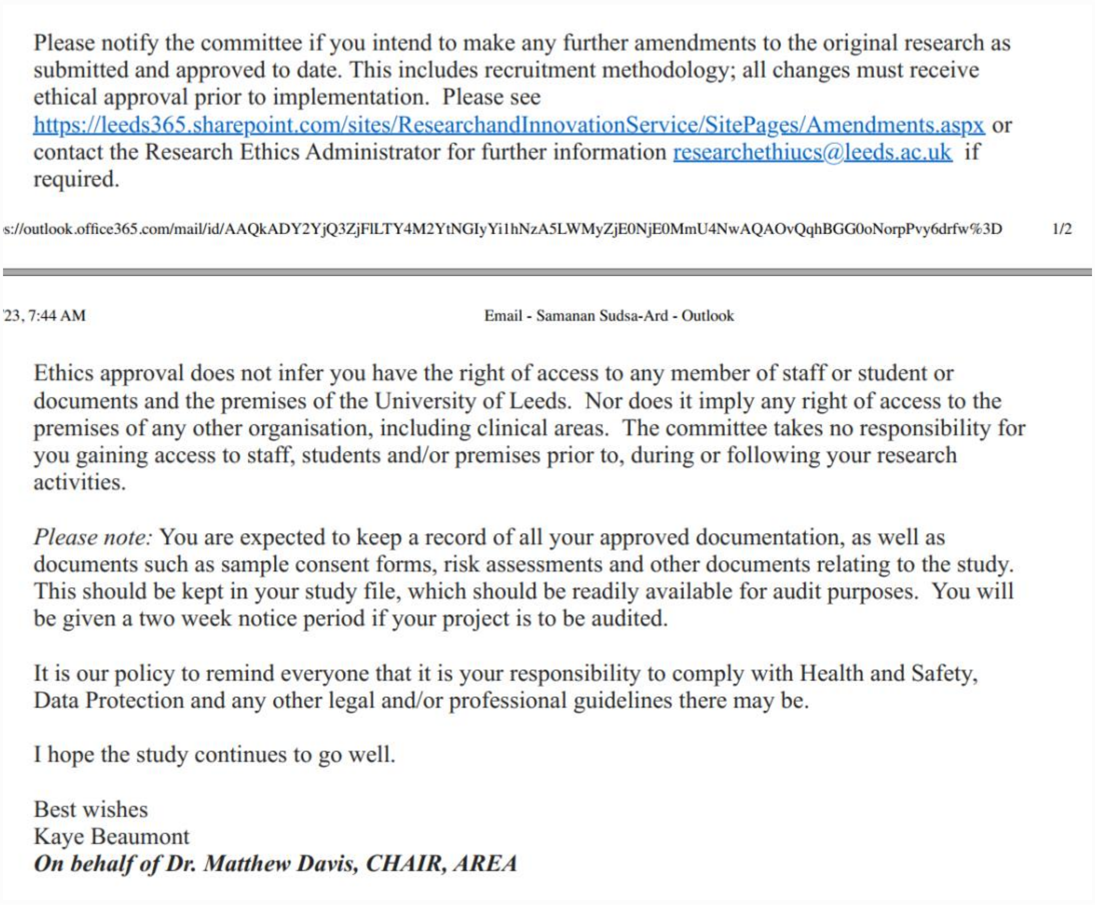
> In order to facilitate this requirement, you have approval to implement the change *subject to the following conditions which must be fulfilled prior to the amendment being implemented:*

**Interviews and/or or focus groups should be undertaken by secure electronic means**

1. **The verbal consenting protocol should be implemented – see http://ris.leeds.ac.uk/wp-content/uploads/2020/07/Verbal_Consent_Protocol.pdf . It is suggested that the participant should be informed the recording will commence with obtaining verbal consent via the participant reading out each consent statement and state 'I agree' after each one and this part of the recording should be stored separately to other study files**
2. **The interviews and /or focus groups should be recorded using an encrypted device and uploaded to a secure University of Leeds server as soon as practical to do so and deleted from the recording device**
3. **The Participant Information Sheet should be updated to reflect the change in how the interviews and/or focus groups will be undertaken**

The study documentation must be amended where required to meet the above conditions and submitted for file and possible future audit. *Once you have addressed the conditions and submitted for file/future audit, you may implement the study amendment and further confirmation of approval is not provided. Please note*, failure to comply with the above conditions will be considered a breach of ethics approval and may result in disciplinary action.

***Please retain this email as evidence of conditional approval. Once you have met the conditions and submitted for file/audit, the amendment may be implemented with immediate effect.***



Please notify the committee if you intend to make any further amendments to the original research as submitted and approved to date. This includes recruitment methodology; all changes must receive ethical approval prior to implementation. Please see https://leeds365.sharepoint.com/sites/ResearchandInnovationService/SitePages/Amendments.aspx or contact the Research Ethics Administrator for further information researchethiucs@leeds.ac.uk if required.

s://outlook.office365.com/mail/id/AAQkADY2YjQ3ZjFlLTY4M2YtNGIyYi1hNzA5LWMyZjE0NjE0MmU4NwAQAOvQqhBGG0oNorpPvy6drfw%3D 1/2

'23, 7:44 AM Email - Samanan Sudsa-Ard - Outlook

Ethics approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

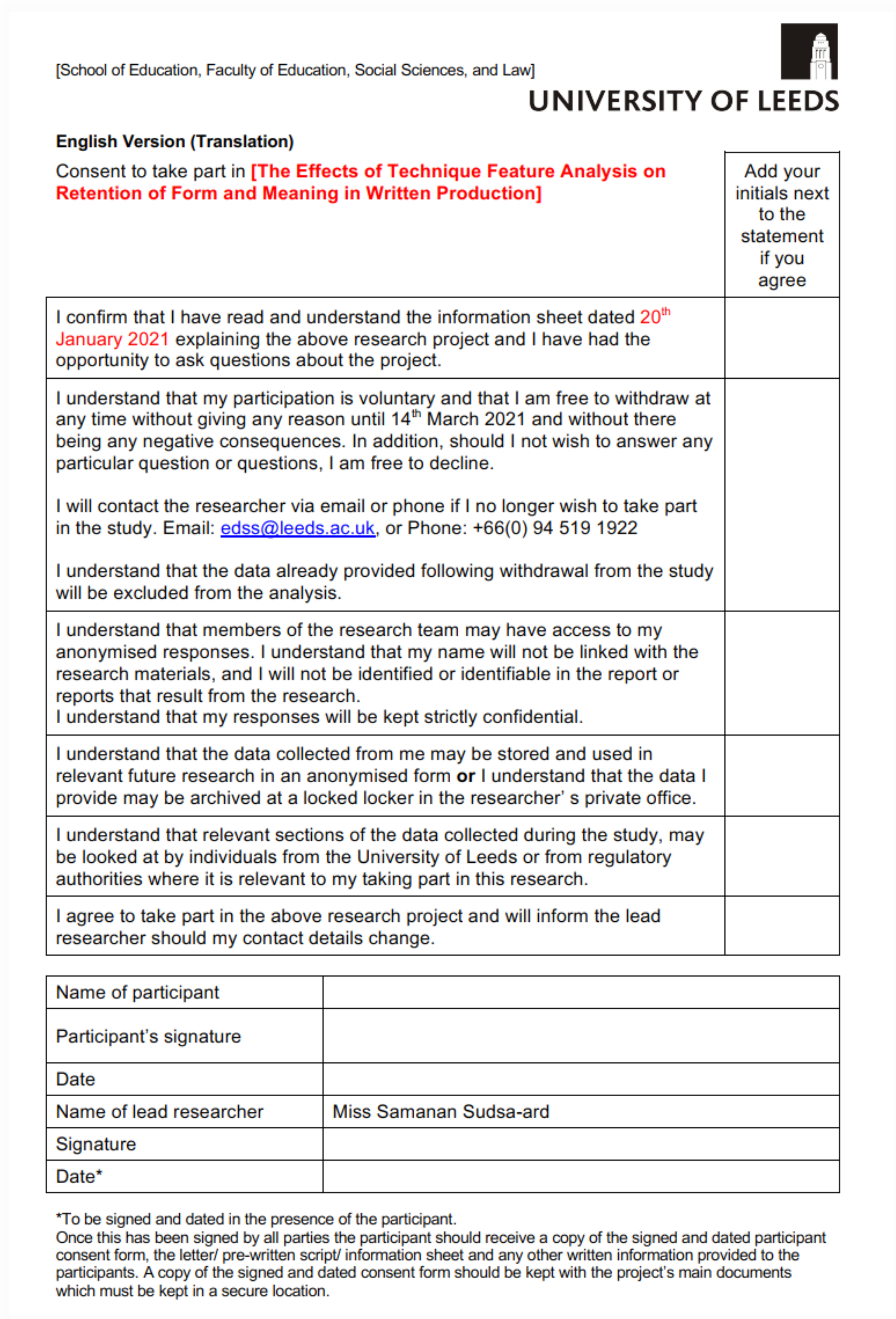
*Please note:* You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, risk assessments and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

I hope the study continues to go well.

Best wishes<br>
Kaye Beaumont<br>
***On behalf of Dr. Matthew Davis, CHAIR, AREA***

# Appendix 5a: Consent Form and Information Sheet of the Pilot Study (AREA20-019_V.2.1_amd20-01-21_in_RED)



[School of Education, Faculty of Education, Social Sciences, and Law]

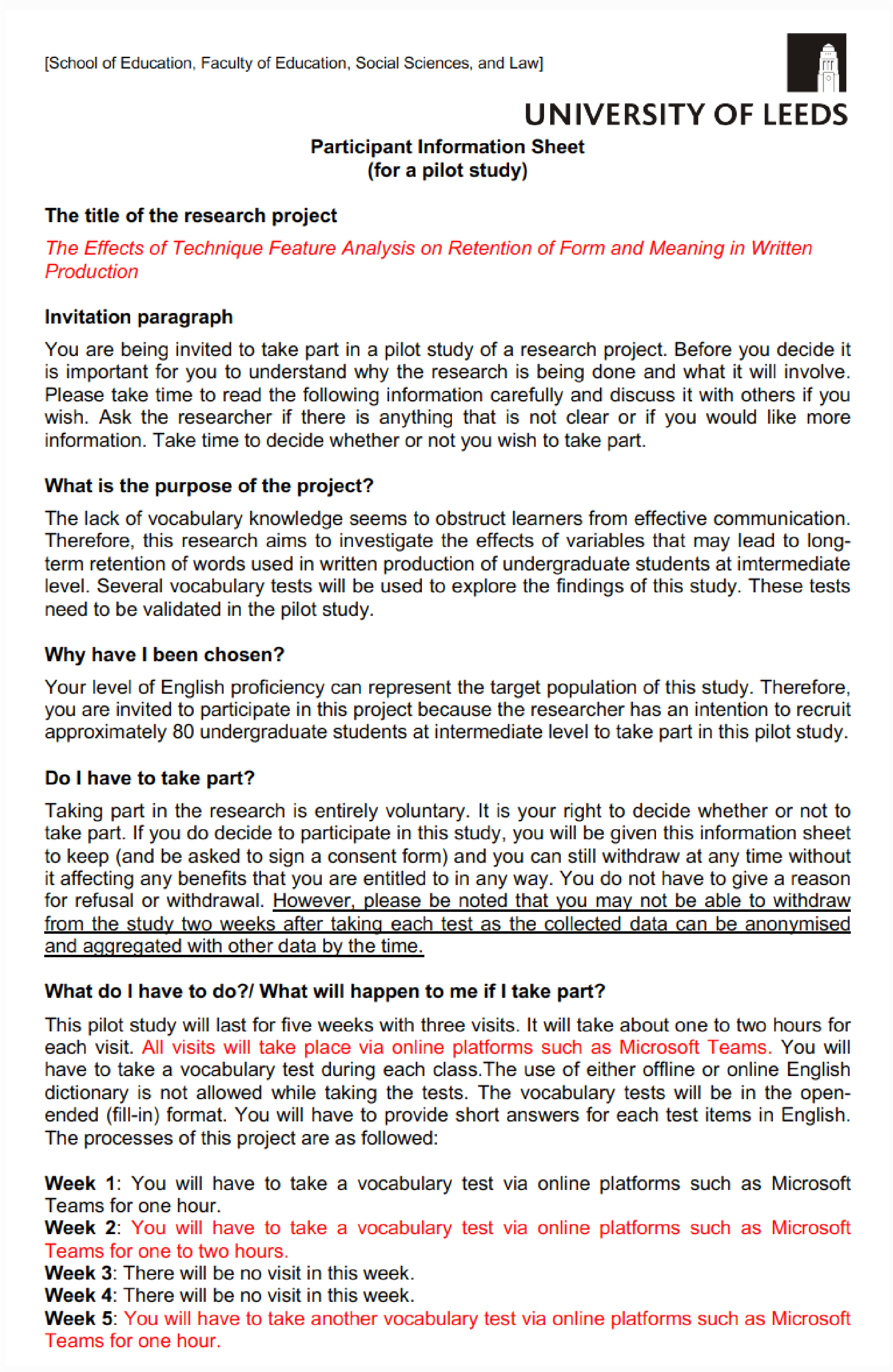
UNIVERSITY OF LEEDS

**English Version (Translation)**

| Consent to take part in **[The Effects of Technique Feature Analysis on Retention of Form and Meaning in Written Production]** | Add your initials next to the statement if you agree |
|---|---|
| I confirm that I have read and understand the information sheet dated 20th January 2021 explaining the above research project and I have had the opportunity to ask questions about the project. | |
| I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason until 14th March 2021 and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline.<br>I will contact the researcher via email or phone if I no longer wish to take part in the study. Email: edss@leeds.ac.uk, or Phone: +66(0) 94 519 1922<br>I understand that the data already provided following withdrawal from the study will be excluded from the analysis. | |
| I understand that members of the research team may have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.<br>I understand that my responses will be kept strictly confidential. | |
| I understand that the data collected from me may be stored and used in relevant future research in an anonymised form **or** I understand that the data I provide may be archived at a locked locker in the researcher' s private office. | |
| I understand that relevant sections of the data collected during the study, may be looked at by individuals from the University of Leeds or from regulatory authorities where it is relevant to my taking part in this research. | |
| I agree to take part in the above research project and will inform the lead researcher should my contact details change. | |

| | |
|---|---|
| Name of participant | |
| Participant's signature | |
| Date | |
| Name of lead researcher | Miss Samanan Sudsa-ard |
| Signature | |
| Date* | |

*To be signed and dated in the presence of the participant.
Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/ pre-written script/ information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be kept with the project's main documents which must be kept in a secure location.



[School of Education, Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

### Participant Information Sheet
### (for a pilot study)

**The title of the research project**

*The Effects of Technique Feature Analysis on Retention of Form and Meaning in Written Production*

**Invitation paragraph**

You are being invited to take part in a pilot study of a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask the researcher if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

**What is the purpose of the project?**

The lack of vocabulary knowledge seems to obstruct learners from effective communication. Therefore, this research aims to investigate the effects of variables that may lead to long-term retention of words used in written production of undergraduate students at imtermediate level. Several vocabulary tests will be used to explore the findings of this study. These tests need to be validated in the pilot study.

**Why have I been chosen?**

Your level of English proficiency can represent the target population of this study. Therefore, you are invited to participate in this project because the researcher has an intention to recruit approximately 80 undergraduate students at intermediate level to take part in this pilot study.

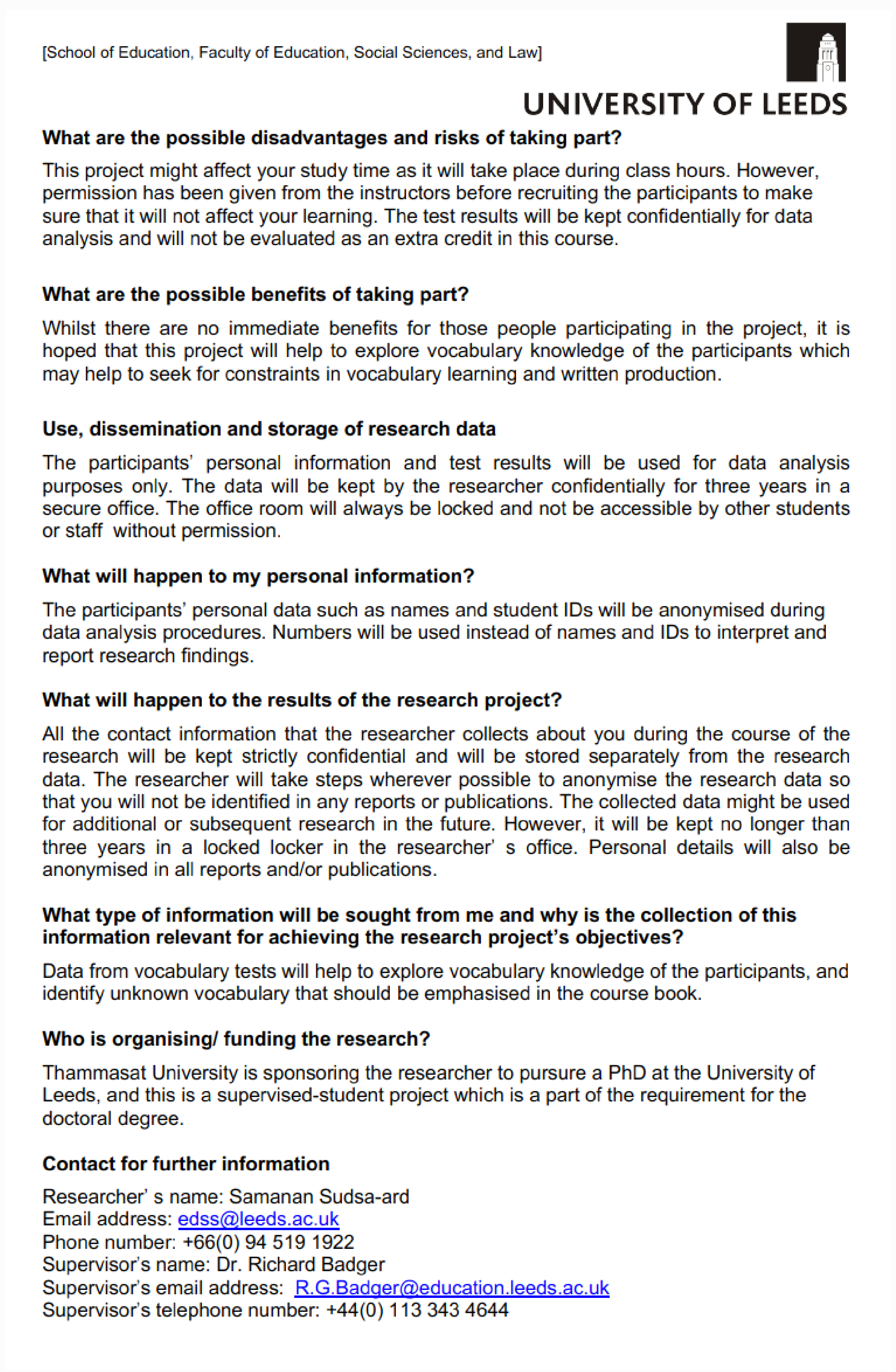
**Do I have to take part?**

Taking part in the research is entirely voluntary. It is your right to decide whether or not to take part. If you do decide to participate in this study, you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason for refusal or withdrawal. However, please be noted that you may not be able to withdraw from the study two weeks after taking each test as the collected data can be anonymised and aggregated with other data by the time.

**What do I have to do?/ What will happen to me if I take part?**

This pilot study will last for five weeks with three visits. It will take about one to two hours for each visit. All visits will take place via online platforms such as Microsoft Teams. You will have to take a vocabulary test during each class.The use of either offline or online English dictionary is not allowed while taking the tests. The vocabulary tests will be in the open-ended (fill-in) format. You will have to provide short answers for each test items in English. The processes of this project are as followed:

**Week 1**: You will have to take a vocabulary test via online platforms such as Microsoft Teams for one hour.
**Week 2**: You will have to take a vocabulary test via online platforms such as Microsoft Teams for one to two hours.
**Week 3**: There will be no visit in this week.
**Week 4**: There will be no visit in this week.
**Week 5**: You will have to take another vocabulary test via online platforms such as Microsoft Teams for one hour.



[School of Education, Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

**What are the possible disadvantages and risks of taking part?**

This project might affect your study time as it will take place during class hours. However, permission has been given from the instructors before recruiting the participants to make sure that it will not affect your learning. The test results will be kept confidentially for data analysis and will not be evaluated as an extra credit in this course.

**What are the possible benefits of taking part?**

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this project will help to explore vocabulary knowledge of the participants which may help to seek for constraints in vocabulary learning and written production.

**Use, dissemination and storage of research data**

The participants' personal information and test results will be used for data analysis purposes only. The data will be kept by the researcher confidentially for three years in a secure office. The office room will always be locked and not be accessible by other students or staff without permission.

**What will happen to my personal information?**

The participants' personal data such as names and student IDs will be anonymised during data analysis procedures. Numbers will be used instead of names and IDs to interpret and report research findings.

**What will happen to the results of the research project?**

All the contact information that the researcher collects about you during the course of the research will be kept strictly confidential and will be stored separately from the research data. The researcher will take steps wherever possible to anonymise the research data so that you will not be identified in any reports or publications. The collected data might be used for additional or subsequent research in the future. However, it will be kept no longer than three years in a locked locker in the researcher' s office. Personal details will also be anonymised in all reports and/or publications.

**What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?**

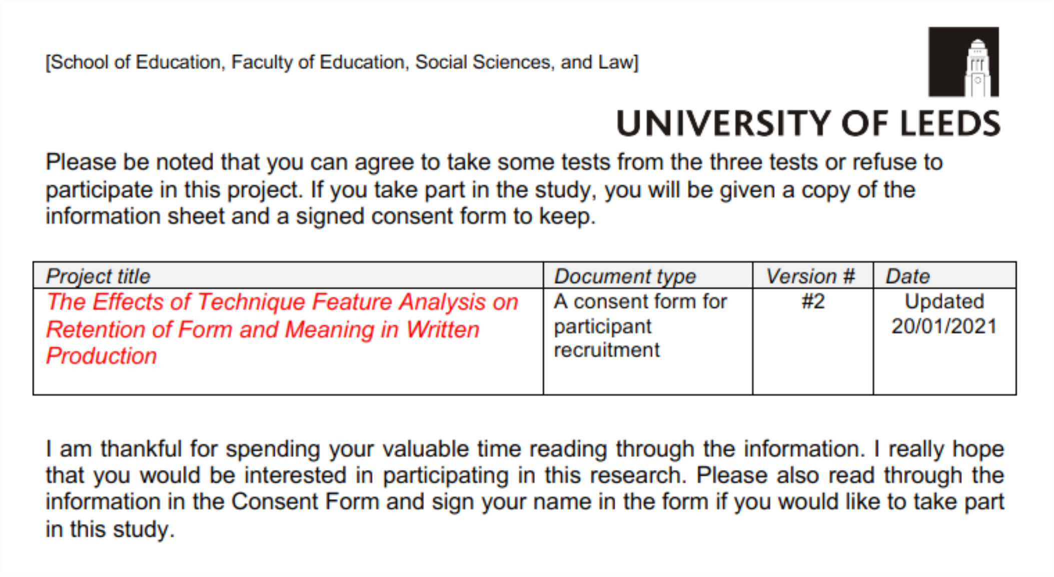
Data from vocabulary tests will help to explore vocabulary knowledge of the participants, and identify unknown vocabulary that should be emphasised in the course book.

**Who is organising/ funding the research?**

Thammasat University is sponsoring the researcher to pursure a PhD at the University of Leeds, and this is a supervised-student project which is a part of the requirement for the doctoral degree.

**Contact for further information**

Researcher' s name: Samanan Sudsa-ard
Email address: edss@leeds.ac.uk
Phone number: +66(0) 94 519 1922
Supervisor's name: Dr. Richard Badger
Supervisor's email address: R.G.Badger@education.leeds.ac.uk
Supervisor's telephone number: +44(0) 113 343 4644



[School of Education, Faculty of Education, Social Sciences, and Law]

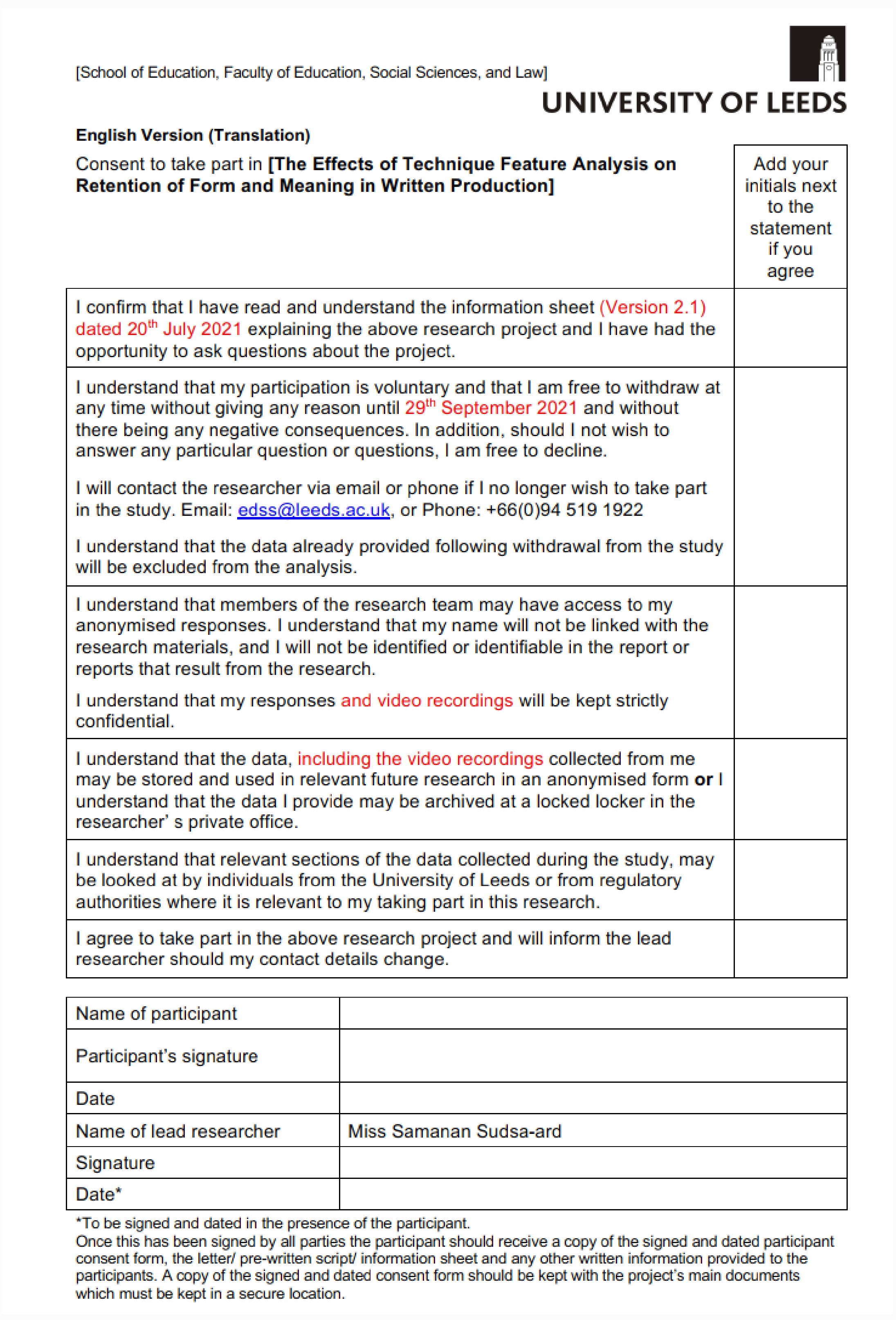
UNIVERSITY OF LEEDS

Please be noted that you can agree to take some tests from the three tests or refuse to participate in this project. If you take part in the study, you will be given a copy of the information sheet and a signed consent form to keep.

| *Project title* | *Document type* | *Version #* | *Date* |
|---|---|---|---|
| *The Effects of Technique Feature Analysis on Retention of Form and Meaning in Written Production* | A consent form for participant recruitment | #2 | Updated 20/01/2021 |

I am thankful for spending your valuable time reading through the information. I really hope that you would be interested in participating in this research. Please also read through the information in the Consent Form and sign your name in the form if you would like to take part in this study.

# Appendix 5b: Consent Form and Information Sheet of the Main Study (AREA20-019_V.2.1_amd20-07-21_in_RED)



[School of Education, Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

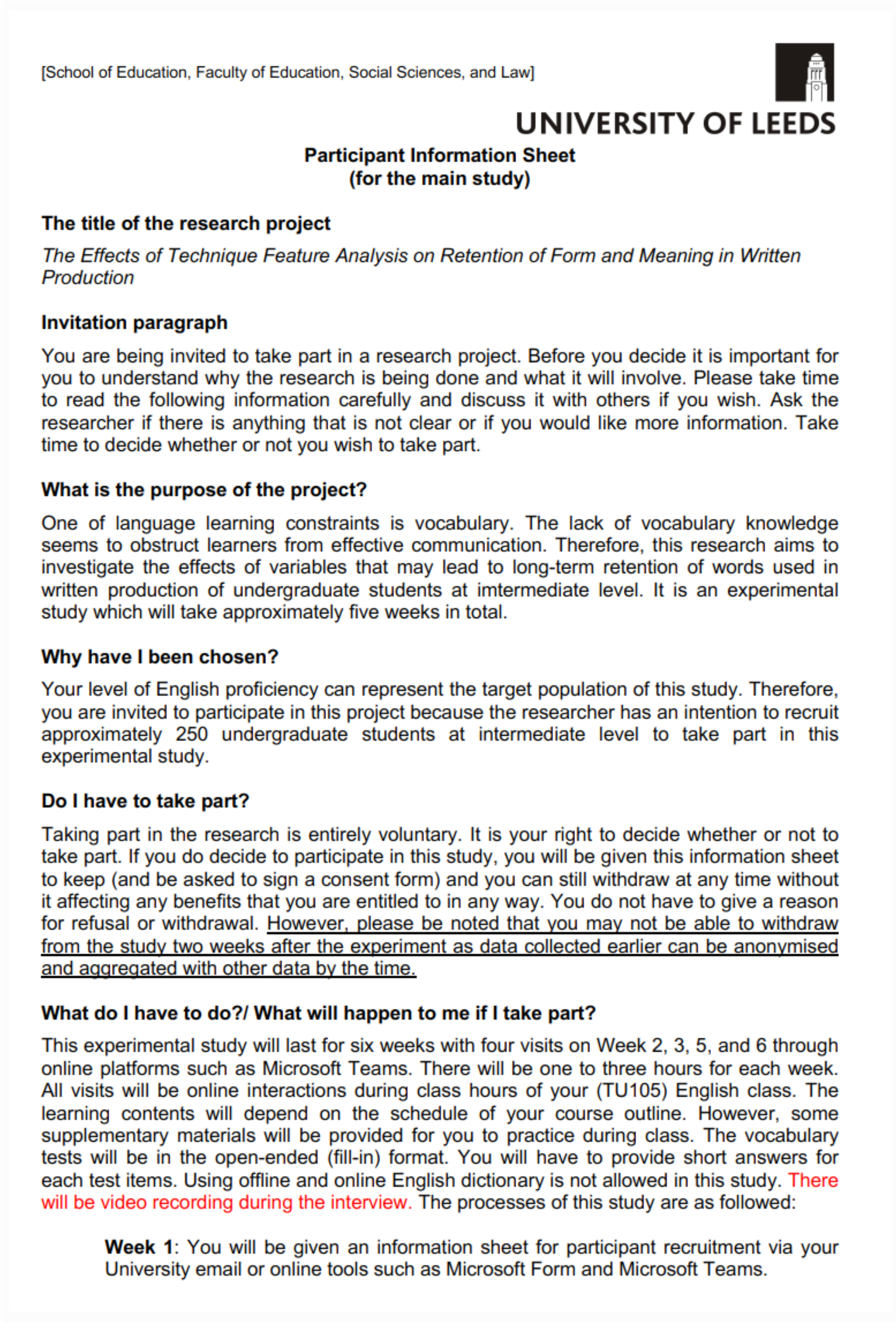
**English Version (Translation)**

| Consent to take part in **[The Effects of Technique Feature Analysis on Retention of Form and Meaning in Written Production]** | Add your initials next to the statement if you agree |
|---|---|
| I confirm that I have read and understand the information sheet (Version 2.1) dated 20th July 2021 explaining the above research project and I have had the opportunity to ask questions about the project. | |
| I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason until 29th September 2021 and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline.<br>I will contact the researcher via email or phone if I no longer wish to take part in the study. Email: edss@leeds.ac.uk, or Phone: +66(0)94 519 1922<br>I understand that the data already provided following withdrawal from the study will be excluded from the analysis. | |
| I understand that members of the research team may have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.<br>I understand that my responses and video recordings will be kept strictly confidential. | |
| I understand that the data, including the video recordings collected from me may be stored and used in relevant future research in an anonymised form **or** I understand that the data I provide may be archived at a locked locker in the researcher' s private office. | |
| I understand that relevant sections of the data collected during the study, may be looked at by individuals from the University of Leeds or from regulatory authorities where it is relevant to my taking part in this research. | |
| I agree to take part in the above research project and will inform the lead researcher should my contact details change. | |

| | |
|---|---|
| Name of participant | |
| Participant's signature | |
| Date | |
| Name of lead researcher | Miss Samanan Sudsa-ard |
| Signature | |
| Date* | |

*To be signed and dated in the presence of the participant.

Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/ pre-written script/ information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be kept with the project's main documents which must be kept in a secure location.



[School of Education, Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

**Participant Information Sheet**
**(for the main study)**

### The title of the research project

*The Effects of Technique Feature Analysis on Retention of Form and Meaning in Written Production*

### Invitation paragraph

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask the researcher if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

### What is the purpose of the project?

One of language learning constraints is vocabulary. The lack of vocabulary knowledge seems to obstruct learners from effective communication. Therefore, this research aims to investigate the effects of variables that may lead to long-term retention of words used in written production of undergraduate students at imtermediate level. It is an experimental study which will take approximately five weeks in total.

### Why have I been chosen?

Your level of English proficiency can represent the target population of this study. Therefore, you are invited to participate in this project because the researcher has an intention to recruit approximately 250 undergraduate students at intermediate level to take part in this experimental study.

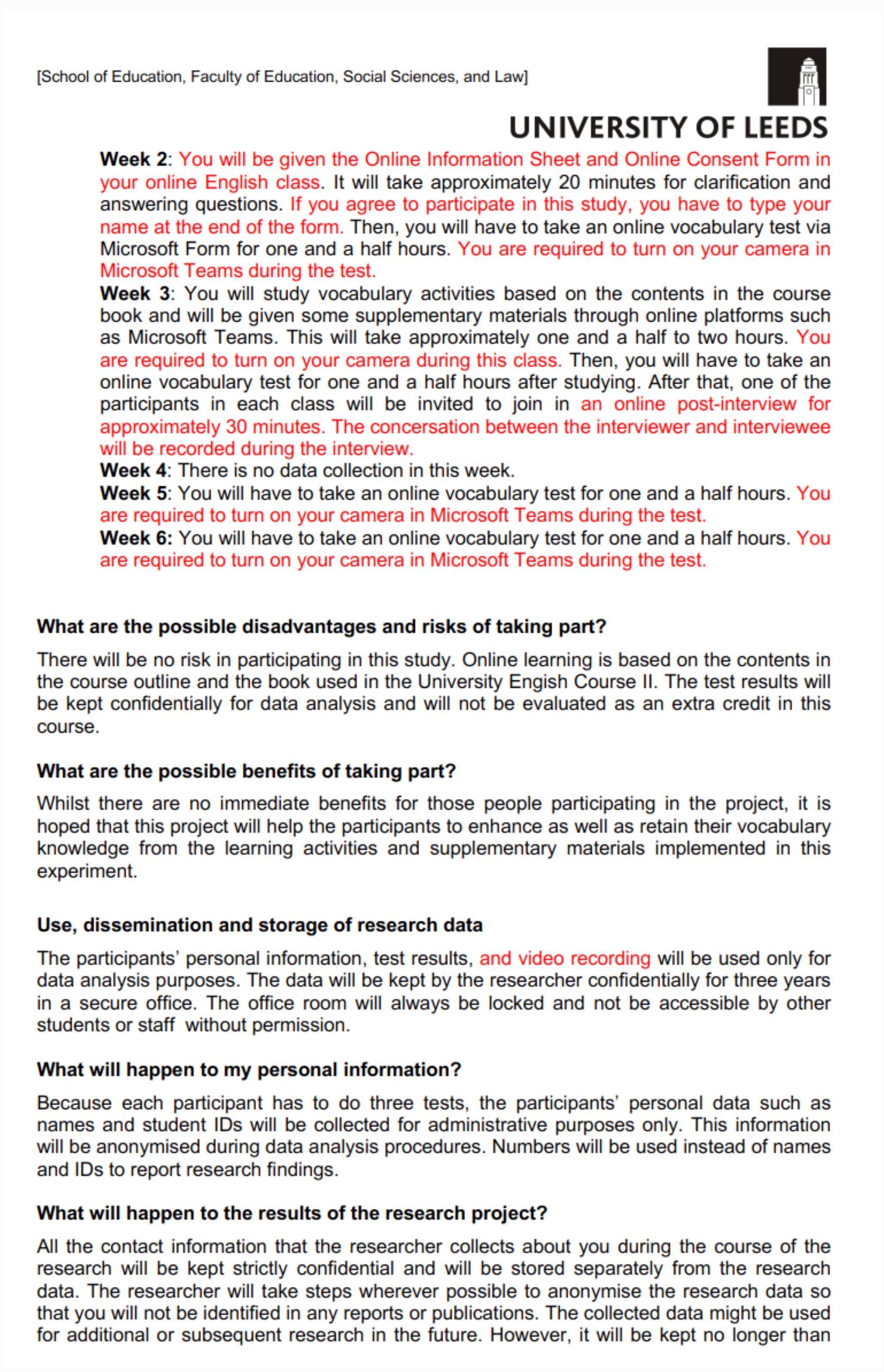
### Do I have to take part?

Taking part in the research is entirely voluntary. It is your right to decide whether or not to take part. If you do decide to participate in this study, you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason for refusal or withdrawal. However, please be noted that you may not be able to withdraw from the study two weeks after the experiment as data collected earlier can be anonymised and aggregated with other data by the time.

### What do I have to do?/ What will happen to me if I take part?

This experimental study will last for six weeks with four visits on Week 2, 3, 5, and 6 through online platforms such as Microsoft Teams. There will be one to three hours for each week. All visits will be online interactions during class hours of your (TU105) English class. The learning contents will depend on the schedule of your course outline. However, some supplementary materials will be provided for you to practice during class. The vocabulary tests will be in the open-ended (fill-in) format. You will have to provide short answers for each test items. Using offline and online English dictionary is not allowed in this study. There will be video recording during the interview. The processes of this study are as followed:

**Week 1**: You will be given an information sheet for participant recruitment via your University email or online tools such as Microsoft Form and Microsoft Teams.



[School of Education, Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

**Week 2**: You will be given the Online Information Sheet and Online Consent Form in your online English class. It will take approximately 20 minutes for clarification and answering questions. If you agree to participate in this study, you have to type your name at the end of the form. Then, you will have to take an online vocabulary test via Microsoft Form for one and a half hours. You are required to turn on your camera in Microsoft Teams during the test.
**Week 3**: You will study vocabulary activities based on the contents in the course book and will be given some supplementary materials through online platforms such as Microsoft Teams. This will take approximately one and a half to two hours. You are required to turn on your camera during this class. Then, you will have to take an online vocabulary test for one and a half hours after studying. After that, one of the participants in each class will be invited to join in an online post-interview for approximately 30 minutes. The concersation between the interviewer and interviewee will be recorded during the interview.
**Week 4**: There is no data collection in this week.
**Week 5**: You will have to take an online vocabulary test for one and a half hours. You are required to turn on your camera in Microsoft Teams during the test.
**Week 6:** You will have to take an online vocabulary test for one and a half hours. You are required to turn on your camera in Microsoft Teams during the test.

### What are the possible disadvantages and risks of taking part?

There will be no risk in participating in this study. Online learning is based on the contents in the course outline and the book used in the University Engish Course II. The test results will be kept confidentially for data analysis and will not be evaluated as an extra credit in this course.

### What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this project will help the participants to enhance as well as retain their vocabulary knowledge from the learning activities and supplementary materials implemented in this experiment.

### Use, dissemination and storage of research data

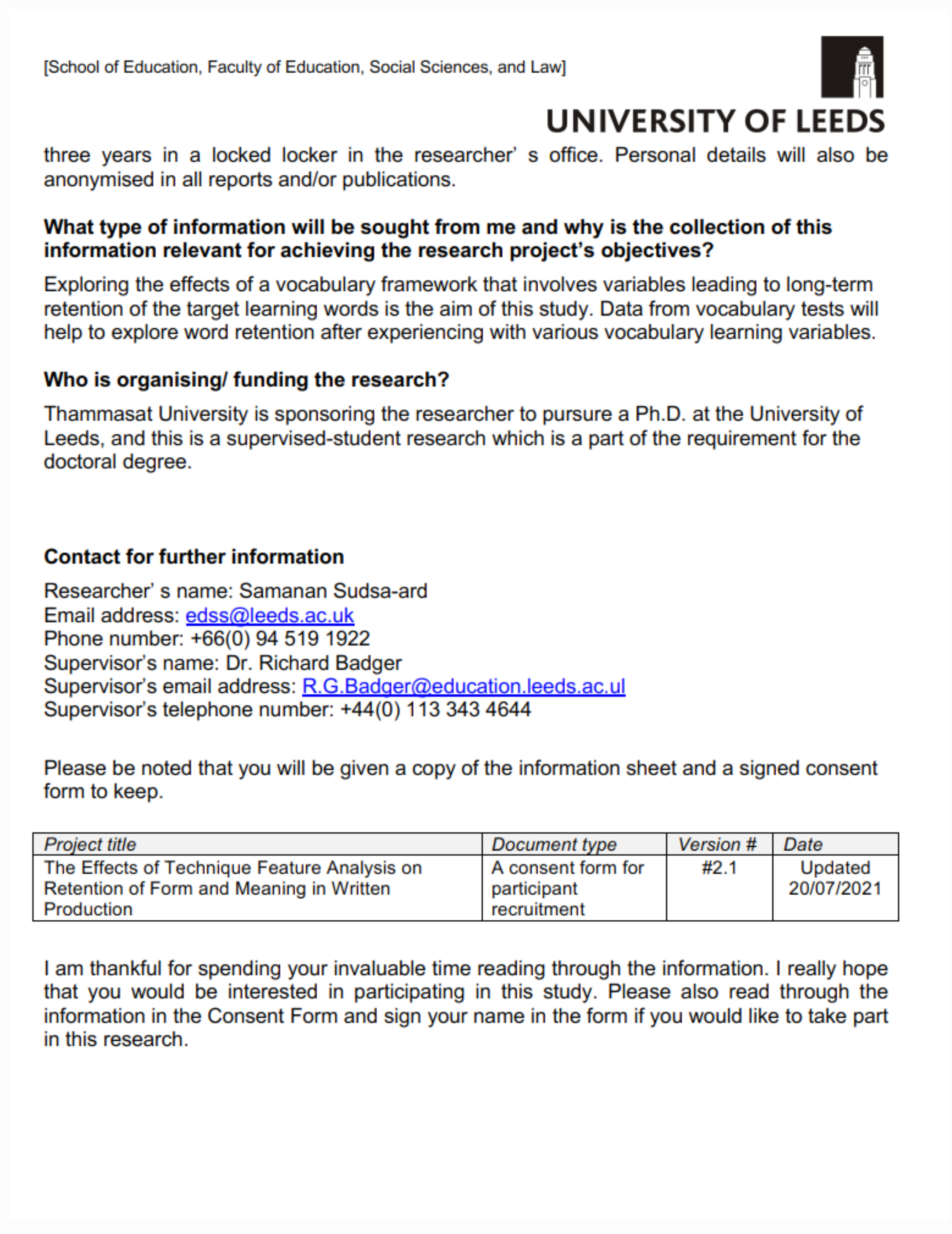
The participants' personal information, test results, and video recording will be used only for data analysis purposes. The data will be kept by the researcher confidentially for three years in a secure office. The office room will always be locked and not be accessible by other students or staff  without permission.

### What will happen to my personal information?

Because each participant has to do three tests, the participants' personal data such as names and student IDs will be collected for administrative purposes only. This information will be anonymised during data analysis procedures. Numbers will be used instead of names and IDs to report research findings.

### What will happen to the results of the research project?

All the contact information that the researcher collects about you during the course of the research will be kept strictly confidential and will be stored separately from the research data. The researcher will take steps wherever possible to anonymise the research data so that you will not be identified in any reports or publications. The collected data might be used for additional or subsequent research in the future. However, it will be kept no longer than



[School of Education, Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

three years in a locked locker in the researcher' s office. Personal details will also be anonymised in all reports and/or publications.

**What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?**

Exploring the effects of a vocabulary framework that involves variables leading to long-term retention of the target learning words is the aim of this study. Data from vocabulary tests will help to explore word retention after experiencing with various vocabulary learning variables.

**Who is organising/ funding the research?**

Thammasat University is sponsoring the researcher to pursue a Ph.D. at the University of Leeds, and this is a supervised-student research which is a part of the requirement for the doctoral degree.

**Contact for further information**

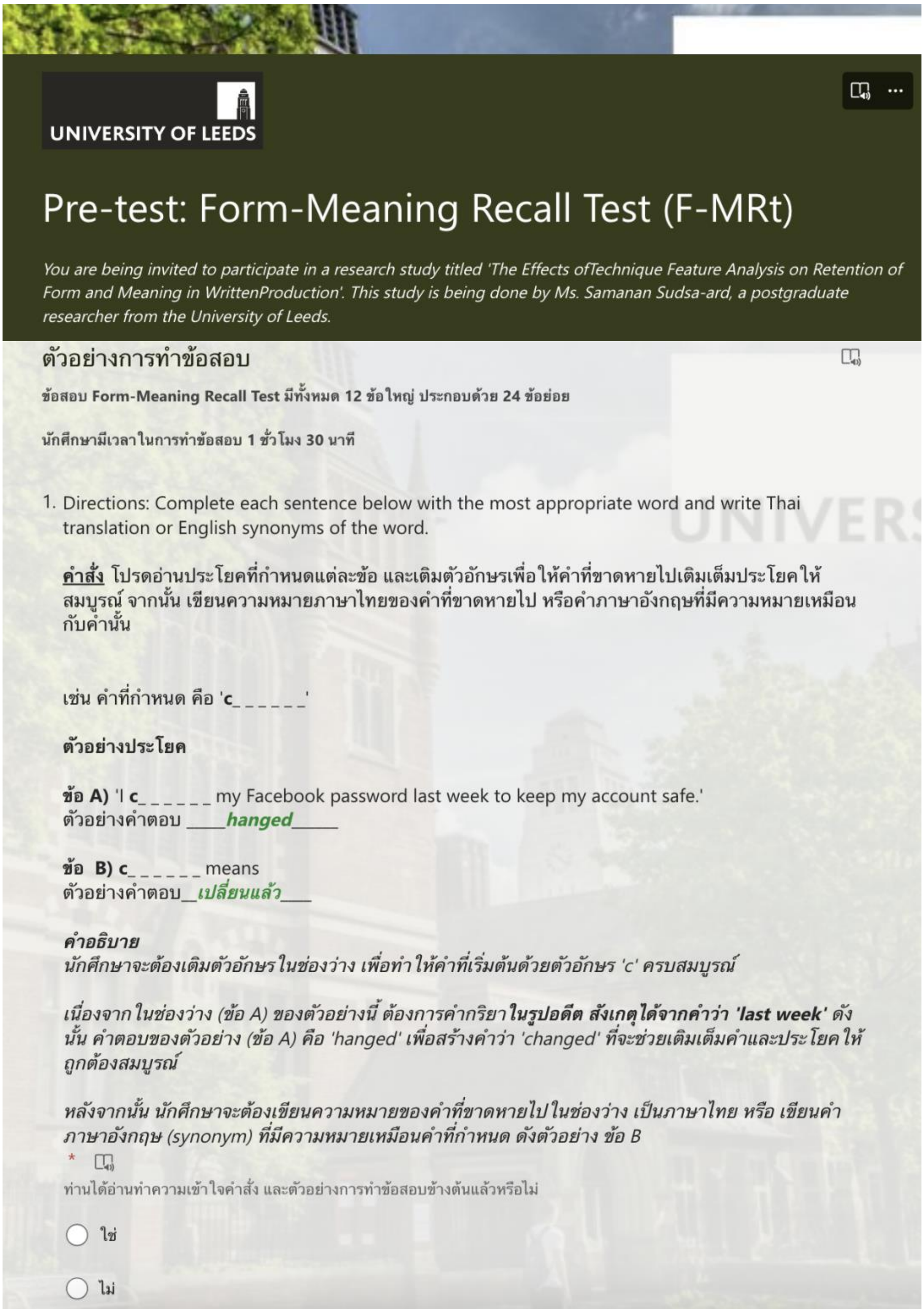
Researcher' s name: Samanan Sudsa-ard
Email address: edss@leeds.ac.uk
Phone number: +66(0) 94 519 1922
Supervisor's name: Dr. Richard Badger
Supervisor's email address: R.G.Badger@education.leeds.ac.ul
Supervisor's telephone number: +44(0) 113 343 4644

Please be noted that you will be given a copy of the information sheet and a signed consent form to keep.

| *Project title* | *Document type* | *Version #* | *Date* |
|---|---|---|---|
| The Effects of Technique Feature Analysis on Retention of Form and Meaning in Written Production | A consent form for participant recruitment | #2.1 | Updated 20/07/2021 |

I am thankful for spending your invaluable time reading through the information. I really hope that you would be interested in participating in this study. Please also read through the information in the Consent Form and sign your name in the form if you would like to take part in this research.

## Appendix 6: Form-Meaning Recall test (F-MRt)



UNIVERSITY OF LEEDS

# Pre-test: Form-Meaning Recall Test (F-MRt)

*You are being invited to participate in a research study titled 'The Effects ofTechnique Feature Analysis on Retention of Form and Meaning in WrittenProduction'. This study is being done by Ms. Samanan Sudsa-ard, a postgraduate researcher from the University of Leeds.*

ตัวอย่างการทำข้อสอบ

**ข้อสอบ Form-Meaning Recall Test มีทั้งหมด 12 ข้อใหญ่ ประกอบด้วย 24 ข้อย่อย**

**นักศึกษามีเวลาในการทำข้อสอบ 1 ชั่วโมง 30 นาที**

1. Directions: Complete each sentence below with the most appropriate word and write Thai translation or English synonyms of the word.

**คำสั่ง** โปรดอ่านประโยคที่กำหนดแต่ละข้อ และเติมตัวอักษรเพื่อให้คำที่ขาดหายไปเติมเต็มประโยคให้สมบูรณ์ จากนั้น เขียนความหมายภาษาไทยของคำที่ขาดหายไป หรือคำภาษาอังกฤษที่มีความหมายเหมือนกับคำนั้น

เช่น คำที่กำหนด คือ '**c**_ _ _ _ _ _'

**ตัวอย่างประโยค**

**ข้อ A)** 'I **c**_ _ _ _ _ _ my Facebook password last week to keep my account safe.'
ตัวอย่างคำตอบ _____***hanged***_____

**ข้อ B) c**_ _ _ _ _ _ means
ตัวอย่างคำตอบ__***เปลี่ยนแล้ว***___

***คำอธิบาย***
*นักศึกษาจะต้องเติมตัวอักษรในช่องว่าง เพื่อทำให้คำที่เริ่มต้นด้วยตัวอักษร 'c' ครบสมบูรณ์*

*เนื่องจาก ในช่องว่าง (ข้อ A) ของตัวอย่างนี้ ต้องการคำกริยา* ***ในรูปอดีต สังเกตุได้จากคำว่า 'last week'*** *ดังนั้น คำตอบของตัวอย่าง (ข้อ A) คือ 'hanged' เพื่อสร้างคำว่า 'changed' ที่จะช่วยเติมเต็มคำและประโยคให้ถูกต้องสมบูรณ์*

*หลังจากนั้น นักศึกษาจะต้องเขียนความหมายของคำที่ขาดหายไปในช่องว่าง เป็นภาษาไทย หรือ เขียนคำภาษาอังกฤษ (synonym) ที่มีความหมายเหมือนคำที่กำหนด ดังตัวอย่าง ข้อ B*

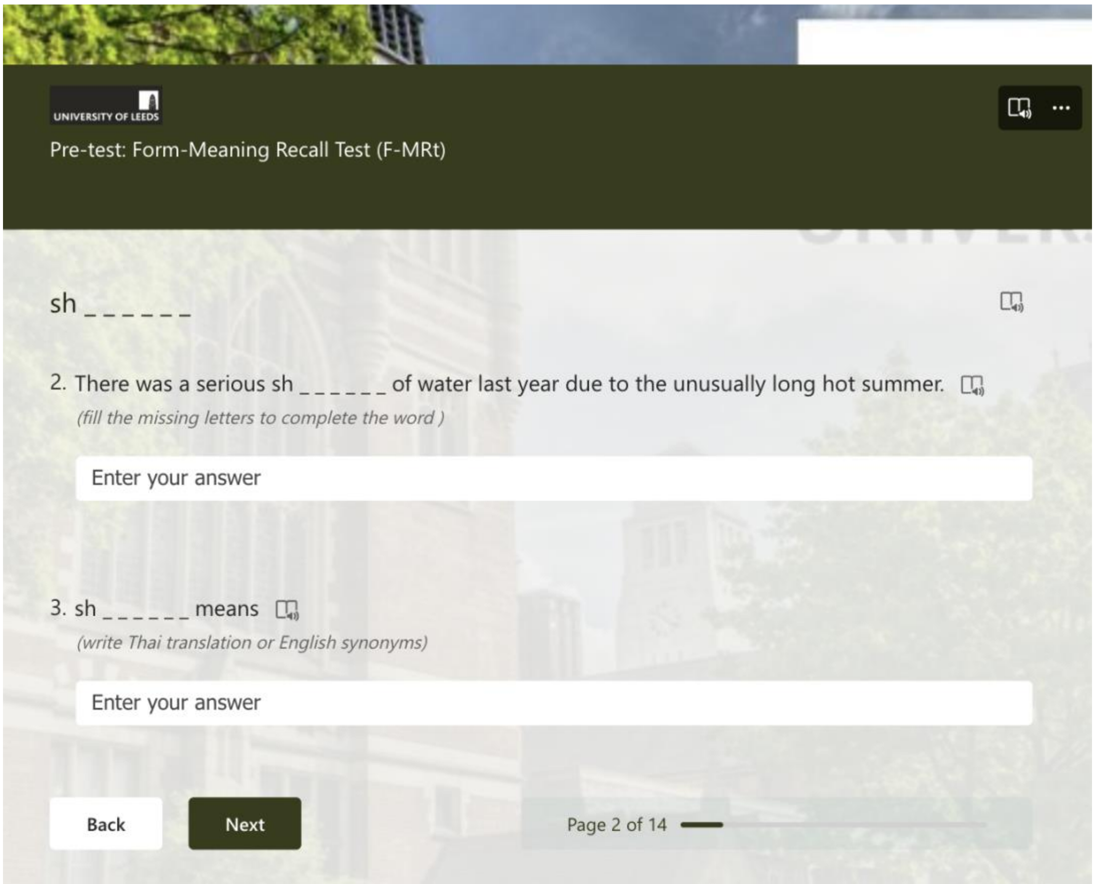
*

ท่านได้อ่านทำความเข้าใจคำสั่ง และตัวอย่างการทำข้อสอบข้างต้นแล้วหรือไม่

○ ใช่

○ ไม่

### Item 1: shortage (K1)



Pre-test: Form-Meaning Recall Test (F-MRt)

sh _ _ _ _ _ _

2. There was a serious sh _ _ _ _ _ _ of water last year due to the unusually long hot summer.

*(fill the missing letters to complete the word )*

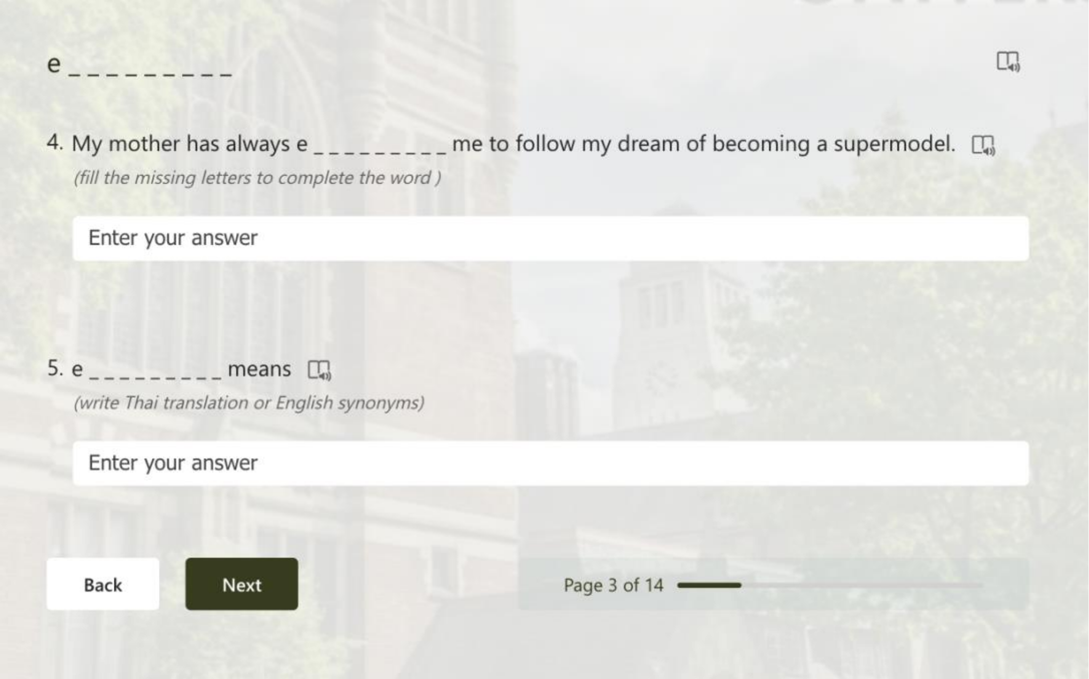
Enter your answer

3. sh _ _ _ _ _ _ means

*(write Thai translation or English synonyms)*

Enter your answer

### Item 2: encouraged (K2)



e _ _ _ _ _ _ _ _ _

4. My mother has always e _ _ _ _ _ _ _ _ _ me to follow my dream of becoming a supermodel.

*(fill the missing letters to complete the word )*

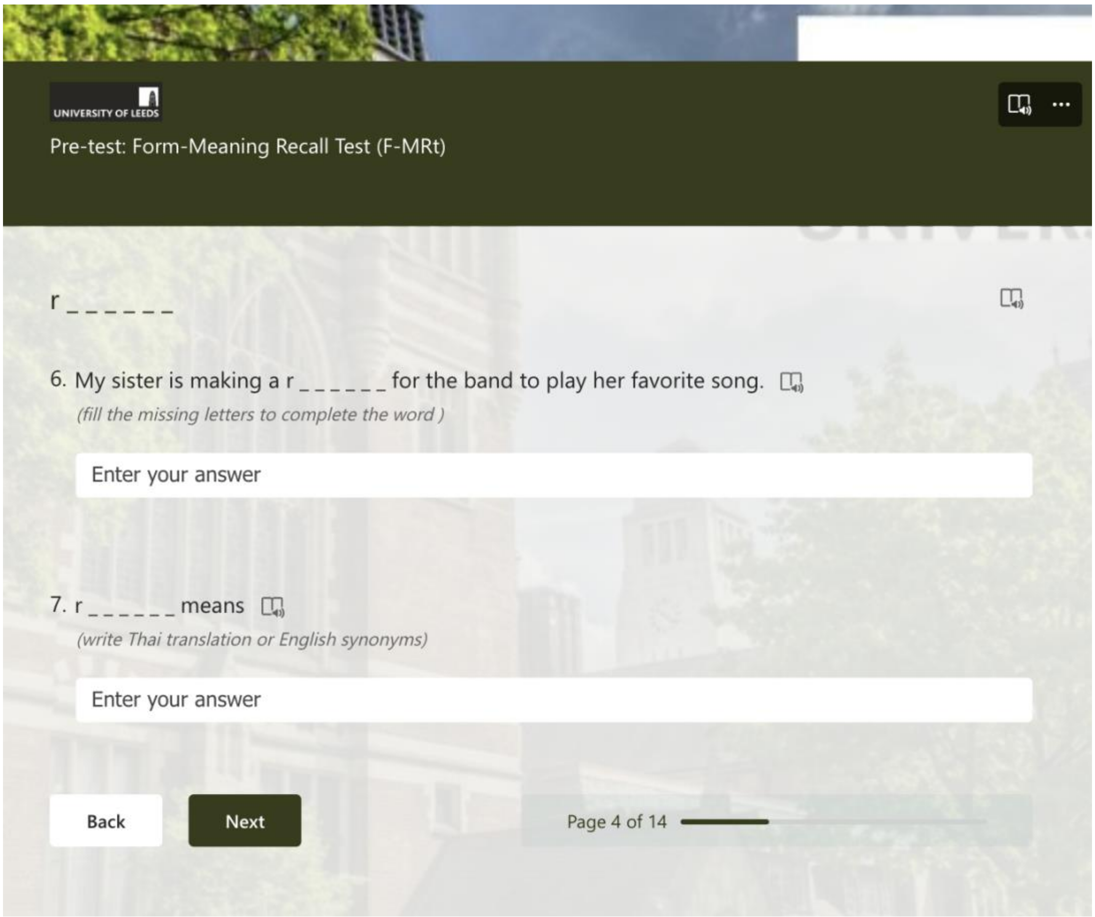
Enter your answer

5. e _ _ _ _ _ _ _ _ _ means

*(write Thai translation or English synonyms)*

Enter your answer

### Item 3: request (K3)



Pre-test: Form-Meaning Recall Test (F-MRt)

r _ _ _ _ _ _

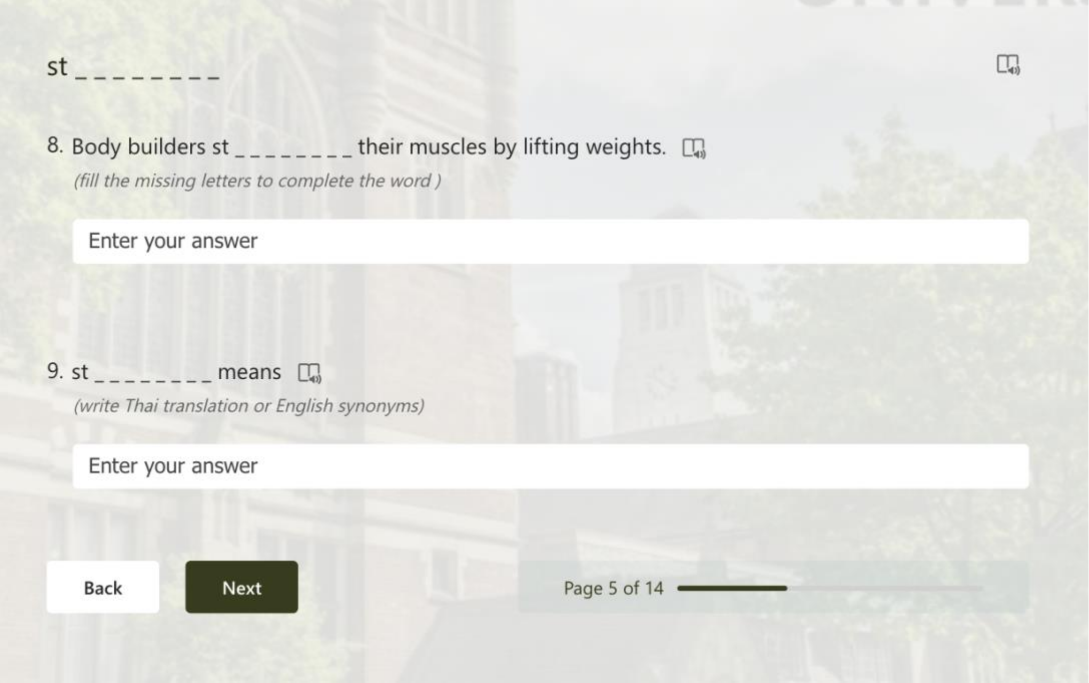
6. My sister is making a r _ _ _ _ _ _ for the band to play her favorite song.
   *(fill the missing letters to complete the word )*

   Enter your answer

7. r _ _ _ _ _ _ means
   *(write Thai translation or English synonyms)*

   Enter your answer

### Item 4: strengthen (K2)



st _ _ _ _ _ _ _ _

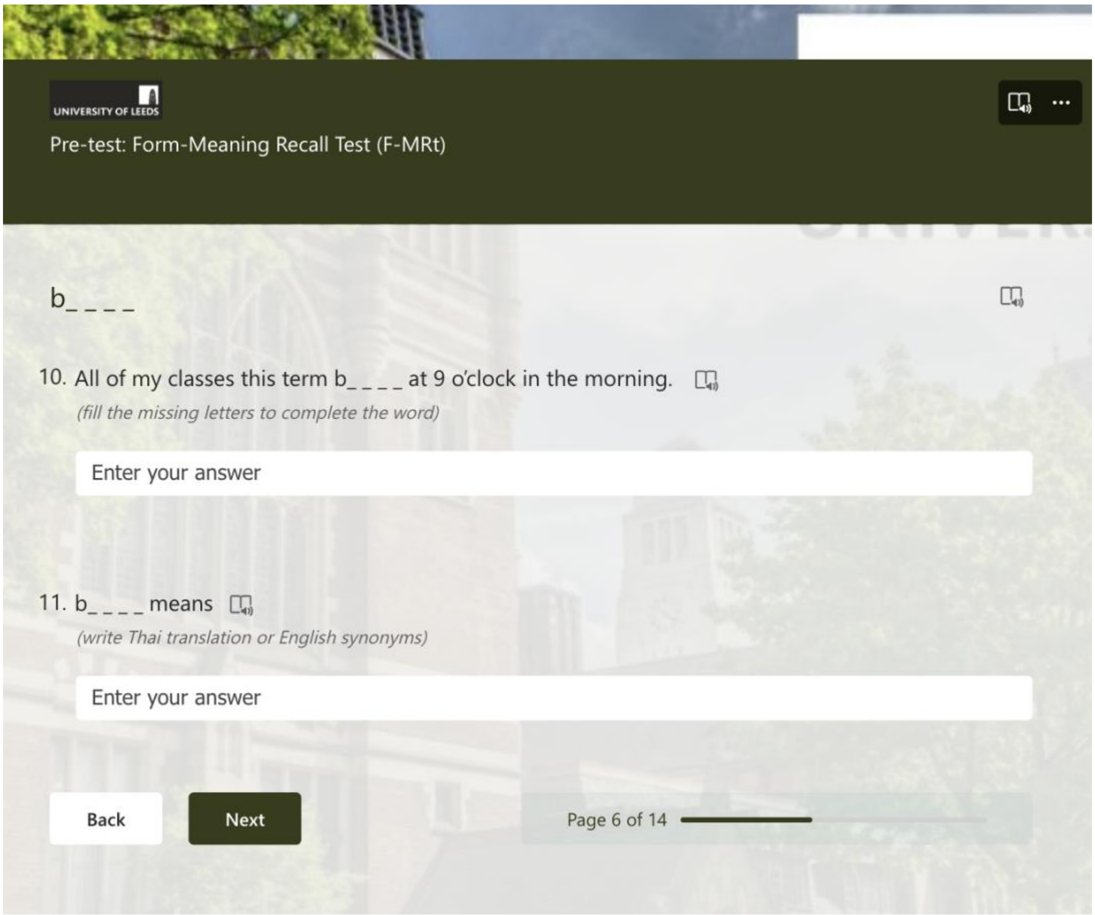
8. Body builders st _ _ _ _ _ _ _ _ their muscles by lifting weights.
   *(fill the missing letters to complete the word )*

   Enter your answer

9. st _ _ _ _ _ _ _ _ means
   *(write Thai translation or English synonyms)*

   Enter your answer

## Item 5: begins (K1, extra word)



Pre-test: Form-Meaning Recall Test (F-MRt)

b_ _ _ _

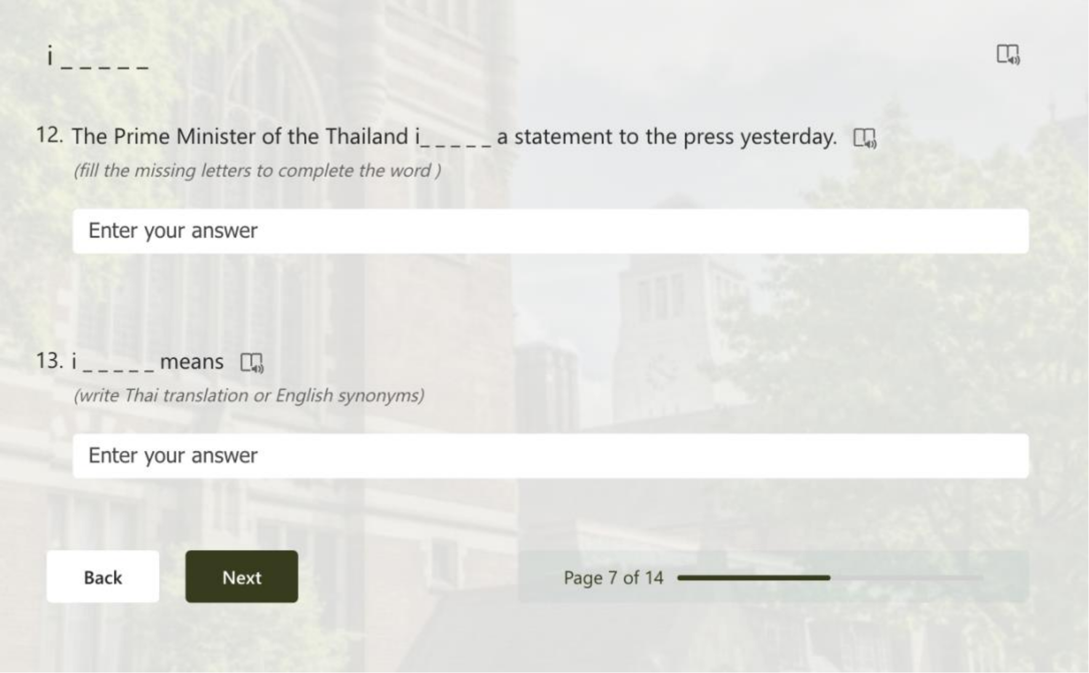
10. All of my classes this term b_ _ _ _ at 9 o'clock in the morning.
(fill the missing letters to complete the word)

Enter your answer

11. b_ _ _ _ means
(write Thai translation or English synonyms)

Enter your answer

## Item 6: issued (K1)



i _ _ _ _ _

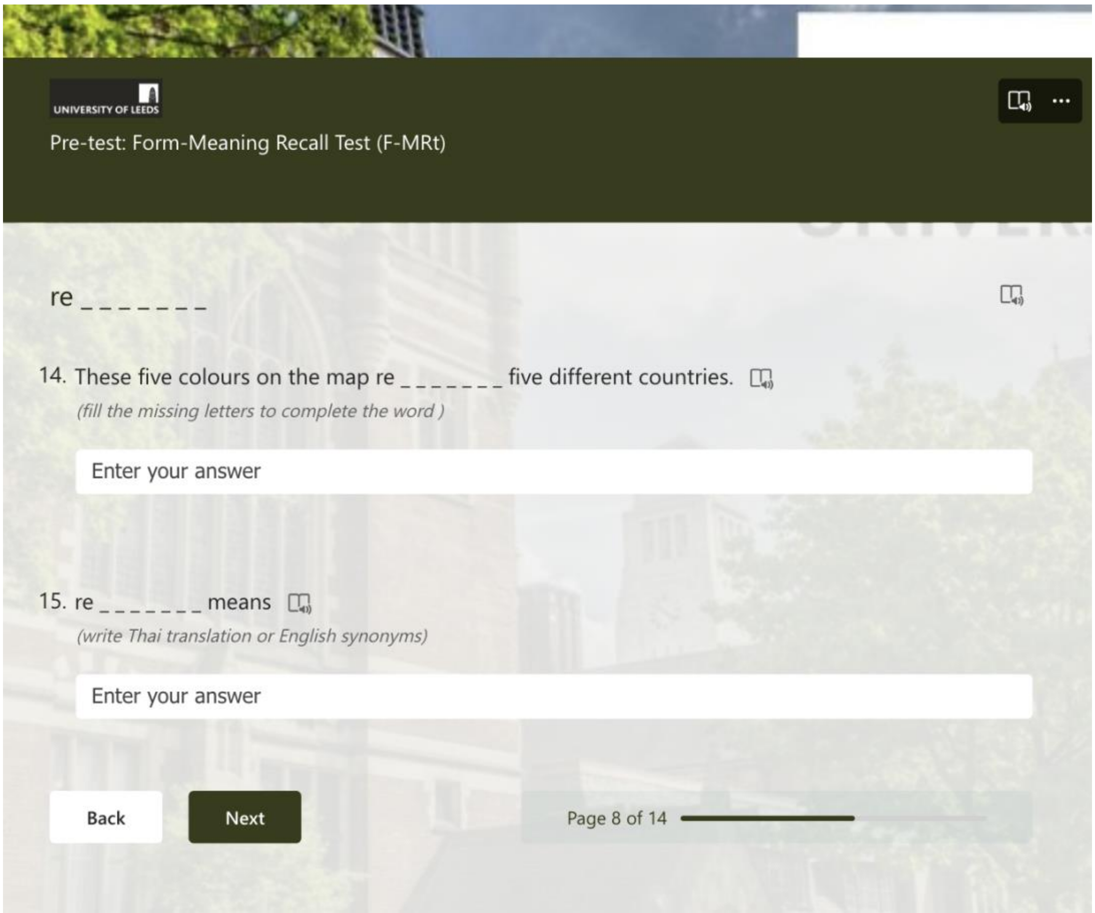
12. The Prime Minister of the Thailand i_ _ _ _ _ a statement to the press yesterday.
(fill the missing letters to complete the word )

Enter your answer

13. i _ _ _ _ _ means
(write Thai translation or English synonyms)

Enter your answer

### Item 7: represent (K2)



UNIVERSITY OF LEEDS

Pre-test: Form-Meaning Recall Test (F-MRt)

re _ _ _ _ _ _ _

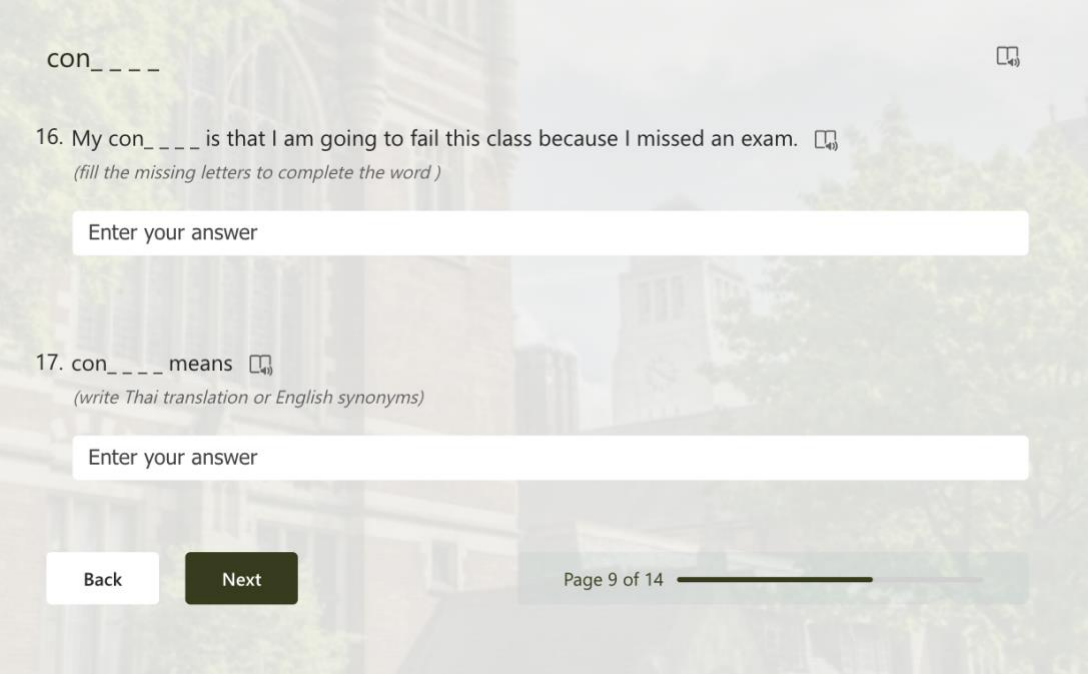
14. These five colours on the map re _ _ _ _ _ _ _ five different countries.
*(fill the missing letters to complete the word )*

Enter your answer

15. re _ _ _ _ _ _ _ means
*(write Thai translation or English synonyms)*

Enter your answer

### Item 8: concern (K1)



con_ _ _ _

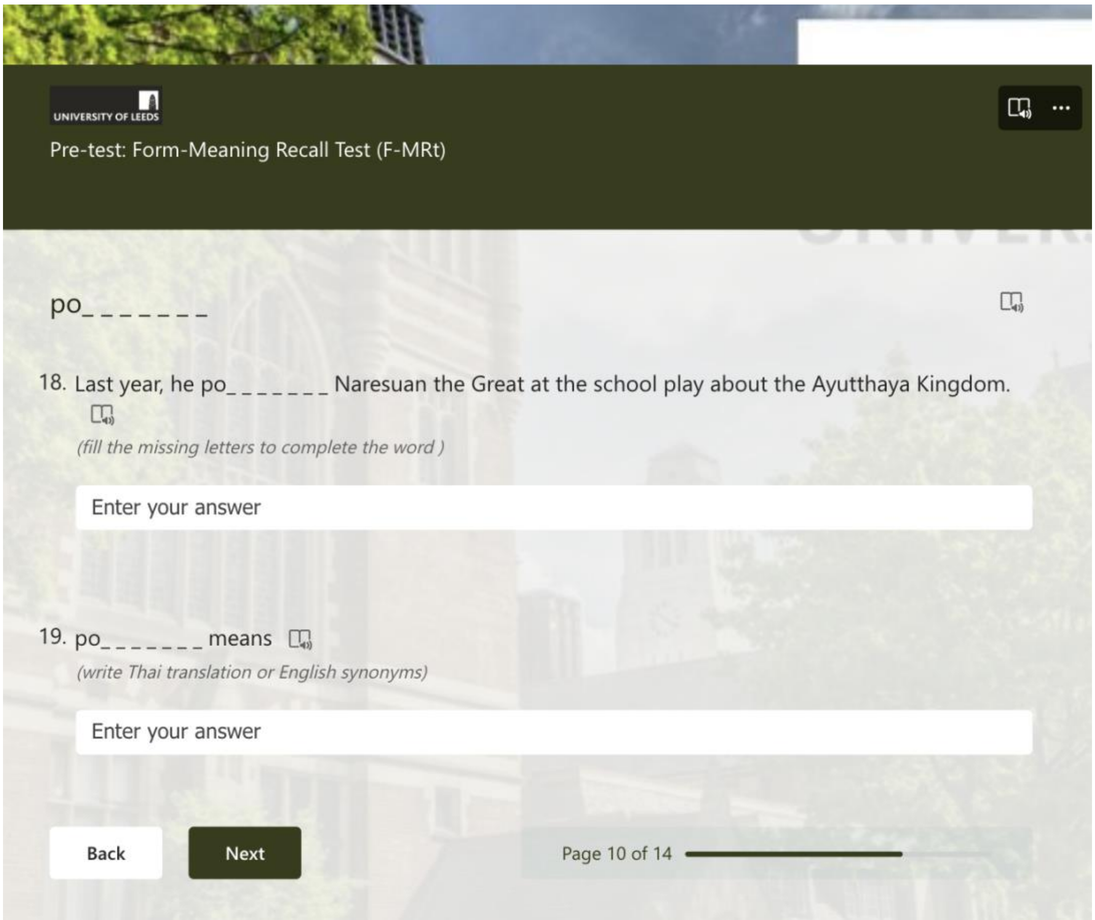
16. My con_ _ _ _ is that I am going to fail this class because I missed an exam.
*(fill the missing letters to complete the word )*

Enter your answer

17. con_ _ _ _ means
*(write Thai translation or English synonyms)*

Enter your answer

### Item 9: portrayed (K4)



Pre-test: Form-Meaning Recall Test (F-MRt)

po_ _ _ _ _ _ _

18. Last year, he po_ _ _ _ _ _ _ Naresuan the Great at the school play about the Ayutthaya Kingdom.

*(fill the missing letters to complete the word )*

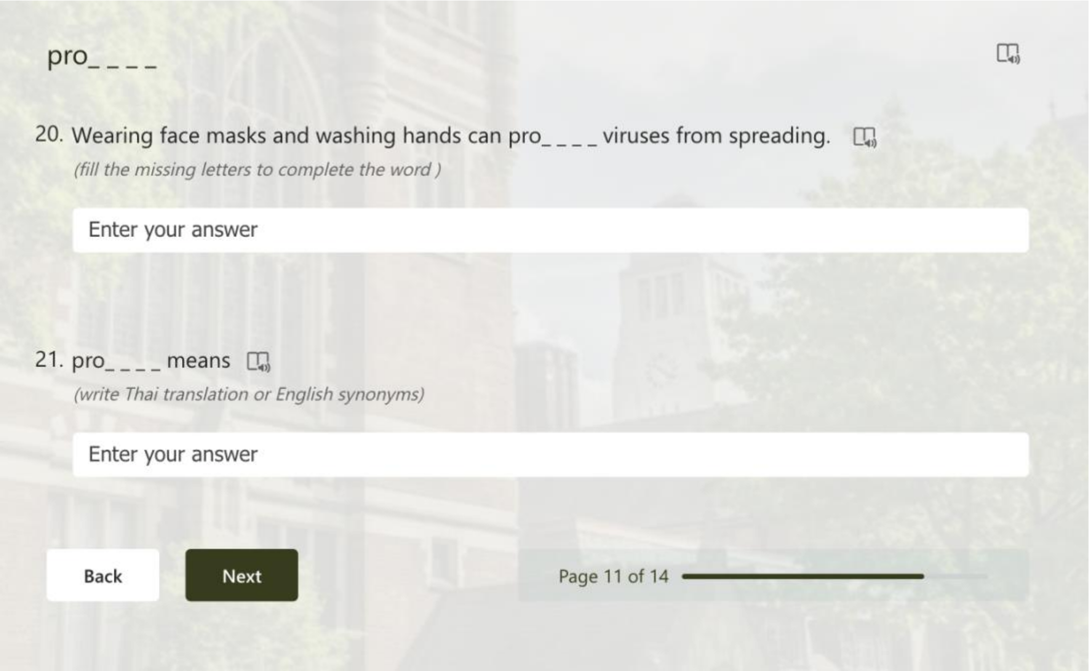
Enter your answer

19. po_ _ _ _ _ _ _ means

*(write Thai translation or English synonyms)*

Enter your answer

### Item 10: protect (K1, extra word)



pro_ _ _ _

20. Wearing face masks and washing hands can pro_ _ _ _ viruses from spreading.

*(fill the missing letters to complete the word )*

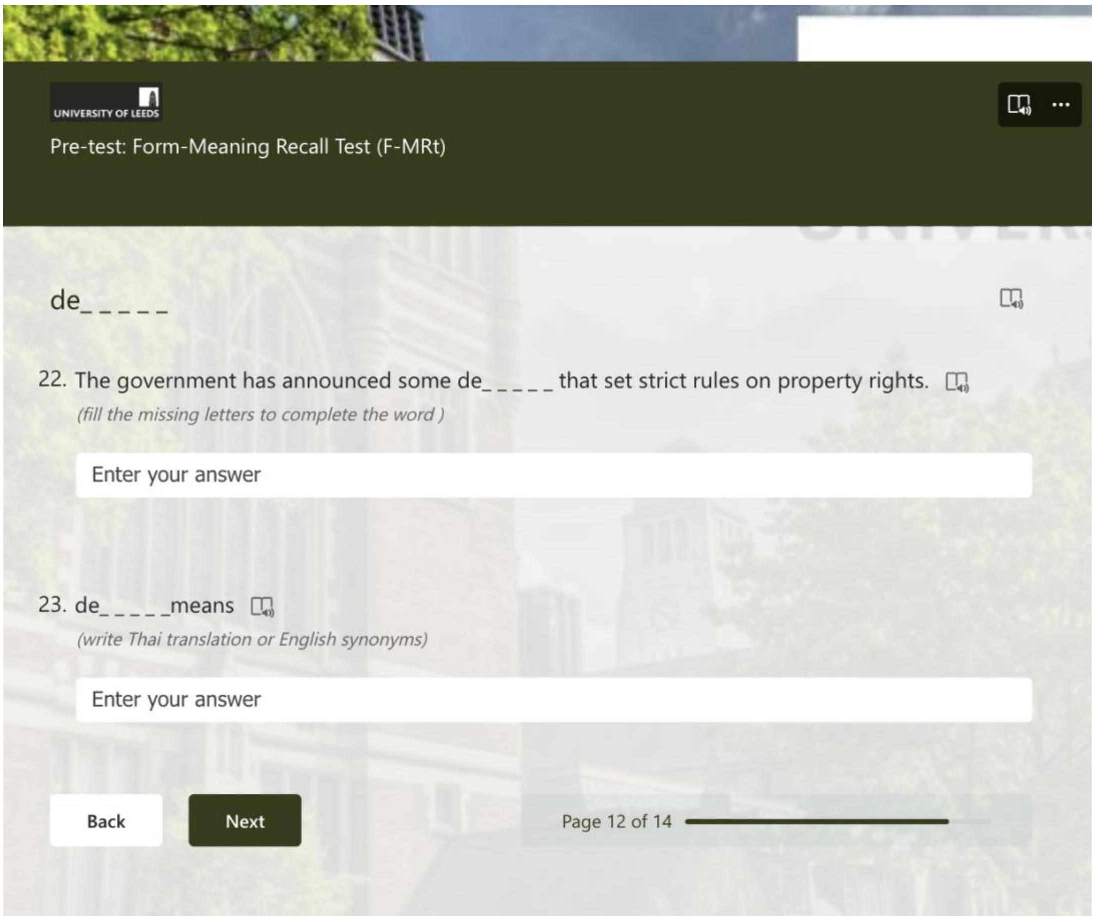
Enter your answer

21. pro_ _ _ _ means

*(write Thai translation or English synonyms)*

Enter your answer

## Item 11: decrees (K5)



Pre-test: Form-Meaning Recall Test (F-MRt)

de_ _ _ _ _

22. The government has announced some de_ _ _ _ _ that set strict rules on property rights.

*(fill the missing letters to complete the word )*

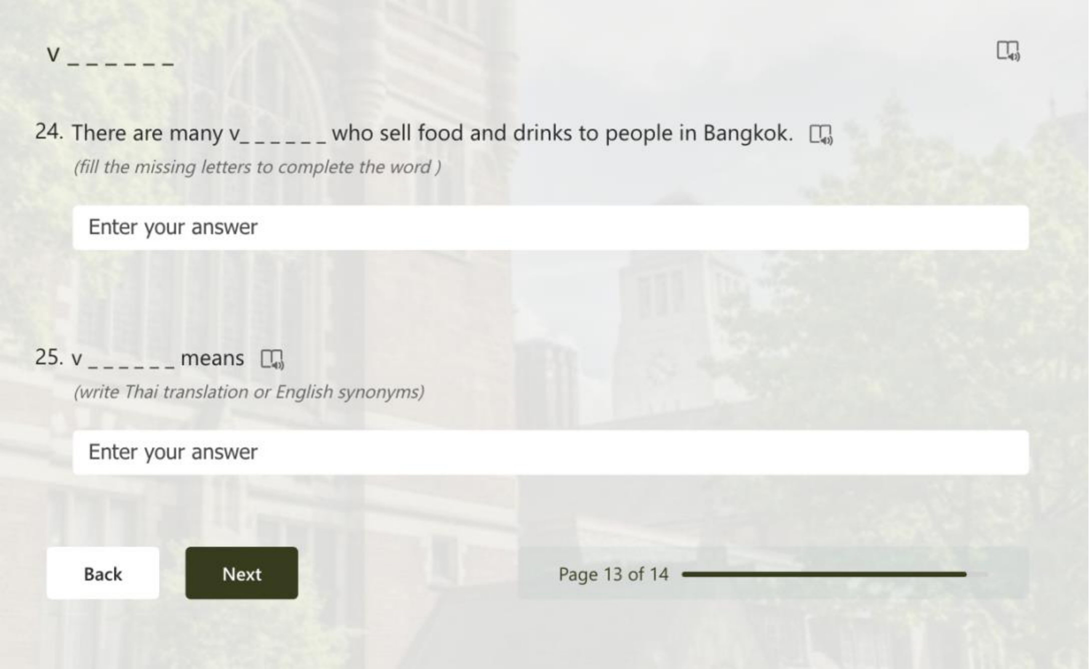
Enter your answer

23. de_ _ _ _ _means

*(write Thai translation or English synonyms)*

Enter your answer

## Item 12: vendors (K4)



v _ _ _ _ _ _

24. There are many v_ _ _ _ _ _ who sell food and drinks to people in Bangkok.

*(fill the missing letters to complete the word )*

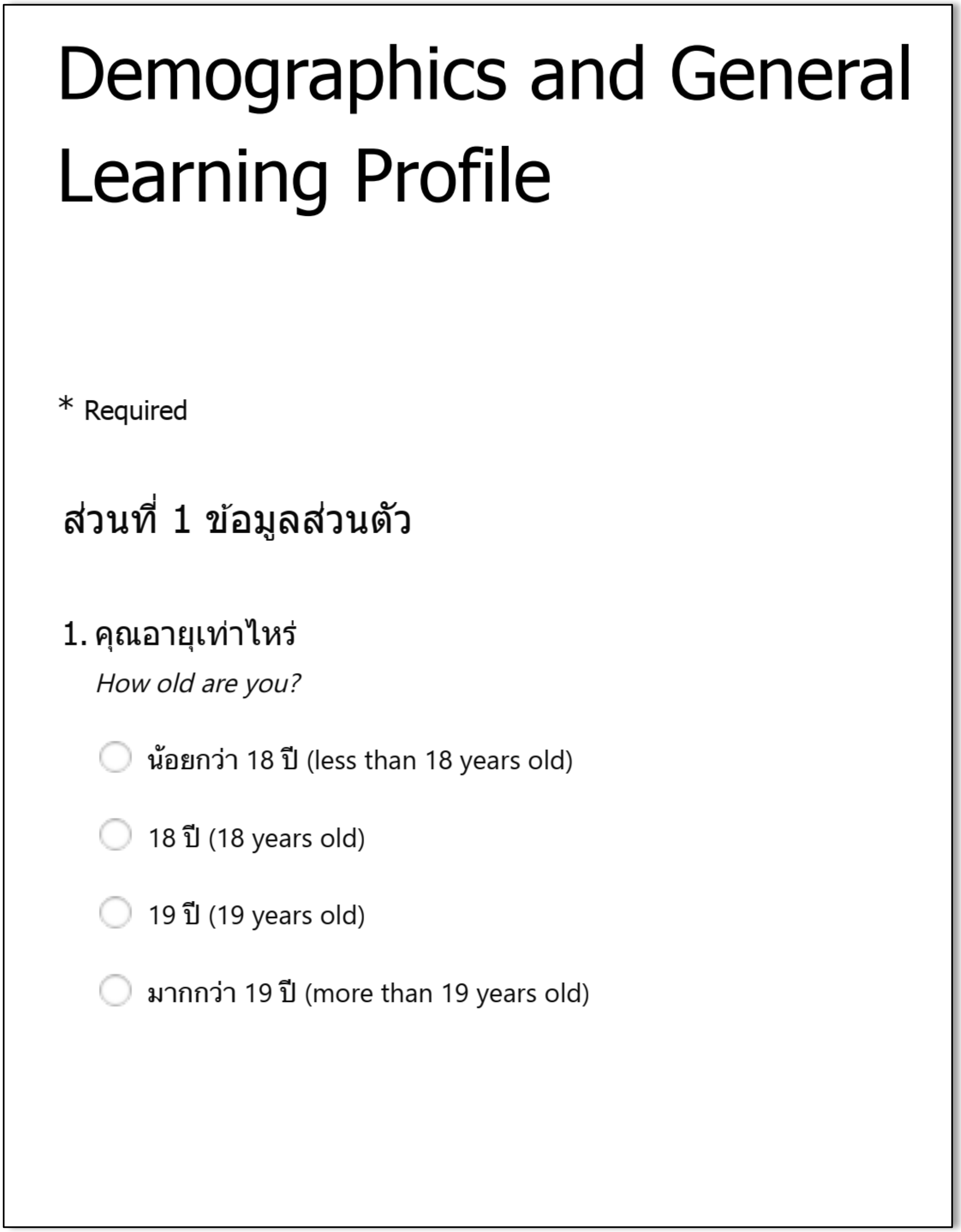
Enter your answer

25. v _ _ _ _ _ _ means

*(write Thai translation or English synonyms)*

Enter your answer

## Appendix 7: Questionnaire (bilingual version)



# Demographics and General Learning Profile

* Required

## ส่วนที่ 1 ข้อมูลส่วนตัว

1. คุณอายุเท่าไหร่

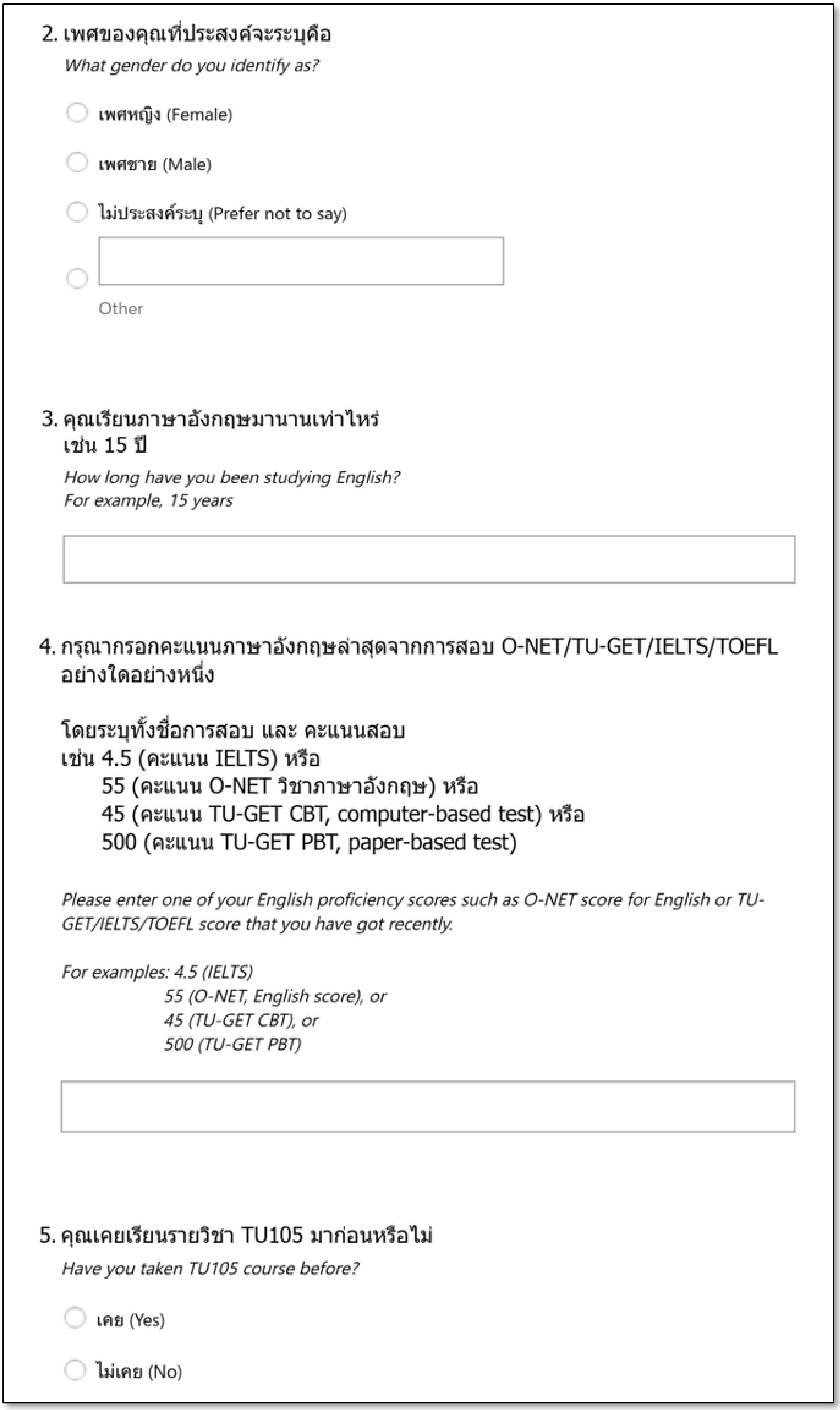
*How old are you?*

○ น้อยกว่า 18 ปี (less than 18 years old)

○ 18 ปี (18 years old)

○ 19 ปี (19 years old)

○ มากกว่า 19 ปี (more than 19 years old)



2. เพศของคุณที่ประสงค์จะระบุคือ

*What gender do you identify as?*

○ เพศหญิง (Female)

○ เพศชาย (Male)

○ ไม่ประสงค์ระบุ (Prefer not to say)

○ Other

3. คุณเรียนภาษาอังกฤษมานานเท่าไหร่
เช่น 15 ปี

*How long have you been studying English?*
*For example, 15 years*

4. กรุณากรอกคะแนนภาษาอังกฤษล่าสุดจากการสอบ O-NET/TU-GET/IELTS/TOEFL อย่างใดอย่างหนึ่ง

โดยระบุทั้งชื่อการสอบ และ คะแนนสอบ
เช่น 4.5 (คะแนน IELTS) หรือ
55 (คะแนน O-NET วิชาภาษาอังกฤษ) หรือ
45 (คะแนน TU-GET CBT, computer-based test) หรือ
500 (คะแนน TU-GET PBT, paper-based test)

*Please enter one of your English proficiency scores such as O-NET score for English or TU-GET/IELTS/TOEFL score that you have got recently.*

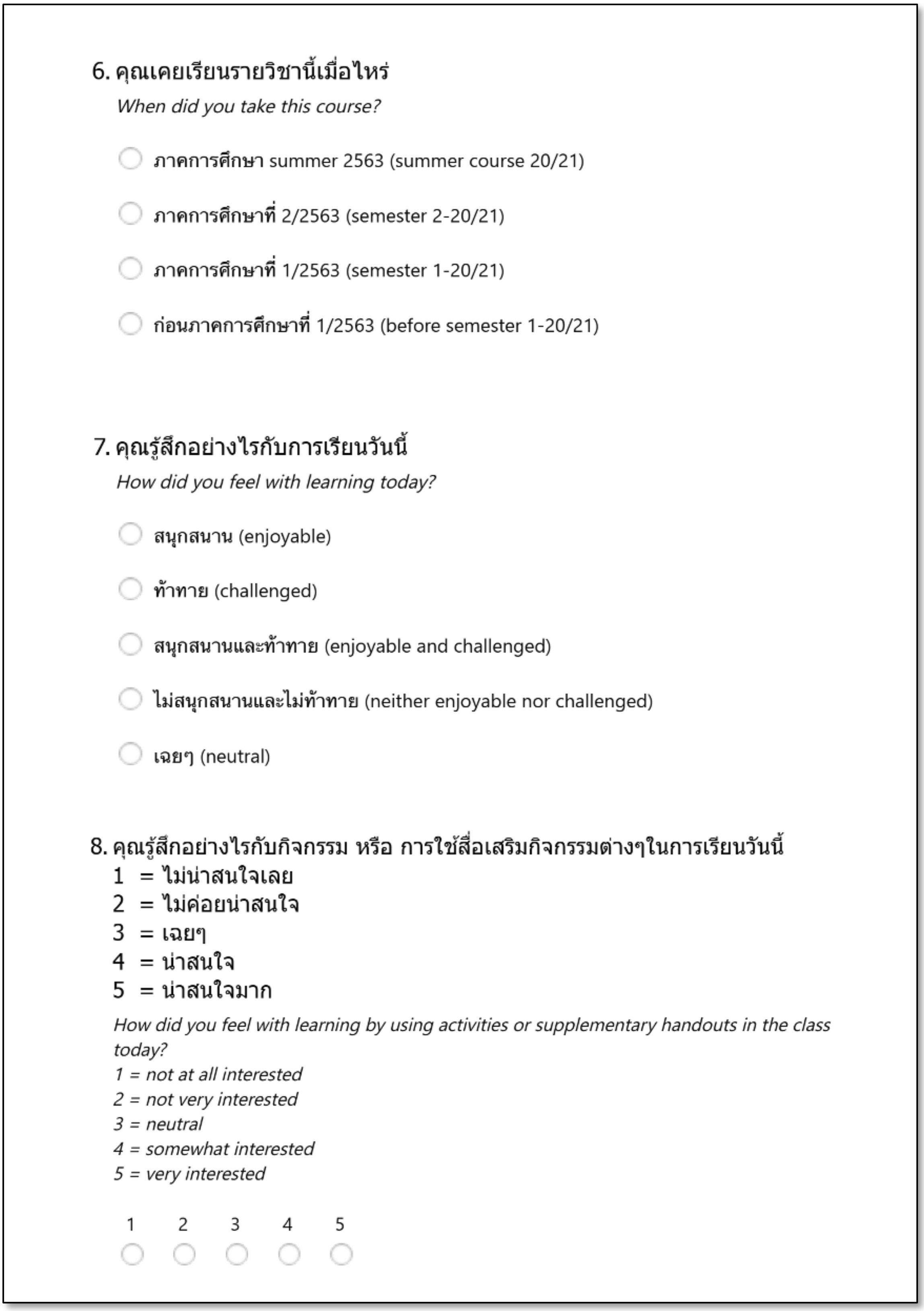
*For examples: 4.5 (IELTS)*
*55 (O-NET, English score), or*
*45 (TU-GET CBT), or*
*500 (TU-GET PBT)*

5. คุณเคยเรียนรายวิชา TU105 มาก่อนหรือไม่

*Have you taken TU105 course before?*

○ เคย (Yes)

○ ไม่เคย (No)



6. คุณเคยเรียนรายวิชานี้เมื่อไหร่

*When did you take this course?*

○ ภาคการศึกษา summer 2563 (summer course 20/21)

○ ภาคการศึกษาที่ 2/2563 (semester 2-20/21)

○ ภาคการศึกษาที่ 1/2563 (semester 1-20/21)

○ ก่อนภาคการศึกษาที่ 1/2563 (before semester 1-20/21)

7. คุณรู้สึกอย่างไรกับการเรียนวันนี้

*How did you feel with learning today?*

○ สนุกสนาน (enjoyable)

○ ท้าทาย (challenged)

○ สนุกสนานและท้าทาย (enjoyable and challenged)

○ ไม่สนุกสนานและไม่ท้าทาย (neither enjoyable nor challenged)

○ เฉยๆ (neutral)

8. คุณรู้สึกอย่างไรกับกิจกรรม หรือ การใช้สื่อเสริมกิจกรรมต่างๆในการเรียนวันนี้
1 = ไม่น่าสนใจเลย
2 = ไม่ค่อยน่าสนใจ
3 = เฉยๆ
4 = น่าสนใจ
5 = น่าสนใจมาก

*How did you feel with learning by using activities or supplementary handouts in the class today?*
*1 = not at all interested*
*2 = not very interested*
*3 = neutral*
*4 = somewhat interested*
*5 = very interested*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ |

## Appendix 8a: Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest sensitive scores of the experimental groups (*Noticing* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.98 | 0.30 | [0.40, 1.55] | 620 | 3.25 | **< .001** |
| Group (Noticing vs. Motivation) | -0.34 | 0.41 | [-1.15, 0.47] | 620 | -0.83 | 0.42 |
| Group (Noticing vs. Retrieval) | 0.06 | 0.42 | [-0.76, 0.87] | 620 | 0.13 | 0.89 |
| Group (Noticing vs. Generative Use) | -0.14 | 0.41 | [-0.95, 0.67] | 620 | -0.34 | 0.74 |
| Group (Noticing vs. All TFA Components) | -0.07 | 0.40 | [-0.86, 0.72] | 620 | -0.17 | 0.87 |
| Time (Pre-test vs. Immediate Posttest) | 6.76 | 0.40 | [5.98, 7.54] | 418 | 16.99 | **< .001** |
| Group (Noticing vs. Motivation): Time (Pre-test vs. Immediate Posttest) | -1.80 | 0.56 | [-2.90, -0.69] | 418 | -3.20 | **0.001** |
| Group (Noticing vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | -1.12 | 0.57 | [-2.23, -0.01] | 418 | -1.98 | 0.06 |
| Group (Noticing vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | -0.23 | 0.56 | [-1.34,0.87] | 418 | -0.41 | 0.68 |
| Group (Noticing vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 0.52 | 0.55 | [-0.56, 1.59] | 418 | 0.94 | 0.35 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 8a: Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest sensitive scores of the experimental groups (*Retrieval* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 1.03 | 0.30 | [0.45, 1.61] | 620 | 3.49 | **< .001** |
| Group (Retrieval vs. Motivation) | -0.40 | 0.42 | [-1.21, 0.42] | 620 | -0.96 | 0.34 |
| Group (Retrieval vs. Noting) | -0.06 | 0.42 | [-0.87, 0.76] | 620 | -0.13 | 0.89 |
| Group (Retrieval vs. Generative Use) | -0.20 | 0.42 | [-1.01, 0.62] | 620 | -0.47 | 0.64 |
| Group (Retrieval vs. All TFA Components) | -0.12 | 0.42 | [-0.92, 0.67] | 620 | -0.31 | 0.76 |
| Time (Pre-test vs. Immediate Posttest) | 5.64 | 0.40 | [4.85, 6.43] | 418 | 14.01 | **< .001** |
| Group (Retrieval vs. Motivation): Time (Pre-test vs. Immediate Posttest) | -0.68 | 0.57 | [1.79, 0.43] | 418 | -1.20 | 0.23 |
| Group (Retrieval vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 1.12 | 0.57 | [0.01, 2.23] | 418 | 1.98 | **0.05** |
| Group (Retrieval vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | 0.89 | 0.57 | [-0.22, 1.99] | 418 | 1.57 | 0.12 |
| Group (Retrieval vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 1.63 | 0.55 | [0.55, 2.71] | 418 | 2.97 | **< .01** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 8a: Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest sensitive scores of the experimental groups (*Generative Use* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.84 | 0.29 | [0.26, 1.41] | 620 | 2.86 | **< .001** |
| Group (Generative Use vs. Motivation) | -0.20 | 0.41 | [-1.01, 0.61] | 620 | -0.49 | 0.63 |
| Group (Generative Use vs. Noticing) | 0.14 | 0.41 | [-0.67, 0.95] | 620 | 0.34 | 0.73 |
| Group (Generative Use vs. Retrieval) | 0.20 | 0.42 | [-0.62, 1.01] | 620 | 0.47 | 0.6 |
| Group (Generative Use vs. All TFA Components) | 0.07 | 0.40 | [-0.72, 0.86] | 620 | 0.18 | 0.86 |
| Time (Pre-test vs. Immediate Posttest) | 6.52 | 0.39 | [5.74, 7.30] | 418 | 16.41 | **< .001** |
| Group (Generative Use vs. Motivation): Time (Pre-test vs. Immediate Posttest) | -1.57 | 0.56 | [-2.67,] -0.46 | 418 | -2.79 | **< .01** |
| Group (Generative Use vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 0.23 | 0.56 | [-0.87, 1.34] | 418 | 0.41 | 0.68 |
| Group (Generative Use vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | -0.89 | 0.57 | [-1.99,0.22] | 418 | -1.57 | 0.12 |
| Group (Generative Use vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 0.75 | 0.55 | [-0.33,1.82] | 418 | 1.37 | 0.17 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 8b: Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest strict scores of the experimental groups (*Noticing* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 6.10 | 3.14 | [-0.01, 1.23] | 620 | 1.94 | **0.05** |
| Group (Noticing vs. Motivation) | 2.44 | 4.44 | [-0.85, 0.90] | 620 | 0.06 | 0.96 |
| Group (Noticing vs. Retrieval) | 2.90 | 4.47 | [-0.59, 1.17] | 620 | 0.65 | 0.52 |
| Group (Noticing vs. Generative Use) | 1.19 | 4.44 | [-0.87, 0.87] | 620 | 0.00 | 1.00 |
| Group (Noticing vs. All TFA Components) | 2.38 | 4.32 | [-0.61, 1.09] | 620 | 0.55 | 0.58 |
| Time (Pre-test vs. Immediate Posttest) | 7.00 | 4.27 | [6.16, 7.84] | 418 | 16.41 | **< .001** |
| Group (Noticing vs. Motivation): Time (Pre-test vs. Immediate Posttest) | -2.78 | 6.03 | [-3.97, -1.60] | 418 | -4.61 | **< .001** |
| Group (Noticing vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | -1.40 | 6.07 | [2.59, -0.21] | 418 | -2.31 | **0.02** |
| Group (Noticing vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | -5.61 | 6.03 | [-1.75,0.62] | 418 | -0.93 | 0.35 |
| Group (Noticing vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 8.69 | 5.87 | [-1.07, 1.24] | 418 | 0.15 | 0.88 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 8b: Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest strict scores of the experimental groups (*Retrieval* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.90 | 0.32 | [0.28, 1.52] | 620 | 2.83 | **< .01** |
| Group (Retrieval vs. Motivation) | -0.27 | 0.45 | [-1.14, 0.61] | 620 | -0.60 | 0.55 |
| Group (Retrieval vs. Noting) | -0.29 | 0.45 | [-1.17, 0.59] | 620 | -0.65 | 0.52 |
| Group (Retrieval vs. Generative Use) | -0.29 | 0.45 | [-1.17, 0.59] | 620 | -0.65 | 0.52 |
| Group (Retrieval vs. All TFA Components) | -0.05 | 0.43 | [-0.90, 0.80] | 620 | -0.12 | 0.90 |
| Time (Pre-test vs. Immediate Posttest) | 5.60 | 0.43 | [4.75, 6.45] | 418 | 12.97 | **< .001** |
| Group (Retrieval vs. Motivation): Time (Pre-test vs. Immediate Posttest) | -1.38 | 0.61 | [-2.57, -0.19] | 418 | -2.27 | **< .01** |
| Group (Retrieval vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 1.40 | 0.61 | [0.21, 2.59] | 418 | 2.31 | **0.02** |
| Group (Retrieval vs. Generative Use): Time (Pre-test vs. Immediate Posttest) | 0.84 | 0.61 | [-0.35, 2.03] | 418 | 1.38 | 0.17 |
| Group (Retrieval vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 1.49 | 0.59 | [0.33, 2.65] | 418 | 2.52 | **0.01** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 8b: Linear Mixed-effects Model comparing the Pre-test and Immediate Posttest strict scores of the experimental groups (*Generative Use* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 6.10 | 3.14 | [0.01, 1.23] | 620 | 1.94 | **0.05** |
| Group (Generative Use vs. Motivation) | 2.44 | 4.44 | [-0.85, 0.90] | 620 | 0.06 | 0.96 |
| Group (Generative Use vs. Noticing) | -2.05 | 4.44 | [-0.87, 0.87] | 620 | 0.00 | 1.00 |
| Group (Generative Use vs. Retrieval) | 2.90 | 4.47 | [-0.58, 1.17] | 620 | 0.65 | 0.51 |
| Group (Generative Use vs. All TFA Components) | 2.38 | 4.32 | [-0.61, 1.09] | 620 | 0.55 | 0.57 |
| Time (Pre-test vs. Immediate Posttest) | 6.44 | 4.27 | [5.00, 6.67] | 418 | 15.10 | **< .001** |
| Group (Generative Use vs. Motivation): Time (Pre-test vs. Immediate Posttest) | -2.22 | 6.03 | [-3.40, -1.03] | 418 | -3.68 | **< .001** |
| Group (Generative Use vs. Noticing): Time (Pre-test vs. Immediate Posttest) | 5.61 | 6.03 | [-0.62,1.75] | 418 | 0.93 | 0.35 |
| Group (Generative Use vs. Retrieval): Time (Pre-test vs. Immediate Posttest) | -8.39 | 6.07 | [-2.03, 0.35] | 418 | -1.38 | 0.16 |
| Group (Generative Use vs. All TFA Components): Time (Pre-test vs. Immediate Posttest) | 6.48 | 5.87 | [-0.50, 1.80] | 418 | 1.11 | 0.26 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 9a: Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest sensitive scores of the experimental groups (*Noticing* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.98 | 0.30 | [0.40, 1.55] | 620 | 3.25 | **< .001** |
| Group (Noticing vs. Motivation) | -0.34 | 0.41 | [-1.15, 0.47] | 620 | -0.83 | 0.42 |
| Group (Noticing vs. Retrieval) | 0.06 | 0.42 | [-0.76, 0.87] | 620 | 0.13 | 0.89 |
| Group (Noticing vs. Generative Use) | -0.14 | 0.41 | [-0.95, 0.67] | 620 | -0.34 | 0.74 |
| Group (Noticing vs. All TFA Components) | -0.07 | 0.40 | [-0.86, 0.72] | 620 | -0.17 | 0.87 |
| Time (Pre-test vs. Delayed Posttest) | 6.86 | 0.40 | [6.08, 7.64] | 418 | 17.26 | **< .001** |
| Group (Noticing vs. Motivation): Time (Pre-test vs. Delayed Posttest) | -1.98 | 0.56 | [-3.09, -0.88] | 418 | -3.53 | **< .01** |
| Group (Noticing vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 0.34 | 0.57 | [-0.77, 1.45] | 418 | 0.60 | 0.55 |
| Group (Noticing vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | -0.99 | 0.56 | [-2.10, 0.11] | 418 | -1.77 | 0.08 |
| Group (Noticing vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 0.12 | 0.57 | [-0.95,1.20] | 418 | 0.23 | 0.82 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 9a: Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest sensitive scores of the experimental groups (*Retrieval* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 1.03 | 0.30 | [0.45, 1.61] | 620 | 3.49 | **< .001** |
| Group (Retrieval vs. Motivation) | -0.40 | 0.42 | [-1.21, 0.42] | 620 | -0.96 | 0.34 |
| Group (Retrieval vs. Noting) | -0.06 | 0.42 | [-0.87, 0.76] | 620 | -0.13 | 0.89 |
| Group (Retrieval vs. Generative Use) | -0.20 | 0.42 | [-1.01, 0.62] | 620 | -0.47 | 0.64 |
| Group (Retrieval vs. All TFA Components) | -0.12 | 0.42 | [-0.92, 0.67] | 620 | -0.31 | 0.76 |
| Time (Pre-test vs. Delayed Posttest) | 7.20 | 0.40 | [6.41, 7.99] | 418 | 17.90 | **< .001** |
| Group (Retrieval vs. Motivation): Time (Pre-test vs. Delayed Posttest) | -2.32 | 0.57 | [-3.43, -1.21] | 418 | -4.11 | **< .001** |
| Group (Retrieval vs. Noticing): Time (Pre-test vs. Delayed Posttest) | -0.34 | 0.57 | [-1.45, 0.77] | 418 | -0.60 | 0.55 |
| Group (Retrieval vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | -1.33 | 0.57 | [-2.44, -0.22] | 418 | -2.36 | 0.02 |
| Group (Retrieval vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | -0.22 | 0.55 | [-1.29, 0.86] | 418 | -0.39 | 0.69 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 9a: Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest sensitive scores of the experimental groups (*Generative Use* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.84 | 0.29 | [0.26, 1.41] | 620 | 2.86 | **< .001** |
| Group (Generative Use vs. Motivation) | -0.20 | 0.41 | [-1.01, 0.61] | 620 | -0.49 | 0.63 |
| Group (Generative Use vs. Noticing) | 0.14 | 0.41 | [-0.67, 0.95] | 620 | 0.34 | 0.73 |
| Group (Generative Use vs. Retrieval) | 0.20 | 0.42 | [-0.62, 1.01] | 620 | 0.47 | 0.6 |
| Group (Generative Use vs. All TFA Components) | 0.07 | 0.40 | [-0.72, 0.86] | 620 | 0.18 | 0.86 |
| Time (Pre-test vs. Delayed Posttest) | 5.87 | 0.40 | [5.09, 6.65] | 418 | 14.76 | **< .001** |
| Group (Generative Use vs. Motivation): Time (Pre-test vs. Delayed Posttest) | -0.99 | 0.56 | [-2.09, 0.12] | 418 | -1.76 | 0.09 |
| Group (Generative Use vs. Noticing): Time (Pre-test vs. Delayed Posttest) | 0.99 | 0.56 | [-0.11,2.09] | 418 | 1.77 | 0.09 |
| Group (Generative Use vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 1.33 | 0.57 | [0.22, 2.44] | 418 | 2.36 | **0.02** |
| Group (Generative Use vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 1.12 | 0.55 | [0.04, 2.19] | 418 | 2.05 | **0.04** |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 9b: Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest strict scores of the experimental groups (*Noticing* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 6.10 | 3.14 | [-0.01, 1.23] | 620 | 1.94 | **0.05** |
| Group (Noticing vs. Motivation) | 2.44 | 4.44 | [-0.85, 0.90] | 620 | 0.06 | 0.96 |
| Group (Noticing vs. Retrieval) | 2.90 | 4.47 | [-0.59, 1.17] | 620 | 0.65 | 0.52 |
| Group (Noticing vs. Generative Use) | 1.19 | 4.44 | [-0.87, 0.87] | 620 | 0.00 | 1.00 |
| Group (Noticing vs. All TFA Components) | 2.38 | 4.32 | [-0.61, 1.09] | 620 | 0.55 | 0.58 |
| Time (Pre-test vs. Delayed Posttest) | 7.02 | 4.27 | [6.19, 7.86] | 418 | 16.47 | **< .001** |
| Group (Noticing vs. Motivation): Time (Pre-test vs. Delayed Posttest) | -2.37 | 6.03 | [-3.55, -1.18] | 418 | 3.92 | **< .001** |
| Group (Noticing vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 1.51 | 6.07 | [-1.04, 1.34] | 418 | 0.25 | 0.80 |
| Group (Noticing vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | -1.20 | 6.03 | [-2.38, -0.01] | 418 | -1.98 | **0.05** |
| Group (Noticing vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | -2.20 | 5.87 | [-1.37, 0.93] | 418 | -0.38 | 0.71 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 9b: Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest strict scores of the experimental groups (*Retrieval* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 0.90 | 0.32 | [0.28, 1.52] | 620 | 2.83 | **< .01** |
| Group (Retrieval vs. Motivation) | -0.27 | 0.45 | [-1.14, 0.61] | 620 | -0.60 | 0.55 |
| Group (Retrieval vs. Noting) | -0.29 | 0.45 | [-1.17, 0.59] | 620 | -0.65 | 0.52 |
| Group (Retrieval vs. Generative Use) | -0.29 | 0.45 | [-1.17, 0.59] | 620 | -0.65 | 0.52 |
| Group (Retrieval vs. All TFA Components) | -0.05 | 0.43 | [-0.90, 0.80] | 620 | -0.12 | 0.90 |
| Time (Pre-test vs. Delayed Posttest) | 7.18 | 0.43 | [6.33, 8.02] | 418 | 16.61 | **< .001** |
| Group (Retrieval vs. Motivation): Time (Pre-test vs. Delayed Posttest) | -2.52 | 0.61 | [-3.71, -1.32] | 418 | -4.15 | **< .001** |
| Group (Retrieval vs. Noticing): Time (Pre-test vs. Delayed Posttest) | -0.15 | 0.61 | [-1.34, 1.04] | 418 | -0.25 | 0.80 |
| Group (Retrieval vs. Generative Use): Time (Pre-test vs. Delayed Posttest) | -1.35 | 0.61 | [-2.54, -0.15] | 418 | -2.22 | **< .05** |
| Group (Retrieval vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | -0.37 | 0.59 | [-1.53, 0.79] | 418 | -0.63 | 0.53 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 9b: Linear Mixed-effects Model comparing the Pre-test and Delayed Posttest strict scores of the experimental groups (*Generative Use* group and Pre-test as Reference Level)

| ***Fixed Effects*** | ***b*** | ***SE*** | ***95% CI*** | ***df*** | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| (Intercept) | 6.10 | 3.14 | [0.01, 1.23] | 620 | 1.94 | **0.05** |
| Group (Generative Use vs. Motivation) | 2.44 | 4.44 | [-0.85, 0.90] | 620 | 0.06 | 0.96 |
| Group (Generative Use vs. Noticing) | -2.05 | 4.44 | [-0.87, 0.87] | 620 | 0.00 | 1.00 |
| Group (Generative Use vs. Retrieval) | 2.90 | 4.47 | [-0.58, 1.17] | 620 | 0.65 | 0.51 |
| Group (Generative Use vs. All TFA Components) | 2.38 | 4.32 | [-0.61, 1.09] | 620 | 0.55 | 0.57 |
| Time (Pre-test vs. Delayed Posttest) | 5.83 | 4.27 | [5.00, 6.67] | 418 | 13.67 | **< .001** |
| Group (Generative Use vs. Motivation): Time (Pre-test vs. Delayed Posttest) | -1.17 | 5.95 | [-2.36, -0.01] | 418 | -1.94 | **0.05** |
| Group (Generative Use vs. Noticing): Time (Pre-test vs. Delayed Posttest) | 1.20 | 5.95 | [0.01, 2.38] | 418 | 1.98 | **0.05** |
| Group (Generative Use vs. Retrieval): Time (Pre-test vs. Delayed Posttest) | 1.35 | 5.98 | [0.15,2.54] | 418 | 2.22 | **0.03** |
| Group (Generative Use vs. All TFA Components): Time (Pre-test vs. Delayed Posttest) | 9.75 | 5.78 | [-0.18, 2.13] | 418 | 1.66 | 0.10 |

*Note:* *P* values are significant at the .05 level of alpha; **Bold** values reflect those with significant difference

## Appendix 10: User's manual for applying the TFA framework to evaluate vocabulary tasks/activities

### User's Manual

This manual is created based on the results of the current research project. The main aim is to provide evaluation guidelines to language teachers or researchers when applying the TFA to measure vocabulary tasks.

The two main processes below consist of 7 steps for the self-evaluation and 11 steps for the between-rater evaluation to follow. The first process is compulsory. The main purpose is to select the tasks that match the objective(s) of learning/evaluation. The second process is optional, however; it is strongly advisable to apply this process to vocabulary materials evaluation since self-evaluation can lead to bias.

**1. Self-evaluation**

1.1. Study the eighteen questions in the TFA framework proposed by Nation and Webb (2011) prior to the evaluation

1.2. Note down the TFA terms that you are not familiar with

1.3. Create the list of descriptions of difficult TFA terms that suit the purpose(s) of your evaluation/study

1.4. Review vocabulary tasks suggested by Nation and Webb (2011) and used by previous studies

1.5. Make a list of tasks to match the objective(s) of learning

1.6. Use the eighteen criteria of the TFA framework together with the list of descriptions applying to your evaluation to measure the selected tasks

1.7. Review the evaluation results (TFA scores) and go back to the previous step if needed.

*(This rechecking step is optional, but advisable for the purpose of validity in assessment.*

**2. Between-raters evaluation**

2.1. Follow steps 1.1 to 1.6 in the self-evaluation process.

2.2. Prepare evaluation materials for the other raters. These include (a) a Table of the TFA framework, (b) the list of descriptions of TFA terms applying to your evaluation/study,

(c) the evaluation form with eighteen TFA criteria and (d) details of the selected tasks with lesson plan of each task. *(see the suggested evaluation forms with an example in the Supplementary Materials: A, B and C)*

2.3. Invite three experienced-raters to the evaluation. *(All raters should have at least five-year experience in the field of vocabulary and/or materials evaluation)*

*2.4.* Operate a training for all raters to be familiar with the materials. (*It is suggested that these raters should receive the materials to study at least one week prior to the training.*)

2.5. Encourage the raters to do a trial during the training by using the example of task that is not the target task for the main evaluation.

2.6. Allow two-hour to one-week for raters to do self-evaluation

*2.7.* Collect the evaluation forms from all raters and compare the results. *(When there is a marked gap or disagreement between three raters, rely on the two raters who have the same agreement.)*

2.8. Evaluate the evaluation forms. *(If the three raters agree, the TFA criterion gets one point. See example of the comparison form in the Supplementary Materials).*

2.9. Ask the raters to provide reasons if two among three raters disagree on any criterion. (If needed invite them to join a training to improve the quality of evaluation again.)

2.10. Analyse the results from the raters again to interpret whether the TFA components have high or low potential. *(All components should be rated high or receive at least 66.68% from the three-rater evaluation to be regarded as having high potential.)*

2.11. Adjust/modify the low potential tasks and do the evaluation following steps 2.6 - 2.10 again until all tasks receive high TFA score in all components.

## Supplementary Materials for Vocabulary Task Evaluation

### A (for the rater): A Table of self-evaluation by using the TFA criteria

**Directions:** Read details of tasks in the lesson plan or task description carefully and answer the following questions.

*Give one point to the criterion if you can answer 'yes' and zero point if your answer is 'no'*

| Component | Criteria | Result |
|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | |
| | Does the activity motivate learning? | |
| | Do the learners select the words? | |
| ***Noticing*** | Does the activity focus attention on the target words? | |
| | Does the activity raise awareness of new vocabulary learning? | |
| | Does the activity involve negotiation? | |
| ***Retrieval*** | Does the activity involve retrieval of the word? | |
| | Is it productive retrieval? | |
| | Is it recall? | |
| | Are there multiple retrievals of each word? | |
| | Is there spacing between retrieval? | |
| ***Generative Use*** | Does the activity involve generative use? | |
| | Is it productive? | |
| | Is there a marked change that involves the use of other words? | |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | |
| | Does the activity involve instantiation? | |
| | Does the activity involve imagination? | |
| | Does the activity avoid interference? | |

## Supplementary Materials

### B (for the researcher): A table of the evaluation form to compare results between the three raters

| Component | Criteria | Raters' scores | | | Result |
|---|---|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | | | | |
| | Does the activity motivate learning? | | | | |
| | Do the learners select the words? | | | | |
| ***Noticing*** | Does the activity focus attention on the target words? | | | | |
| | Does the activity raise awareness of new vocabulary learning? | | | | |
| | Does the activity involve negotiation? | | | | |
| ***Retrieval*** | Does the activity involve retrieval of the word? | | | | |
| | Is it productive retrieval? | | | | |
| | Is it recall? | | | | |
| | Are there multiple retrievals of each word? | | | | |
| | Is there spacing between retrieval? | | | | |
| ***Generative Use*** | Does the activity involve generative use? | | | | |
| | Is it productive? | | | | |
| | Is there a marked change that involves the use of other words? | | | | |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | | | | |
| | Does the activity involve instantiation? | | | | |
| | Does the activity involve imagination? | | | | |
| | Does the activity avoid interference? | | | | |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*

# Supplementary Materials

## C: An example of an evaluation to compare results between the three raters

| Component | Criteria | Raters' scores | | | Result |
|---|---|---|---|---|---|
| ***Motivation*** | Is there a clear vocabulary learning goal? | 1 | 1 | 1 | **1** |
| | Does the activity motivate learning? | 1 | 1 | 1 | **1** |
| | Do the learners select the words? | 0 | 0 | 0 | **0** |
| ***Noticing*** | Does the activity focus attention on the target words? | 1 | 0 | 1 | **1*** |
| | Does the activity raise awareness of new vocabulary learning? | 0 | 0 | 0 | **0** |
| | Does the activity involve negotiation? | 0 | 0 | 1 | **0*** |
| ***Retrieval*** | Does the activity involve retrieval of the word? | 1 | 1 | 1 | **1** |
| | Is it productive retrieval? | 0 | 0 | 0 | **0** |
| | Is it recall? | 0 | 0 | 0 | **0** |
| | Are there multiple retrievals of each word? | 1 | 1 | 1 | **1** |
| | Is there spacing between retrieval? | 0 | 0 | 0 | **0** |
| ***Generative Use*** | Does the activity involve generative use? | 0 | 0 | 0 | **0** |
| | Is it productive? | 0 | 0 | 0 | **0** |
| | Is there a marked change that involves the use of other words? | 0 | 0 | 0 | **0** |
| ***Retention*** | Does the activity ensure successful linking of form and meaning? | 0 | 0 | 0 | **0** |
| | Does the activity involve instantiation? | 1 | 1 | 1 | **1** |
| | Does the activity involve imagination? | 0 | 0 | 0 | **0** |
| | Does the activity avoid interference? | 1 | 1 | 1 | **1** |

*Note: at least two-third of the analysis from the three experts received one score for each criterion; 1* and 0* = not totally agreed by all three raters*